Universidade Federal de São Carlos – UFSCar
Centro de Ciências Exatas e de Tecnologia – CCET
Departamento de Computação – DC

**Júlia Yumi Araujo Sato**

# Aprendizado multilíngue e multimodal para o português do Brasil

São Carlos
2023

**Júlia Yumi Araujo Sato**

# Aprendizado multilíngue e multimodal para o português do Brasil

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Bacharel em Engenharia da Computação.

Área de concentração: Inteligência Artificial

Orientador: Profa. Dra. Helena de Medeiros Caseli

São Carlos
2023

# Agradecimentos

Agradeço imensamente à minha orientadora, Helena Caseli, por me acompanhar durante toda a minha jornada de pesquisa científica, sempre me auxiliando e guiando no desenvolvimento deste projeto. Agradeço à minha co-orientadora, Lucia Specia, por me receber no Imperial College London, onde pude aprender muito ao seu lado. Agradeço também à FAPESP pelo apoio financeiro, uma vez que parte deste trabalho foi desenvolvido com apoio das bolsas #2020/15995-1 e #2022/04442-7. Por fim, agradeço a todos os docentes do Departamento de Computação da UFSCar, que me ajudaram a obter o conhecimento necessário para obter o meu diploma, e agradeço a minha família e namorada por sempre me apoiarem durante meus estudos.

# Resumo

Este trabalho explora o domínio da tradução automática multimodal, que é um processo que combina informações de diferentes modalidades – como texto, imagens e áudio – para realizar traduções entre idiomas. Assim, esta tarefa tem o objetivo de analisar o impacto de diversos tipos de informações associadas a texto e imagens. Baseando-se no framework Visual Translation Language Modelling (CAGLAYAN et al., 2021), aprimoramos suas capacidades para lidar com outros pares de idiomas e cenários mais complexos relativos à relação entre imagem e texto (SATO; CASELI; SPECIA, 2022). Para avaliar a capacidade de generalização do modelo, utilizamos o corpus multimodal e multilíngue How2 (SANABRIA et al., 2018), que inclui dados de vídeos com legendas em inglês e traduções em português. Além disso, tendo em vista que o *masking* – isto é, o processo de ocultação de *tokens* visuais ou linguísticos durante o treinamento – pode aprimorar a compreensão dos modelos, já que torna-se necessário prever os *tokens* ocultos com base no contexto circundante, foram propostas novas estratégias de *masking* considerando padrões linguísticos específicos e diferentes categorias semânticas (SATO; CASELI; SPECIA, 2023). Experimentos extensivos na tarefa de tradução automática multimodal português-inglês demonstram a eficácia das técnicas de *masking* mais informadas. Em particular, descobrimos que o *masking* seletivo relacionado à categoria "pessoa" melhora significativamente o desempenho, indicando seu papel crucial na interpretação de informações visuais. Essas descobertas oferecem *insights* sobre o comportamento do modelo e contribuem para o desenvolvimento de abordagens de *masking* mais eficazes na tradução automática multimodal. Por fim, vale destacar que a abordagem proposta neste trabalho alcança estado-da-arte no conjunto de dados How2 (pontuação BLEU de 53.1) e fornece informações valiosas sobre a interação entre imagens e textos em sistemas de tradução.

**Palavras-chave:** Tradução automática multimodal. Visão e Linguagem. Masking. Português brasileiro. Processamento de Linguagem Natural.

# Abstract

This work explores the domain of multimodal machine translation, which is a process that combines information from different modalities – such as text, images, and audio – to perform translations between languages. Thus, this task aims to analyze the impact of various types of information associated with text and images. Building upon the Visual Translation Language Modelling framework (CAGLAYAN et al., 2021), we enhanced its capabilities to handle other language pairs and more complex scenarios related to the image-text relationship (SATO; CASELI; SPECIA, 2022). To evaluate the model's generalization ability, we used the multimodal and multilingual corpus How2 (SANABRIA et al., 2018), which includes videos with English subtitles and crowdsourced Portuguese translations. Furthermore, considering that masking (i.e., the process of hiding visual or linguistic tokens during training) can enhance model understanding, as it requires predicting the hidden tokens based on the surrounding context, new masking strategies were proposed considering specific linguistic patterns and different semantic categories (SATO; CASELI; SPECIA, 2023). Extensive experiments in the Portuguese-English multimodal machine translation task demonstrate the effectiveness of the more informed masking techniques. In particular, we found that selective masking related to the 'person' category significantly improves performance, indicating its crucial role in interpreting visual information. These findings offer insights into the model's behavior and contribute to the development of more effective masking approaches in multimodal machine translation. Finally, it is worth noting that the approach proposed in this work achieved state-of-the-art results on the How2 dataset (53.1 BLEU) and provided valuable information about the interaction between images and texts in translation systems.

**Keywords:** Multimodal Machine Translation, Vision and Language, Masking, Brazilian Portuguese, Natural Language Processing.

# Lista de ilustrações

# Lista de tabelas

# Lista de siglas

**BERT** *Bidirectional Encoder Representations from Transformers*

**CLM** Modelagem de Linguagem Causal

**ERNIE** *Enhanced Representation through kNowledge IntEgration*

**ELECTRA** *Efficiently Learning an Encoder that Classifiers Token Replacements Accurately*

**MMT** tradução automática multimodal

**MLM** Modelagem de Linguagem Mascarada

**MT** tradução automática

**MRC** classificação de região mascarada

**NSP** Previsão de Próxima Frase

**NMT** tradução automática neural

**PLN** Processamento de Linguagem Natural

**PMI** *Pointwise Mutual Information*

**RNN** rede neural recorrente

**R-CNN** *Region-based Convolutional Neural Network*

**SPP-net** *Spatial Pyramid Pooling*

**SSD** *Single Shot MultiBox Detecto*

**TLM** Modelagem de Linguagem de Tradução

**VTLM** *Vision Translation Language Modelling*

**XLM** *Cross-lingual Language Model*

**XLU** *cross-lingual understanding*

# Sumário

# Capítulo 1

# Introdução

Os humanos lidam constantemente com informações multimodais, ou seja, conjuntos de dados de diferentes modalidades, como texto e imagens. Para as máquinas processarem a informação de forma semelhante aos humanos, elas devem ser capazes de processar dados multimodais e compreender a relação entre essas modalidades, não apenas texto ou imagens de forma isolada, por exemplo. Esse aspecto multimodal do aprendizado pode ser bastante útil em aplicações multilíngue, isto é, aplicações que envolvem dois ou mais idiomas.

Enquanto os modelos multimodais são treinados para serem capazes de interpretar e associar dados de diferentes modalidades – como texto, áudio e imagem simultaneamente – os modelos multilíngues devem entender vários idiomas aprendendo representações multilíngues. Nesse contexto, os modelos que aprendem representações multimodais e multilíngues demonstraram ter um melhor desempenho em muitas tarefas de linguagem natural (CAGLAYAN et al., 2021).

Recentemente, a área de Processamento de Linguagem Natural (PLN) vem vivenciando uma mudança significativa de paradigma com a proposição de diversos modelos neurais (*deep learning*) para processamento da língua (DEVLIN et al., 2019; CONNEAU; LAMPLE, 2019; RADFORD et al., 2018). Esses avanços se baseiam no uso de redes neurais artificiais e estratégias como *transfer learning*, que se baseia na ideia de utilizar o conhecimento aprendido em uma tarefa específica para aplicar em outras tarefas, e mecanismo de atenção (*attention*), que permite a codificação seletiva de informações em um vetor de palavras baseado na relevância de determinada palavra no contexto.

Há diversos exemplos de aplicações de PLN onde essas estratégias alcançaram o estado da arte no processamento monolíngue (LAN et al., 2020; LIU et al., 2019; ROTHE; NARAYAN; SEVERYN, 2020), multilíngue (DEVLIN et al., 2019; CONNEAU; LAM-

PLE, 2019) e multimodal (TAN; BANSAL, 2019; LU et al., 2019; LI et al., 2019; LIN et al., 2021). Contudo, para o português esses avanços ainda são bastante iniciais (SOUZA; NOGUEIRA; LOTUFO, 2020).

Embora o interesse inicial fosse apenas em modelos multimodais (TAN; BANSAL, 2019; LU et al., 2019; LI et al., 2019) ou multilíngues (DEVLIN et al., 2019; CONNEAU; LAMPLE, 2019), desenvolvimentos recentes resultaram em estruturas capazes de fazer os dois e fornecer modelos multilíngues e multimodais (CAGLAYAN et al., 2021; HUANG et al., 2021; NI et al., 2021).

Nesse contexto, a modalidade visual pode ajudar as máquinas a entender melhor as informações textuais. Essa abordagem foi introduzida em uma tarefa de tradução automática neural multimodal (MNMT) (SPECIA et al., 2016; ELLIOTT et al., 2017; BARRAULT et al., 2018), que se concentra principalmente no aprimoramento da tradução somente de texto com *features* visuais. Portanto, a tradução automática multimodal (MMT) melhora a qualidade da tradução usando o contexto da modalidade visual adicional. Com isso, espera-se que a tradução seja mais precisa, já que o contexto visual ajuda a reduzir a ambiguidade.

O *Vision Translation Language Modelling* (VTLM), proposto por Caglayan et al. (2021), mostrou que o pré-treinamento multimodal e multilíngue leva a melhorias consideráveis para a MMT em comparação à performance de modelos que possuem apenas o pré-treinamento multilíngue, sem a parte visual. Isso destaca a eficácia do pré-treinamento multimodal e multilíngue para a tarefa MMT.

No entanto, a maioria dos modelos pré-treinados segue o paradigma de pré-treinamento do BERT (DEVLIN et al., 2019) e adota *masked language modeling* (MLM) e suas variantes para aprender representações mascarando *tokens* e fazendo previsões com base em seu contexto. O MLM convencional depende da seleção aleatória de *tokens* a serem mascarados e, portanto, pode não considerar informações linguísticas que podem ser úteis para algumas tarefas de PLN, como MMT.

Assim, este trabalho propõe uma maneira de atacar o problema de processamento multimodal e multilíngue envolvendo o idioma português, por meio de uma extensão da *framework* VTLM adaptada para o português brasileiro e circunstâncias mais desafiadoras relacionadas à relação texto-imagem. Para tanto, realizou-se um estudo aprofundado de estratégias que pudessem resultar em um pré-treinamento mais eficiente, como a análise do efeito do *masking* de regiões visuais para a tarefa *Masked Region Classification* (MRC), no contexto da tradução automática multimodal.

O *masking* é o processo de ocultação de *tokens* visuais ou linguísticos durante o treinamento. Esse procedimento tem o potencial de aprimorar a compreensão dos modelos, já que torna-se necessário prever os *tokens* ocultos com base no contexto circundante. A análise do efeito do *masking* se torna particularmente pertinente neste cenário devido ao emprego frequente do Modelagem de Linguagem Mascarada (MLM) no treinamento dos

modelos de linguagem. Então, aprimorar a técnica de *masking* pode refletir em avanços na qualidade e eficácia de modelos voltados para tarefas como a MMT.

A análise do *masking* proposta neste trabalho envolve propor novas estratégias de mascaramento para explorar abordagens mais informadas de mascarar *tokens* visuais. Essas abordagens estão relacionadas à mudança na forma como o mascaramento é realizado de modo a não mascarar *tokens* aleatórios e apenas mascarar *tokens* da categoria desejada, focando em situações que podem ser favorecidas por uma melhor compreensão da informação visual. Assim, será possível focar no aprendizado de elementos visuais que podem auxiliar em situações mais difíceis de interpretar apenas com informações textuais.

Por exemplo, como a maioria dos modelos de linguagem pré-treinados é baseada no inglês, eles não conseguem entender alguns padrões linguísticos comuns em muitos outros idiomas, como o gênero das palavras. A língua inglesa trata o gênero das palavras de forma diferente de línguas como francês, espanhol, português ou italiano. Enquanto algumas línguas possuem palavras diferentes com o mesmo significado que se encontram nas formas feminina e masculina, isso não acontece na língua inglesa. Por exemplo, considerando a tradução inglês-português, o pronome “they” pode ser traduzido para “elas” (feminino) ou “eles” (masculino). Outro exemplo é o adjetivo “beautiful”, que pode ser traduzido como “bonita” (feminino) ou “bonito” (masculino) dependendo de quem ou a que se refere.

Nesse contexto, propomos três estratégias de *masking* seletivo – *masking* visual mais informado, *masking* textual mais informado e *masking* visual e textual mais informado – cada uma com foco no mascaramento de *tokens* linguísticos e visuais específicos que podem contribuir para uma melhor compreensão de alguns desses diferentes padrões. Estas novas estratégias foram incorporadas ao VTLM e um extenso conjunto de experimentos foram executados para verificar o impacto das mudanças para a tarefa MMT Português-Inglês.

Por fim, exploramos o uso do *masking* em diferentes categorias semânticas para melhorar a tradução automática multimodal. Inicialmente, estudamos o corpus How2, analisando elementos visuais e textuais que aparecem com mais frequência neste conjunto de dados. Em seguida, aumentamos o *threshold* na detecção de objetos para aprimorar o mascaramento seletivo de determinados *tokens* visuais e avaliamos os efeitos dessa abordagem na tradução. Além disso, investigamos o impacto de diferentes proporções de *masking* em categorias semânticas-chave, como “person”, “clothing” e “furniture”, que aparecem de forma predominante no corpus.

De forma geral, este trabalho destaca a relevância do *masking* seletivo em categorias semânticas específicas para aprimorar a tradução automática multimodal e proporciona *insights* sobre o comportamento do modelo em relação a diferentes tipos de informações semânticas. Essas descobertas podem contribuir para o desenvolvimento de técnicas mais eficazes de *masking* em abordagens multimodais e enriquecer a compreensão sobre a interação entre imagens e textos em sistemas de tradução automática.

## 1.1 Objetivo

Neste contexto, este trabalho visa investigar o impacto do *masking* textual e visual mais linguisticamente informado na Tradução Automática Multimodal português-inglês. Para tanto, foram realizados experimentos usando o conjunto de dados How2 (SANABRIA et al., 2018) e uma extensão da *framework* VTLM (CAGLAYAN et al., 2021) proposta pela autora deste trabalho e suas orientadoras (SATO; CASELI; SPECIA, 2022).

## 1.2 Organização da monografia

Este texto está organizado como segue. No Capítulo 2, são descritos os principais conceitos relacionados a modelos multilíngues e multimodais. Em seguida, o Capítulo 3 descreve brevemente algumas das propostas da literatura para a tradução automática multimodal e as estratégias de *masking* em modelos pré-treinados. O Capítulo 4 descreve a extensão da *framework* VTLM para legendas de vídeo e um novo par de língua, e a exploração de estratégias de *masking* mais informadas e com diferentes categorias semânticas. O Capítulo 5 descreve os recursos e os métodos utilizados para avaliar a performance dos modelos com as mudanças propostas, assim como a apresentação dos resultados obtidos em cada etapa do desenvolvimento. Por fim, o último capítulo evidencia as conclusões deste trabalho.

# Capítulo 2

# Fundamentação teórica

Este capítulo apresenta os principais conceitos relacionados a modelos multilíngues e multimodais, com ênfase para a estratégia selecionada para ser usada neste trabalho: o VTLM.

A *framework* VTLM, junto a outros modelos *transformers*[1], como BERT, LXMERT e XLM são apresentados brevemente a seguir com o intuito de explicitar seu modo de funcionamento e suas limitações.

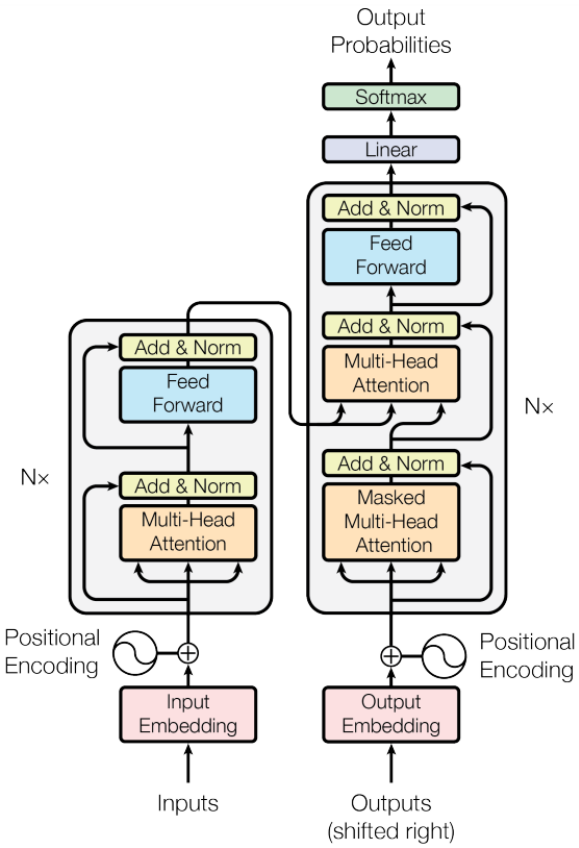
**BERT** O *Bidirectional Encoder Representations from Transformers* (BERT) (DEVLIN et al., 2019) é um modelo de linguagem baseado no pré-treinamento de representações bidirecionais de texto. Sua arquitetura é fundamentada na arquitetura *Transformer* (VASWANI et al., 2017a), ilustrada na Figura 1, que se baseia em mecanismos de atenção para capturar relações contextuais entre palavras em textos, o que elimina a necessidade de sequencialidade.

Assim, a arquitetura do BERT inclui várias camadas de codificadores do *Transformer*. Cada camada é responsável por transformar a representação dos *tokens* de entrada em representações mais ricas e contextuais. Os *tokens* de entrada são alimentados em várias camadas, e a saída da última camada é usada para uma variedade de tarefas específicas, como classificação de sentimento, reconhecimento de entidades nomeadas, tradução, entre outras.

Em termos de treinamento, o BERT possui duas etapas: pré-treinamento e ajuste fino (*fine-tuning*). Na etapa de pré-treinamento, o BERT é treinado usando duas tarefas não supervisionadas: (1) MLM e (2) Previsão de Próxima Frase (NSP). A

[1] Outras fontes de informação, além dos artigos científicos que descrevem os modelos, são os repositórios como o https://huggingface.co/transformers/

Figura 1 – Arquitetura *Transformer*




Fonte: Vaswani et al. (2017a).

primeira tarefa tem o objetivo de permitir representações bidirecionais profundas pré-treinadas mascarando de forma aleatória uma porcentagem dos *tokens* de entrada para serem posteriormente preditos com base apenas em seu contexto. E a segunda tarefa tem o intuito de entender a relação entre as frases durante o pré-treinamento, predizendo quem seria a próxima frase dada a frase atual.

Para o ajuste fino, o modelo é inicializado com os parâmetros pré-treinados, sendo que, para cada tarefa, as entradas e saídas específicas da tarefa são conectadas e, em seguida, todos os parâmetros são ajustados de ponta a ponta[2].

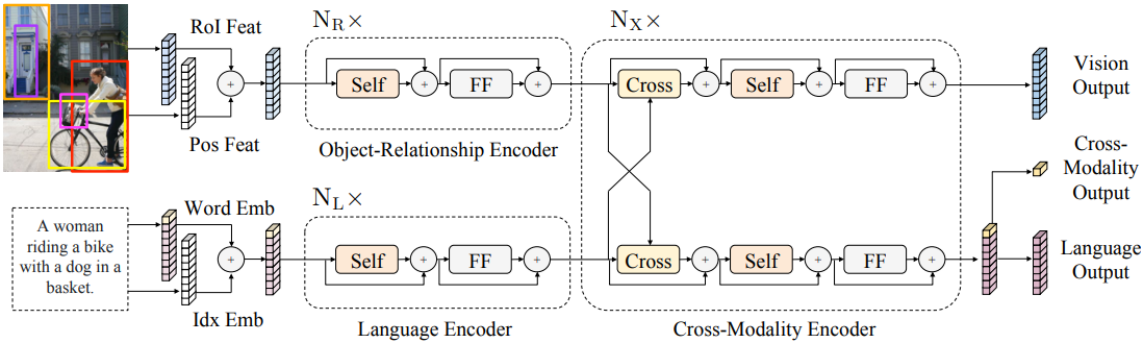
O BERT supera os métodos propostos antes dele porque é o primeiro sistema não supervisionado e profundamente bidirecional para pré-treinamento. Isso significa que o BERT foi treinado usando apenas um corpus de texto sem anotações e usa contextos da esquerda e da direita para representar uma palavra, ao contrário dos modelos tradicionais de PLN em que cada palavra é contextualizada usando apenas as palavras à sua esquerda (ou direita).

[2] Todas as camadas e parâmetros do modelo são ajustados em conjunto, permitindo que o modelo aprenda a extrair *features* relevantes diretamente das entradas e aplicá-las para gerar as saídas desejadas, sem depender de etapas intermediárias

Além disso, o BERT pode ser ajustado com uma camada de saída adicional para criar modelos para diversas tarefas de PLN. Dessa forma, o BERT é usado como base para muitos modelos dessa área (LIU et al., 2019; LAN et al., 2020; CLARK et al., 2020).

**LXMERT** *Learning Cross-Modality Encoder Representations from Transformers* (LXMERT) (TAN; BANSAL, 2019) é uma *framework* baseada no BERT, adaptada para construir um modelo *cross-modal* para aprendizado de relações entre as modalidades (visão e linguagem). Sua arquitetura é apresentada na Figura 2. O modelo consiste em três codificadores *Transformer*: (1) um codificador de relacionamento de objeto, (2) um codificador de linguagem e (3) um codificador de modalidade cruzada.

Figura 2 – Arquitetura do LXMERT




Fonte: Tan e Bansal (2019).

As duas entradas do modelo são uma imagem e uma frase relacionada a esta imagem. Cada frase é representada como uma sequência de palavras e cada imagem como uma sequência de objetos, que são obtidos previamente a partir da detecção de objetos nas imagens de entrada e extração de *features*. E as três saídas são: (1) representações de linguagem, (2) representações de imagem e (3) representações de modalidade cruzada.

Seu treinamento multimodal permite a inferência de *tokens* mascarados tanto entre elementos da mesma modalidade quanto em outra modalidade considerando os componentes alinhados. Essa abordagem ajuda a construir relacionamentos intra e intermodais.

**XLM** A abordagem do *Cross-lingual Language Model* (XLM), proposta por (CONNEAU; LAMPLE, 2019), engloba dois métodos para aprendizagem de modelos de linguagem multilíngue: (1) um não supervisionado, que envolve dois objetivos de pré-treinamento monolíngues; e (2) um supervisionado, que usa dados paralelos para a geração de modelo de linguagem multilíngue. Dados paralelos são dados em um idioma acompanhados de suas traduções para outro idioma. Por exemplo, um cor-

pus paralelo português-inglês é composto por textos em português acompanhados de suas traduções para o inglês ou vice-versa.

São apresentados, então, três estratégias de treinamento de modelos de linguagem: Modelagem de Linguagem Causal (CLM), MLM e Modelagem de Linguagem de Tradução (TLM). As duas primeiras são não supervisionadas e apenas requerem dados monolíngues, mas têm a desvantagem de não poderem utilizar informações de dados paralelos, mesmo quando estes estão disponíveis; e a terceira é uma extensão do MLM em que frases paralelas são concatenadas em vez de utilizar uma série de textos monolíngues, ou seja, é uma estratégia supervisionada e requer dados paralelos. Como resultado, as abordagens MLM e CLM oferecem recursos multilíngues que podem ser usados para modelos de pré-treinamento e a TLM, em conjunto com a MLM, melhora o pré-treinamento multilíngue quando dados paralelos estão disponíveis.

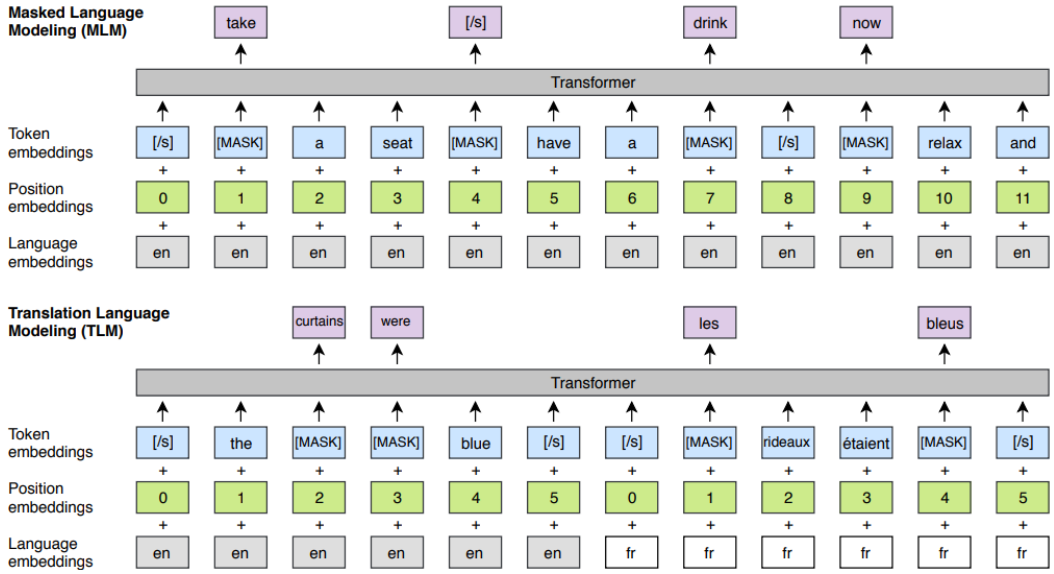
A Figura 3 ilustra as estratégias MLM e TLM. Os *position embeddings*, mostrados na figura, são uma parte essencial das arquiteturas de modelos de linguagem. Esses *embeddings* são introduzidos para permitir que o modelo leve em consideração a ordem das palavras em uma sequência de texto. Em uma rede neural, cada palavra ou *token* é representado como um vetor numérico (*embedding*) que codifica suas características semânticas. No entanto, para capturar a informação de posição, especialmente em sequências mais longas, é necessário fornecer ao modelo uma maneira de diferenciar as posições relativas das palavras na sequência. Portanto, estes *embeddings* permitem que o modelo saiba qual é a posição de cada palavra em relação às outras, o que é importante para a compreensão adequada do contexto e da estrutura da frase.

O XLM também fornece um método de pré-treinamento eficiente para tarefas de *cross-lingual understanding* (XLU). Suas aplicações incluem tradução automática não supervisionada e supervisionada e classificação *cross-lingual* (XNLI), isto é, depois do ajuste fino de um modelo XLM em um corpus de treinamento de um determinado idioma, o modelo ainda é capaz de fazer previsões precisas em outras línguas, para as quais há pouco ou nenhum dado de treinamento.

**VTLM** O VTLM (*Visual Translation Language Modelling*) (CAGLAYAN et al., 2021) combina aprendizado multimodal e multilíngue para gerar representações *cross-lingual* e multimodal com o intuito de melhorar a eficácia na tradução automática multimodal. Para isso, o modelo faz a junção do TLM (*Translation Language Modelling*) proposto por (CONNEAU; LAMPLE, 2019) com classificação de região mascarada (MRC) (CHEN et al., 2020; SU et al., 2020).

A arquitetura do VTLM (Figura 4) estende o TLM adicionando uma modalidade visual ao lado dos pares de tradução, e o modelo final processa pares de tradução

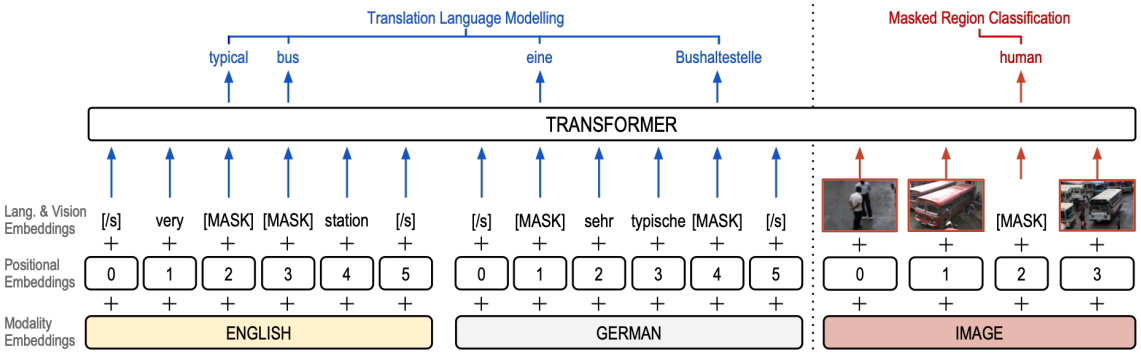
Figura 3 – Pré-treinamento do XLM, ilustrando as estratégias MLM e TLM




Fonte: Conneau e Lample (2019).

e *features* de região projetada em um fluxo único. Essas *features* são extraídas por meio de um modelo Faster R-CNN (REN et al., 2015) pré-treinado no dataset Open Images (KUZNETSOVA et al., 2020).

Figura 4 – Arquitetura do VTLM




Fonte: Caglayan et al. (2021).

O VTLM define a entrada $x$ como a concatenação de sentenças de idioma de origem de comprimento $m$ ($s^{(1)}_{1:m}$), sentenças de idioma de destino de comprimento $n$ ($s^{(2)}_{1:n}$), e $\{v_1, \cdots, v_o\}$ *features* de imagem correspondentes:

$$x = [s^{(1)}_1, \cdots, s^{(1)}_m, s^{(2)}_1, \cdots, s^{(2)}_n, v_1, \cdots, v_o] \tag{1}$$

O modelo final combina o erro do TLM com o erro do MRC de acordo com a seguinte equação:

$$\mathcal{L} = \frac{1}{|X|} \sum_{x \in \mathcal{X}} \log Pr(\{\hat{y}, \hat{v}\} | \tilde{x}; \theta) \tag{2}$$

onde $\tilde{x}$ é a sequência de entrada mascarada, $\hat{y}$ são os alvos reais para posições mascaradas, $\hat{v}$ são os rótulos de detecção e $\theta$ são os parâmetros do modelo.

Nesta abordagem, o *masking* é aleatório e se aplica a *tokens* linguísticos e visuais. Sua proporção é de 15% e é aplicado separadamente nos fluxos visual e de linguagem. O VTLM substitui seus vetores de *features* projetadas pelo *token* [MASK], sendo que 10% do *masking* equivale ao uso de *features* de região selecionadas aleatoriamente de todas as imagens no *batch*, e os 10% restantes das regiões são deixados intactos.

Seu pré-treinamento conta com recursos visuais e *cross-lingual* e realiza modelagem de linguagem mascarada e classificação de região mascarada em um dataset de visão e linguagem paralelo de três vias, que trata-se de uma extensão do corpus *Conceptual Captions* (CC) (SHARMA et al., 2018) com traduções automáticas alemãs.

Após o pré-treinamento, o codificador VTLM é transferido para um modelo de tradução automática multimodal baseado em *Transformer* e ajustado para a tarefa de tradução automática multimodal. Nesta etapa, o corpus Multi30k (ELLIOTT et al., 2016) foi utilizado.

Neste trabalho, optamos por utilizar o VTLM de Caglayan et al. (2021). O VTLM estende a *framework* TLM (CONNEAU; LAMPLE, 2019) com *features* regionais e introduz uma abordagem de pré-treinamento que combina pré-treinamento multilíngue e visual. Ele executa modelagem de linguagem mascarada e classificação de região mascarada em um conjunto de dados de visão e linguagem paralela de três vias, que é uma extensão do corpus Conceptual Captions (SHARMA et al., 2018) com traduções automáticas em alemão. O VTLM alcançou uma pontuação BLEU de 44,0 e uma pontuação METEOR de 61,3 no conjunto de teste inglês-alemão (En-De) 2016 do Multi30k para a tarefa MMT, demonstrando a eficácia do pré-treinamento multimodal e multilíngue.

# Capítulo 3

# Trabalhos relacionados

Este capítulo descreve trabalhos que investigaram a tradução automática multimodal por meio de várias estratégias, bem como os que experimentaram outras estratégias de *masking* em modelos pré-treinados.

## 3.1 Tradução automática multimodal

A tradução automática multimodal, diferente da tradução automática (MT) tradicional, considera outras modalidades além da informação do texto para traduzir melhor as sentenças de origem para as de destino.

Trabalhos anteriores propõem vários modelos e métodos de tradução automática multimodal. Huang et al. (2016b) concatena *features* visuais globais e regionais com texto para atender a imagem e o texto durante a decodificação. Calixto e Liu (2017) usam *features* de imagem globais para inicializar os estados ocultos do codificador/decodificador da rede neural recorrente (RNN). Elliott e Kádár (2017) apresentam uma *framework* de aprendizado multitarefa para aprender representações visualmente fundamentadas e aprender a traduzir. Zhou et al. (2018) aprimora o aprendizado de um *embedding* de linguagem visual compartilhado e um tradutor baseado em atenção multimodal por meio de um mecanismo de atenção visual. Calixto, Rios e Aziz (2019) apresentam um modelo de variável latente para aprender as interações entre *features* visuais e textuais. Ive, Madhyastha e Specia (2019) propõem um método de traduzir e refinar baseado no Transformer (VASWANI et al., 2017b), onde as imagens são usadas apenas por um decodificador de segundo estágio. Yin et al. (2020) utilizam um grafo multimodal unificado para capturar diferentes relações semânticas.

Ao contrário dos trabalhos anteriores, Yao e Wan (2020) propõem a autoatenção multimodal no Transformer para resolver o problema de importância relativa entre diferentes modalidades. Este problema refere-se à questão de como atribuir pesos ou relevância a diferentes tipos de informações provenientes de várias modalidades ao lidar com tarefas multimodais, pois nem todas as informações de todas as modalidades contribuem igualmente para o resultado final. Eles mostram uma abordagem para incorporar informações de outra modalidade com base em uma perspectiva gráfica do Transformer e evitam codificar informações irrelevantes em imagens aprendendo as representações de imagens com base no texto. Essa proposta foi avaliada no conjunto de dados Multi30k (ELLIOTT et al., 2016), que contém 29.000 instâncias para treinamento, 1.024 para validação e 1.000 para teste (Test2016). O modelo gerado atingiu uma pontuação METEOR de 55,7 e uma pontuação BLEU de 38,7 no conjunto de teste inglês-alemão (En-De) 2016, demonstrando o benefício da modalidade visual ao superar sua *baseline* somente de texto em mais de 1 ponto BLEU.

Nessa linha, Liu, Cao e Zhao (2021) introduzem um método de seleção em cenários multimodais denominado Gumbel-Attention. Ele seleciona as partes relacionadas ao texto das *features* da imagem e remove as informações irrelevantes usando um método diferenciável. Eles também usam o conjunto de dados Multi30k e apresentam seus resultados no conjunto de teste inglês-alemão. O modelo Gumbel-Attention MMT obteve um melhor desempenho (39,2 BLEU e 57,8 METEOR no Multi30k Test2016) em relação ao Multimodal Transformer (YAO; WAN, 2020), que atingiu 38,7 BLEU e 55,7 METEOR.

Em contraste com estudos anteriores, Long, Wang e Li (2021) introduzem um método de tradução automática que só precisa da frase de origem no momento da inferência. Eles criaram um modelo generativo baseado em imaginação chamado ImagiT, que aprende a produzir representação visual a partir da frase de origem e, em seguida, gera a frase da língua-alvo usando a frase de origem e a "representação imaginada". Semelhante ao trabalho anterior, os experimentos foram conduzidos no conjunto de dados Multi30k. Seus melhores resultados foram no conjunto de teste inglês-alemão (En-De) 2017, alcançando uma pontuação BLEU de 32,4 e uma pontuação METEOR de 52,5, e no conjunto de teste inglês-francês (En-Fr) 2016, obtendo 59,9 BLEU e 74,3 METEOR. Seus resultados mostram melhorias em relação às *baselines* NMT somente de texto, demonstrando a eficácia de seu modelo.

## 3.2 *Masking* em modelos pré-treinados

A modelagem de língua por meio do mascaramento (*masking*), uma das estratégias propostas com o BERT (DEVLIN et al., 2019), visa aprender com eficiência as representações bidirecionais, mascarando um conjunto de *tokens* de entrada aleatoriamente e prevendo-os posteriormente. Nesta abordagem, 15% dos *tokens* de entrada são seleciona-

dos aleatoriamente para o *masking*, dos quais 80% são substituídos pelo *token* `[MASK]`, 10% são substituídos por um *token* aleatório e 10% são deixados intactos.

Após o BERT, várias abordagens foram propostas para otimizar modelos de linguagem pré-treinados. Devlin et al. (2019) posteriormente propôs o *whole word masking* (wwm) em uma tentativa de resolver as desvantagens do *masking* de *token* aleatório na tarefa MLM. Nessa abordagem, ao invés de selecionar aleatoriamente *tokens* WordPiece (WU et al., 2016) para mascarar, todos os *tokens* correspondentes a uma palavra completa são mascarados de uma vez. Isso força o modelo a recuperar a palavra completa na tarefa MLM, ao invés de apenas recuperar *tokens* WordPiece. Os autores obtiveram resultados estado-da-arte em onze tarefas de PLN, incluindo pontuação GLUE de 80.5, acurácia MultiNLI de 86.7%, e pontuação de 93.2 no SQuAD v1.1 e 83.1 no SQuAD v2.0.

Zhang et al. (2019) apresenta o *Enhanced Representation through kNowledge IntEgration* (ERNIE) para otimizar o processo de *masking* do BERT aplicando *masking* de entidade/frase. Em vez de selecionar palavras de entrada aleatoriamente, o *masking* em nível de frase mascara palavras consecutivas e o *masking* em nível de entidade mascara as entidades nomeadas. Os resultados experimentais demonstraram que o ERNIE é comparável ao BERT em tarefas comuns de PLN.

Joshi et al. (2020) propõem *masking* aleatório de *span*, onde eles selecionam iterativamente *spans* aleatórios contíguos para o *masking*. A taxa de *masking* padrão de 15% é mantida, mas neste caso os *spans* são as unidades. Eles amostram um comprimento de *span* de uma distribuição geométrica em cada iteração e selecionam aleatoriamente o ponto inicial para o *span* a ser mascarado. Dessa forma, com os mesmos dados de treinamento do BERT, os resultados obtidos foram de 94.6 no SQuAD v1.1 e 88.7 no SQuAD v2.0.

Em paralelo, Clark et al. (2020) apresentaram o *Efficiently Learning an Encoder that Classifiers Token Replacements Accurately* (ELECTRA), que usa uma *framework* gerador-discriminador. Enquanto o gerador aprende a prever as palavras originais dos *tokens* mascarados, o discriminador usa a *Replaced Token Detection* para discriminar se o *token* de entrada é substituído pelo gerador. Os autores mostraram que os ganhos são especialmente notáveis para modelos pequenos, que podem superar o desempenho de um modelo treinado com 30 vezes mais capacidade de processamento. Eles também mostraram que esta abordagem funciona bem em grande escala, onde tem um desempenho comparável ao RoBERTa (ZHUANG et al., 2021) e XLNet (YANG et al., 2020), usando menos de 1/4 de seu poder de processamento, e superando-os quando usando a mesma quantidade de capacidade de processamento.

Levine et al. (2021) destacam a ineficiência do *masking* aleatório de *tokens* e propõem uma estratégia de *masking* baseada no conceito de *Pointwise Mutual Information* (PMI). *PMI-masking* que conjuntamente mascara um *token* $n$-gram se ele exibe alta colocação sobre o corpus. Os experimentos mostraram que o *PMI-masking* atinge a performance de

métodos anteriores de *masking* em metade do tempo de treinamento e pode melhorar a performance do modelo no final do treinamento.

Xiao et al. (2021) introduz um método $n$-gram de *masking* chamado ERNIE-Gram para focar nas intra-dependências e inter-relações de informações linguísticas de granulação grossa. Nesta abordagem, $n$-gramas são mascarados com símbolos `[MASK]`, e preditos diretamente usando identidades explícitas de $n$-gramas ao invés de sequências contíguas de $n$ *tokens*. Os resultados obtidos ultrapassaram modelos anteriores, como XLNet (YANG et al., 2020) e RoBERTa (ZHUANG et al., 2021), e atingiram resultados comparáveis com métodos estado-da-arte.

Por fim, no caso do VTLM, executa-se *masked language modeling* e *masked region classification* em um conjunto de dados de visão e linguagem paralelo de três vias. A taxa de *masking* padrão é mantida (ou seja, 15%) e é aplicada separadamente aos fluxos visuais e de linguagem. Neste trabalho, nós incorporamos estratégias de *masking* mais formadas ao VTLM, de forma a não selecionar os *tokens* de maneira aleatória para o *masking* e focar em mascarar *tokens* específicos para aprender determinados padrões de linguagem com eficiência.

# Capítulo 4

# Desenvolvimento

Este capítulo descreve o corpus (seção 4.1) e os métodos utilizados neste trabalho para geração dos modelos *baseline* (seção 4.2) e dos modelos com *masking* mais informado para a tradução multimodal envolvendo os idiomas português e inglês.

## 4.1 Corpus How2

Para permitir a avaliação das estratégias investigadas neste trabalho, utilizou-se o corpus How2 (SANABRIA et al., 2018). Trata-se de uma coleção multimodal e multilíngue de cerca de 80.000 vídeos instrucionais (cerca de 2.000 horas), acompanhados de legendas em inglês, e cerca de 300 horas de traduções coletadas em português obtidas com *crowdsourcing*, além de resumos de cada vídeo, em inglês. O corpus How2 foi utilizado em todas as etapas de experimentação com o VTLM.

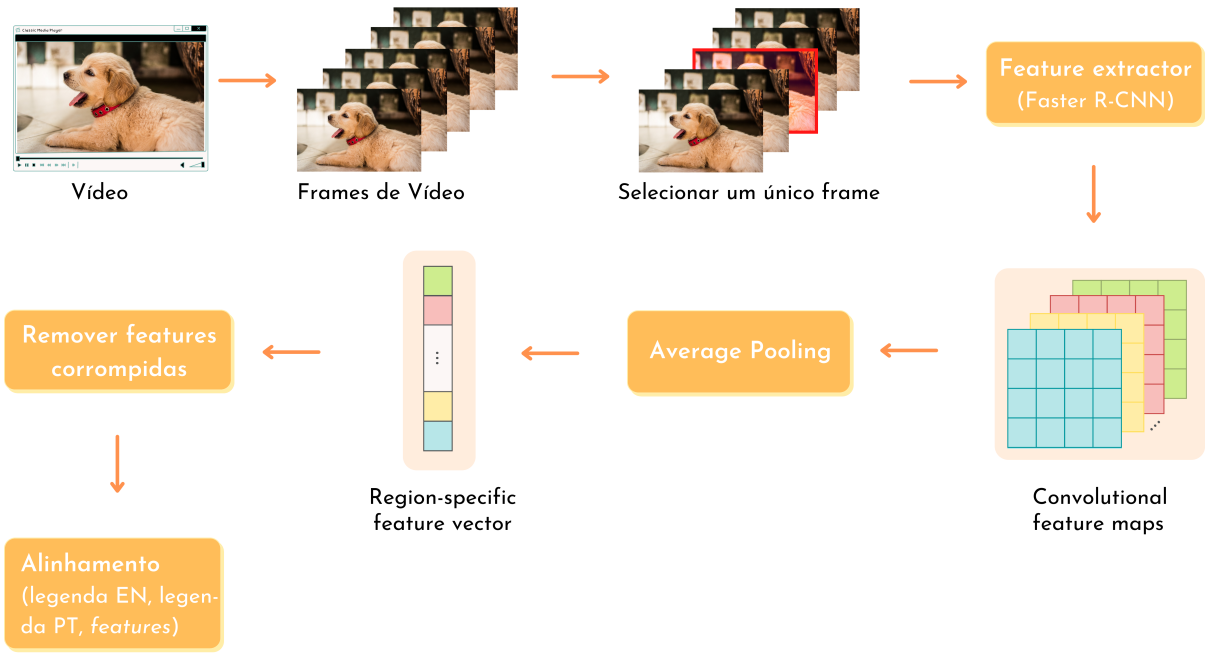
Como descrito anteriormente, o How2 é uma coleção de vídeos – o que significa que o texto associado a cada imagem (*frame* de vídeo) não é uma simples descrição da mesma, mas sim uma legenda que pode não estar relacionada ao seu *frame* correspondente. Esta característica do How2 difere do corpus usado originalmente no VTLM, que é uma coleção de imagens estáticas associadas a suas respectivas legendas, ou seja, cada imagem possui uma única frase que está semanticamente alinhada com ela (a frase é uma descrição da imagem). Como consequência, além da diferença de idiomas, o uso deste corpus traz novas questões que devem ser tratadas para que ele possa ser utilizado na experimentação com o VTLM.

Para isso, foram propostas e realizadas algumas etapas de pré-processamento ilustradas na Figura 5. Este processo iniciou-se com a extração de *features* a partir dos *frames* de vídeo. Para cada segmento, selecionou-se o frame do meio e extraiu-se *convolutional*

*feature maps* das 36 regiões mais confiáveis usando o modelo Faster R-CNN (REN et al., 2016) pré-treinado no dataset Open Images (KUZNETSOVA et al., 2020) e depois foi feito *average pool* de cada um para obter um vetor de *features* específicas da região. O Faster R-CNN é um modelo de detecção de objetos que é capaz de localizar e classificar objetos em uma imagem, atribuindo a cada região da imagem uma pontuação de confiança que indica a probabilidade de conter um objeto. Assim, as regiões mais confiáveis referem-se a áreas específicas em um *frame* de vídeo que foram identificadas pelo modelo Faster R-CNN como sendo altamente relevantes para a tarefa em questão.

Este processo teve continuidade com o alinhamento bilíngue e multimodal, ou seja, as *features* obtidas no processo de extração foram associadas aos seus textos correspondentes em inglês e português. Além disso, também houve o desenvolvimento de um *script* para desconsiderar *features* que poderiam estar corrompidas, pois o VTLM não oferece suporte a *features* inexistentes ou não carregáveis, então isto poderia prejudicar os experimentos. Este processo se deu pela identificação de todas as *features* disponíveis e verificação automática de cada uma, utilizando um *script* Python, para identificar possíveis problemas e, caso necessário, realizar sua remoção.

Figura 5 – Etapas de pré-processamento




Assim, após identificar todas as *features* válidas, os conjuntos de treinamento, teste e validação do How2 foram modificados para eliminar os segmentos que apresentavam *features* corrompidas. Além disso, os scripts MOSES[1] também foram utilizados para

[1] https://github.com/moses-smt/mosesdecoder

pré-processar o conjunto de dados e, em seguida, foi aplicado *byte pair encoding* (BPE) (SENNRICH; HADDOW; BIRCH, 2016) para converter *tokens* em subpalavras.

Após o pré-processamento, os scripts de pré-treinamento, ajuste fino e decodificação do VTLM foram alterados para adaptá-lo ao idioma português[2], tornando Português-Inglês o par de idiomas padrão. Como resultado, foi possível que as *features* da região projetada e os pares de tradução inglês-português fossem processados em um único fluxo pelo VTLM.

Para realizar a extração de *features*, as imagens foram divididas de forma a ser possível realizar este processo em diferentes máquinas, com o intuito de reduzir a duração desta etapa. Assim, utilizou-se a máquina virtual *c2-standard-8* da Google Cloud Platform, com 8 CPUs e 32 GB RAM, e uma máquina do Laboratório LALIC acessada remotamente, que possui uma GPU NVIDIA, 8 CPUs e 16 GB RAM.

## 4.2 Treinamento dos modelos baseline

A próxima atividade deste projeto foi a experimentação com o VTLM para Português-Inglês. Os procedimentos apresentados nesta seção foram descritos e publicados em (SATO; CASELI; SPECIA, 2022). Como forma de comparação, além do modelo VTLM, os experimentos também foram realizados de maneira a verificar a eficácia de outros dois modelos para as tarefas de tradução automática neural (NMT) e MMT, totalizando, assim, três modelos distintos:

- ❑ ***Visual Translation Language Modelling* (VTLM):** o modelo possui uma modalidade visual ao lado dos pares de tradução e processa as *features* de região projetada e os pares de tradução em um único fluxo.

- ❑ ***Translation Language Modelling* (TLM):** o modelo possui uma arquitetura equivalente à arquitetura do VTLM sem as *features* de região.

- ❑ ***Baseline Transformers*:** modelos treinados do zero sem transferir pesos dos modelos TLM ou VTLM pré-treinados. Esses modelos foram treinados apenas no conjunto de dados de tradução automática (MT).

Assim, após as mudanças realizadas para adaptar os modelos a uma nova categoria de corpus (coleção de vídeos com legendas) e a um novo par de idiomas (Português-Inglês), os experimentos foram realizados conforme (CAGLAYAN et al., 2021) e são descritos a seguir para o VTLM (os procedimentos realizados para os outros dois modelos citados acima ocorreram de forma análoga).

[2] https://github.com/LALIC-UFSCar/VTLM_English-Portuguese/tree/master/scripts

### 4.2.1 Pré-treinamento

Para o pré-treinamento, utilizou-se um conjunto do corpus How2 que contém 155 mil *features* e seus textos correspondentes em inglês e português. O pré-treinamento foi realizado por 690 épocas, usando uma única GPU NVIDIA GeForce GTX 1070, e os melhores *checkpoints* foram selecionados em relação à precisão do conjunto de validação.

Semelhante às configurações originais do VTLM, definimos a dimensão do modelo para 512, a dimensão da camada de *feedforward* para 2048, o número de camadas para 6 e o número de *attention heads* para 8. Além disso, os parâmetros do modelo também foram inicializados aleatoriamente e usamos Adam (KINGMA; BA, 2014) com o tamanho do *minibatch* definido como 32 e a taxa de aprendizado definida como 0,0001. A taxa de *dropout* (SRIVASTAVA et al., 2014) foi fixada em 0,1 em todas as camadas.

### 4.2.2 Ajuste fino

O codificador e o decodificador dos modelos MMT e NMT foram inicializados com pesos do VTLM e ajustados com uma taxa de aprendizado menor. Os mesmos hiperparâmetros da fase de pré-treinamento foram usados, exceto o tamanho do *batch* e a taxa de aprendizado, que foram reduzidos para 16 e 1e-5, respectivamente. O ajuste fino foi realizado por 54 épocas para o modelo MMT e 84 épocas para o modelo NMT.

Para avaliação, foram utilizados os modelos com menor perplexidade do conjunto de validação para decodificar traduções com tamanho de *beam* igual a 8.

## 4.3 VTLM com estratégias de *masking* mais informadas

A próxima atividade deste projeto foi a experimentação com novas estratégias para o pré-treinamento do VTLM para Português-Inglês. Assim, nesta seção, são apresentados os experimentos que foram realizados referentes às novas estratégias de *masking* propostas. Os procedimentos apresentados nesta seção foram descritos e publicados em (SATO; CASELI; SPECIA, 2023).

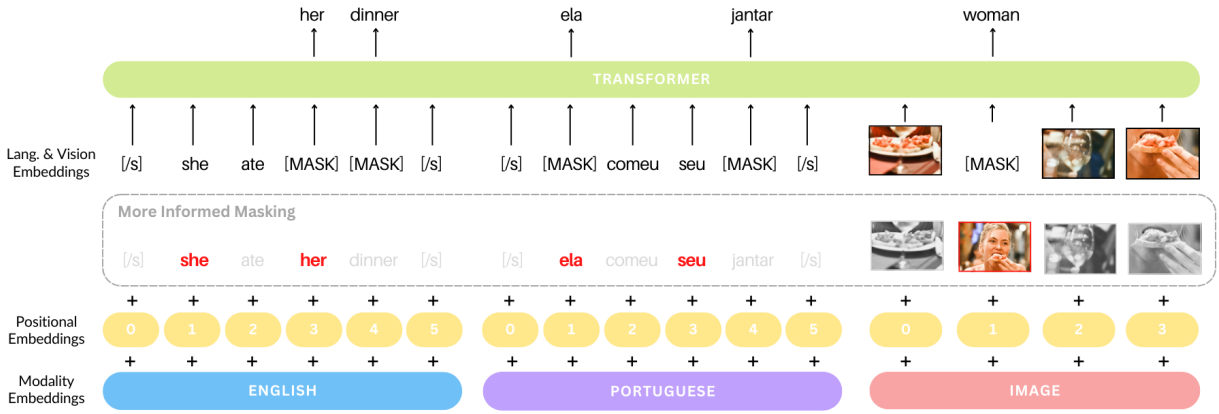
Originalmente, o *masking* no VTLM é realizado de maneira aleatória e se aplica a *tokens* linguísticos e visuais. Para o *masking* textual, 15% das entradas são selecionadas aleatoriamente, sendo 80% substituídas pelo *token* [MASK], 10% substituídas por um *token* aleatório e 10% permanecem inalteradas. Já para o *masking* visual, o VTLM substitui seus vetores de *features* projetadas pelo *token* [MASK], com a mesma proporção de 15%, dos quais 10% se equivalem ao uso de *features* de região selecionadas aleatoriamente de todas as imagens no *batch*, e os 10% restantes das regiões são deixados intactos.

Ao contrário da abordagem original, para as estratégias de *masking* mais informadas propostas neste trabalho, os *tokens* não foram selecionados aleatoriamente para *masking*.

Em vez disso, focou-se em mascarar *tokens* específicos para aprender determinados padrões de linguagem com eficiência. Assim, foram propostas três estratégias de *masking* que exploram formas mais informadas de mascarar *tokens* linguísticos e visuais.

Essas abordagens partem da hipótese de que, ao realizar um mascaramento mais informado – por exemplo, mascarar *tokens* que revelam o gênero das palavras – o modelo poderia chegar a uma melhor compreensão desses conceitos – por exemplo, obtendo melhor desempenho na tradução de pronomes e palavras designadas como masculino, feminino ou neutro. A arquitetura geral do modelo é representada na Figura 6.

Figura 6 – Arquitetura do VTLM, destacando a estratégia de *masking* visual e textual mais informados.




A seguir, são descritas as etapas que foram necessárias para tornar possível a implementação de novas estratégias de *masking* no treinamento do VTLM (4.3.1), assim como os experimentos que envolveram o *masking* visual mais informado (4.3.2), o *masking* textual mais informado (4.3.3) e a combinação de ambos (4.3.4).

### 4.3.1 Adaptação do VTLM para identificar categorias de objetos durante o treinamento

Para selecionar quais elementos serão mascarados durante o pré-treinamento com base em suas categorias, é necessário identificar a categoria de cada elemento na imagem. Assim, o primeiro passo foi investigar e implementar diferentes técnicas de detecção de objetos no VTLM a fim de encontrar a mais adequada e viável computacionalmente para identificar os componentes da imagem durante o treinamento. Isso foi necessário para permitir o desenvolvimento de uma nova estratégia de *masking* visual mais informada, pois a escolha adequada do detector de objetos e sua implementação no VTLM impacta diretamente no desenvolvimento de uma nova estratégia de mascaramento que visa melhorar a eficácia do pré-treinamento.

Após o estudo de diversas ferramentas de detecção de objetos – *Region-based Convolutional Neural Network* (R-CNN) (GIRSHICK et al., 2013), *Spatial Pyramid Pooling* (SPP-net) (HE et al., 2014), Fast R-CNN (GIRSHICK, 2015), Faster R-CNN (REN

et al., 2015), *Single Shot MultiBox Detecto* (SSD) (LIU et al., 2016) e YOLO (REDMON et al., 2015) – as seguintes ferramentas foram escolhidas para experimentos iniciais:

- SSD usando MobileNet v2 (HUANG et al., 2016a) e SSD usando Resnet-101 (HE et al., 2016), para que fosse possível verificar o *tradeoff* entre velocidade e precisão para SSD;
- Faster R-CNN usando Inception Resnet v2 (SZEGEDY et al., 2017), porque Faster R-CNN requer alto poder de computação, então seria interessante reduzir o custo computacional e aumentar sua precisão;
- YOLOv3 (REDMON; FARHADI, 2018), que apresentou melhorias na precisão e velocidade em relação ao YOLO originalmente proposto. Neste caso, a rede para realização da extração de *features* é denominada Darknet-53, que é uma rede proposta pelos autores do YOLOv3 que combina a rede utilizada no YOLOv2, Darknet-19 e a rede residual.

#### 4.3.1.1 Incorporação de detectores de objeto no VTLM

Para analisar a performance de cada ferramenta no VTLM, o modelo foi adaptado para incorporar cada uma das ferramentas de detecção de objetos para que pudessem ser usadas durante o pré-treinamento. Mais especificamente, o *masking* visual foi alterado para realizar a detecção de objetos nas imagens de um *batch* antes de selecionar os *tokens* que seriam mascarados. O objetivo dessa alteração é obter a categoria de cada *token* candidato a ser mascarado antes de selecionar os *tokens* que serão mascarados.

Nesta etapa, o *checkpoint* do modelo de detecção de objetos escolhido é obtido. Em seguida, é criado um índice de categoria a partir do mapa de rótulos do Open Images Dataset e, por fim, as imagens do *batch* são convertidas em *tensors* do tipo *uint8* para que possam ser fornecidos como entrada para cada modelo. A partir disso, o modelo gera previsões sobre *bounding boxes*, rótulos de classe e pontuações de confiança e essas informações (junto com o índice de categoria) são usadas para obter a categoria de cada *token* visual.

Para analisar a eficácia dessa mudança, o *script* de pré-treinamento do VTLM foi executado até a etapa de *masking* visual e a API *TensorFlow Object Detection*[3] foi usada para permitir a visualização de *bounding boxes* e rótulos nas imagens do *batch*. Uma pontuação *threshold* de 0,5 foi escolhida para desenhar apenas as *bounding boxes* que possuem um rótulo com uma pontuação de confiança acima de 50%. Isso possibilitou verificar se a incorporação de cada ferramenta foi bem-sucedida, bem como seu desempenho na identificação de rótulos de classe para cada *token* durante o treinamento. O código-fonte está disponível publicamente[4] e os resultados são apresentados na seção 5.3.1.1.

[3] https://github.com/tensorflow/models/tree/master/research/object_detection
[4] https://github.com/jusato/VTLM_Object_Detection

#### 4.3.1.2 Adaptação do VTLM para identificar as categorias de objetos durante o treinamento usando *features* de objeto

Por fim, foi realizado um último experimento para verificar se seria possível identificar a categoria dos *tokens* visuais durante o treinamento de forma mais eficiente. Para isso, utilizamos *features* de objetos previamente extraídas do corpus How2.

A extração de *features* consiste em transformar dados brutos (imagens, neste caso) em recursos numéricos utilizáveis para aprendizado de máquina, permitindo o armazenamento de informações relevantes da imagem. Com base nisso, usamos as *features* que foram extraídas anteriormente usando o modelo Faster R-CNN pré-treinado no Open Images Dataset V4 para recuperar as informações necessárias para identificar as categorias de *tokens* visuais durante o treinamento.

Diferentemente da implementação anterior, neste caso a detecção de objetos nas imagens não ocorre durante o treinamento. Em vez disso, as informações de detecção necessárias para realizar a identificação dos rótulos de classe são recuperadas de *features* que já foram extraídas pelo modelo de detecção de objetos.

A primeira etapa deste experimento consistiu em analisar como as informações obtidas com a extração de *features* são armazenadas. Após a análise, descobriu-se que as informações relacionadas à detecção – caixas de detecção, classes de detecção, *features* de detecção, pontuações multiclasse de detecção, pontuações de detecção, número de detecções, caixas de detecção brutas e pontuações brutas de detecção – são armazenadas como *tensors* de valores numéricos quando um *batch* de imagens é selecionado durante o treinamento. Com base nisso, o VTLM foi alterado para que essas informações passassem para a etapa de *masking* visual.

O próximo passo foi mudar a forma como o *masking* visual é feito para que no início do processo fosse possível identificar os rótulos de classe de cada objeto da imagem. Assim, foi feita uma alteração para obter o índice de categoria do mapa de rótulos do Open Images Dataset e obter as variáveis contendo as *bounding boxes*, previsões de classes e pontuações de confiança para cada imagem do *batch* por meio das informações que foram armazenadas com o extração de *features*.

Além disso, outra alteração foi feita para identificar o índice associado a cada imagem, bem como a posição de cada *token* visual em relação ao conjunto de imagens do *batch*. Assim, para cada *token* candidato a ser mascarado, foi possível obter seu rótulo de classe e sua respectiva pontuação de confiança. O código-fonte está disponível publicamente[5] e os resultados são apresentados na seção 5.3.1.2.

[5] https://github.com/jusato/VTLM_Object_Detection_Features

### 4.3.2 *Masking* visual mais informado

A proposta de uma nova estratégia de *masking* visual mais informado se tornou possível com os resultados obtidos anteriormente. As mudanças feitas no VTLM para que fosse possível identificar a categoria de *tokens* visuais de forma rápida e eficiente durante o treinamento possibilitaram a escolha dos *tokens* que seriam mascarados com base em sua categoria, permitindo, assim, a realização do *masking* visual mais informado.

A seleção inicial de *tokens* para o *masking* foi alterada para não ser mais realizada de forma aleatória, e selecionar uma maior proporção de *tokens* relacionados a pessoas, como objetos na imagem categorizados como *man*, *woman*, *boy* ou *girl*. Assim, a proporção original do *masking* continuou sendo a mesma, de 15%, mas foram feitos três experimentos de forma a favorecer mais objetos relacionados a pessoas entre os *tokens* que seriam mascarados.

Esta mudança se baseou na hipótese de que ao realizar um mascaramento mais informado, aplicando o *masking* em *tokens* que revelam o suposto gênero das pessoas na imagem, o modelo poderia possivelmente aprender a relacionar melhor as informações textual e visual, obtendo uma melhor performance na tradução de pronomes e palavras no feminino ou masculino.

Esta motivação também surgiu porque estamos analisando a tradução automática multimodal que envolve os idiomas inglês e português, ou seja, idiomas que tratam de forma diferente o gênero das palavras. Enquanto o idioma português possui palavras distintas para se referir a determinados pronomes, isso não acontece no inglês. Por exemplo, o pronome *they* pode ser traduzido tanto para *eles* como para *elas*. Da mesma forma, o idioma português possui muitas palavras com o mesmo significado que se encontram na forma feminina e masculina, o que também não acontece no idioma inglês. Por exemplo, o adjetivo *beautiful* pode ser traduzido para *bonita* ou *bonito*. Portanto, supõe-se que um melhor aprendizado de determinadas informações visuais pode ajudar na tradução de certas palavras.

Assim, no primeiro experimento, a proporção de *tokens* relacionados a pessoas que foram selecionados para o *masking* foi de 5%, com os 10% restantes sendo relacionados a outras categorias e escolhidos de maneira aleatória; no segundo experimento, a proporção foi de 7,5%, ou seja, metade dos *tokens* mascarados; e, no terceiro experimento, a proporção foi de 10%. Esta variação na proporção teve o objetivo de analisar o comportamento do modelo ao favorecer cada vez mais *tokens* relacionados a pessoas, de forma a verificar até que ponto isso poderia beneficiar o modelo.

O código fonte está disponível publicamente[6] e os resultados obtidos são mostrados na seção 5.3.2.

[6] https://github.com/jusato/more_informed_visual_masking

### 4.3.3 *Masking* textual mais informado

A proposta de uma nova estratégia de *masking* textual mais informado seguiu o mesmo caminho da abordagem anterior, isto é, optou-se por mascarar uma maior quantidade de *tokens* que revelam o gênero gramatical das palavras em uma determinada frase.

Da mesma forma, esta mudança se baseou na hipótese de que ao realizar um mascaramento mais informado, aplicando o *masking* em *tokens* que revelam o gênero gramatical das palavras, o modelo poderia possivelmente aprender melhor estes conceitos, obtendo uma melhor performance na tradução de pronomes e palavras no feminino ou masculino.

Assim, a seleção inicial de *tokens* para o *masking* foi alterada para não ser mais realizada de forma aleatória, e favorecer mais pronomes, como ele/ela, dele/dela e seu/sua, entre os *tokens* que seriam mascarados, mantendo a proporção de 15%.

Para isso, primeiro foi necessário encontrar uma maneira de identificar os pronomes a partir da entrada recebida. O VTLM armazena o fluxo textual de entrada na forma de *Tensors*[7] do tipo inteiro, ou seja, o armazenamento do texto ocorre de forma numérica. Assim, a arquitetura do VTLM foi alterada para fazer a conversão deste fluxo inicial de números para palavras e, em seguida, determinar as frases do conjunto de dados obtido para poder identificar os pronomes pessoais, possessivos e demonstrativos.

Após esta conversão, é feita a identificação dos pronomes em cada frase utilizando o casamento direto entre palavras (isto é, igualdade com pronomes como “ele”, “ela”, “they”), e as posições do pronomes são armazenadas em um *tensor* para permitir que eles sejam associados a sua forma numérica original e, posteriormente, identificados na seleção de palavras para o *masking*.

Por fim, durante o *masking*, é feita a identificação dos elementos da entrada referentes a pronomes e os *tokens* são escolhidos de maneira seletiva para serem mascarados, com uma maior proporção de *tokens* referentes a pronomes sendo mascarados.

Foram feitos três experimentos. No primeiro, a proporção de *tokens* referentes a pronomes que foram selecionados para o *masking* foi de 5%, com os 10% restantes sendo relacionados a outras categorias e escolhidos de maneira aleatória; no segundo experimento, a proporção foi de 7,5%; e, no terceiro experimento, a proporção foi de 10%. Da mesma maneira, esta variação na proporção teve o objetivo de analisar o comportamento do modelo ao favorecer cada vez mais *tokens* referentes a pronomes, de forma a verificar até que ponto isso poderia beneficiar o modelo.

O código fonte está disponível publicamente[8] e os resultados obtidos são mostrados na seção 5.3.3.

---

[7] *Tensor* é semelhante a um *array numpy* com a diferença de que pode rodar operações tanto em CPU como em GPU. Assim, é apenas um *array* n-dimensional que pode ser usado para computações numéricas.

[8] https://github.com/jusato/more_informed_textual_masking

### 4.3.4 *Masking* visual e textual mais informados

A proposta de uma nova estratégia de *masking* visual e textual mais informados se tratou de uma combinação das duas abordagens anteriores, isto é, mascarou-se uma maior quantidade de *tokens* relacionados a pessoas, como objetos na imagem categorizados como *man*, *woman*, *boy* ou *girl*, assim como *tokens* que revelam o gênero gramatical das palavras em uma determinada frase.

Esta abordagem teve como objetivo verificar o comportamento do modelo ao aplicar em conjunto o *masking* visual mais informado e o *masking* textual mais informado, baseando-se na hipótese de que uma melhor performance poderia ser obtida ao combinar as duas estratégias concomitantemente.

O código fonte está disponível publicamente[9] e os resultados obtidos são mostrados na seção 5.3.4.

## 4.4 Exploração do *masking* com diferentes categorias semânticas

Nesta seção, abordaremos a exploração do *masking* com outras categorias semânticas, além da categoria "pessoas", para analisar a performance na tradução automática multimodal. Ao explorar o *masking* em categorias adicionais, nosso objetivo é analisar como o modelo de tradução automática multimodal se comporta ao processar imagens e textos que envolvem outros objetos e conceitos específicos. A variação das categorias mascaradas proporcionará um entendimento sobre a capacidade do modelo em lidar com diferentes tipos de informações semânticas e sua habilidade para realizar traduções precisas e contextualmente coerentes.

### 4.4.1 Estudo do corpus

Nesta seção, é descrito o estudo realizado no conjunto de treinamento do corpus How2. O objetivo principal foi aprofundar a compreensão dos termos e classes mais frequentes presentes no corpus, com o propósito de orientar os experimentos relacionados à exploração de técnicas de *masking* com diferentes categorias semânticas. Para obter percepções significativas sobre o corpus, realizamos uma exploração detalhada dos dados, concentrando-nos na exploração da parte visual e textual separadamente.

#### 4.4.1.1 *Frames* de vídeo

A exploração da parte visual iniciou-se com a contagem das classes de objetos nos *frames* de vídeo. Como sabemos, os *frames* de vídeo são anotados com caixas delimitadoras

[9] https://github.com/jusato/more_informed_visual_textual_masking

(*bounding boxes*) que identificam objetos específicos presentes nas cenas. Assim, nesta etapa, nosso objetivo foi identificar as classes de objetos mais frequentes presentes nos frames. Implementamos um procedimento que contou o número de ocorrências de cada classe de objeto presente nas anotações das *bounding boxes*. Isso nos permitiu determinar quais categorias de objetos são mais representativas e comuns no conjunto de dados.

Em seguida, foi realizada uma separação hierárquica das classes de objeto, com o intuito de investigar diferentes níveis de granularidade na categorização. Por exemplo, se um objeto for rotulado como "infant bed", ele também faz parte da classe "bed", que também faz parte da classe "furniture". Assim, identificar a estrutura hierárquica das classes permite a organização dos objetos em níveis mais gerais e específicos. Essa abordagem é valiosa para entender a distribuição dos objetos em diferentes níveis de detalhes e pode orientar a implementação de estratégias de *masking* mais informadas.

#### 4.4.1.2 Texto

Nesta seção, prosseguimos com um estudo detalhado do corpus How2, focando agora na análise dos textos presentes no conjunto de dados.

Para obter uma compreensão mais detalhada do texto, realizamos uma análise que contou o número total de palavras e frases presentes no conjunto de dados. Utilizamos um conjunto de técnicas de pré-processamento utilizando a biblioteca *spacy* para tratar os textos inglês e português, como transformar todas as palavras em letras minúsculas e remover pontuações, garantindo assim uma contagem precisa.

Dentro da análise de palavras, concentramos nossos esforços em contabilizar o número de palavras relacionadas a seres humanos. Definimos uma função que identifica palavras específicas, como "person", "man", "woman", "child", entre outras, que se referem a seres humanos ou a pronomes pessoais.

Com a contagem de palavras, descobrimos o tamanho do vocabulário presente no conjunto, que é importante para entender a diversidade lexical do texto e pode ser útil para futuras análises linguísticas. A contagem de frases, por sua vez, nos permitiu entender o número total de unidades de texto no corpus. Isso é relevante para dimensionar o tamanho geral do conjunto de dados e para a criação de amostras representativas para análises mais aprofundadas.

### 4.4.2 Aumento do *threshold*

Anteriormente, ao identificar as classes dos objetos nos *frames* de vídeo durante o treinamento, utilizava-se um *threshold* de 40% para considerar uma detecção válida de uma determinada classe. Isso significa que apenas as detecções com uma precisão igual ou superior a 40% eram consideradas como representantes da classe em questão. No

entanto, para esta nova análise, optou-se por aumentar o *threshold* para 60% na etapa de identificação de cada classe durante o treinamento do modelo.

O objetivo principal é mascarar corretamente a categoria de interesse ao realizar o *masking* visual mais informado para essa categoria específica. Ao aumentar o *threshold*, priorizamos a detecção de objetos com maior confiança, garantindo que apenas as detecções mais precisas e confiáveis sejam consideradas para o processo de *masking*. Dessa forma, o objetivo foi reduzir a presença de falsos positivos na identificação das classes de objetos, evitando mascarar incorretamente regiões que não correspondem à categoria em análise, o que poderia ser um ruído para o treinamento do modelo.

### 4.4.3 Experimentação com classes mais frequentes

Através da contagem das classes de objetos nas imagens, identificamos as categorias mais prevalentes no conjunto de dados e, ao realizar a separação hierárquica das classes de objetos, pudemos entender como as categorias são organizadas em diferentes níveis de granularidade.

A análise detalhada das classes de objetos mais frequentes resultou na identificação das seguintes categorias predominantes:

1. Person
2. Clothing
3. Furniture
4. Mammal
5. Plant

Essas categorias destacam-se pela sua frequência e representatividade no conjunto de dados. No entanto, devido ao tempo limitado e à demanda computacional necessária para treinar os modelos, não foi possível realizar experimentos detalhados com todas as categorias identificadas. Assim, dada a relevância das categorias “person”, “clothing” e “furniture”, decidimos realizar uma exploração mais aprofundada do *masking*, variando a proporção dessas categorias nos experimentos.

Para realizar a exploração mais aprofundada do *masking*, adotamos uma abordagem seletiva, mantendo a proporção global de 15% de *tokens* mascarados, mas variando as proporções de *tokens* relacionados às categorias “person”, “clothing” e “furniture”. As proporções selecionadas foram de 5%, 7,5% e 10% para cada uma dessas categorias.

Essa abordagem nos permitiu investigar o impacto de diferentes níveis de *masking* em *tokens* específicos, proporcionando uma análise mais refinada e detalhada sobre o comportamento do modelo em relação a essas categorias semânticas-chave. Ao variar as

proporções de *masking* para *tokens* categorizados como “person”, “clothing” ou “furniture”, buscamos entender como o modelo reage a diferentes níveis de informação dessas categorias durante o processo de treinamento.

# Capítulo 5

# Avaliação

Esta seção é destinada à avaliação dos experimentos realizados. Inicialmente, são descritas as medidas de avaliação que foram utilizadas (5.1) e, em seguida, a avaliação dos experimentos realizados: VTLM adaptado para legendas de vídeo e novo par de línguas (5.2), VTLM com estratégias de *masking* mais informadas (5.3) e Exploração do *masking* com diferentes categorias semânticas (5.4).

## 5.1 Medidas de avaliação

Os resultados obtidos com as experimentações descritas no capítulo anterior foram avaliados através de duas métricas: BLEU (PAPINENI et al., 2002) e METEOR (BANERJEE; LAVIE, 2005).

BLEU é uma métrica de avaliação automática da tarefa de tradução de uma língua natural para outra automaticamente – seu cálculo é baseado no número de correspondências entre o texto gerado e os textos de referência usando o método de n-gramas e possui uma pontuação que varia entre 0 e 1 (1 é a pontuação máxima). Assim, trata-se de uma medida para avaliar a qualidade do texto traduzido automaticamente com base na intersecção de seus n-gramas com os de uma ou mais traduções de referência.

Já METEOR, além de ser uma métrica de avaliação automática, é a métrica oficial de tradução automática multimodal – foi a métrica utilizada na competição WMT (2016-2018) para tarefa de tradução automática multimodal. Seu cálculo também é realizado a partir do alinhamento entre as hipóteses geradas e os textos de referência, mas o BLEU foca principalmente na sobreposição de palavras ou frases, enquanto o METEOR considera uma gama mais ampla de fatores linguísticos, incluindo semelhança de palavras, ordem das palavras e aspectos gramaticais.

## 5.2 VTLM adaptado para legendas de vídeo e novo par de línguas

Esta seção está relacionada ao experimento descrito na Seção 4.2, que foi realizado com o intuito de testar a capacidade de generalização do *Visual Translation Language Modelling* para outros idiomas e corpora.

Como apresentado anteriormente, além de resultar em uma mudança de pares de idiomas (Português-Inglês), a utilização do corpus How2 resultou também na introdução de um cenário mais desafiador, em que o texto associado a cada imagem/*frame* de vídeo não é apenas uma descrição desta, e sim uma legenda que pode não estar relacionada ao seu *frame* correspondente. Como consequência, a tradução automática multimodal usando o How2 é consideravelmente mais difícil em comparação à tradução automática multimodal usando o Multi30k (ELLIOTT et al., 2016), corpus originalmente utilizado no VTLM.

### 5.2.1 Resultados

Os modelos treinados foram avaliados para as tarefas de MMT e NMT. A Tabela 1 mostra as pontuações BLEU e METEOR em conjuntos de validação e teste do How2.

Tabela 1 – Pontuações BLEU e METEOR para os *baselines Transformers* (apenas texto), TLM pré-treinado e ajustado usando o How2 (apenas texto) e VTLM pré-treinado e ajustado usando o How2 (texto e imagem) para as tarefas NMT e MMT.

| | | **Teste** | | **Validação** | |
|---|---|---|---|---|---|
| | | BLEU | METEOR | BLEU | METEOR |
| *Baselines* | MMT | 37,57 | 63,51 | 38,34 | 63,60 |
| | NMT | 43,58 | 70,62 | 43,28 | 70,18 |
| TLM | MMT | **51,99** | 77,52 | 52,19 | 77,87 |
| | NMT | 50,61 | 77,67 | 50,72 | 78,01 |
| VTLM | MMT | 51,80 | **78,04** | **52,44** | **78,25** |
| | NMT | **52,20** | **78,20** | **52,81** | **78,70** |

Semelhante à proposta original (CAGLAYAN et al., 2021), os resultados mostram o impacto do pré-treinamento visual multilíngue no desempenho final, pois o modelo MMT supera o *baseline* MMT em aproximadamente 14 pontos BLEU e 14 pontos METEOR quando ajustado para tradução automática multimodal.

Além disso, para os modelos treinados do zero (*Transformers*), o MMT é inferior ao NMT em cerca de 5 pontos BLEU e 7 pontos METEOR, mas quando os *checkpoints* TLM ou VTLM pré-treinados são ajustados para tradução automática, a diferença entre a pontuação dos modelos MMT e NMT diminui ou se torna inexistente.

Por fim, comparado aos melhores resultados obtidos em (CAGLAYAN et al., 2021) com o VTLM Alemão-Inglês – 44,0 BLEU e 61,3 METEOR no conjunto de teste Multi30k para a tarefa de tradução automática multimodal – o VTLM Português-Inglês obteve pontuações mais altas – 51,8 BLEU e 78,04 METEOR no conjunto de teste How2 para a tarefa MMT. No entanto, é importante salientar que uma comparação direta não é possível devido às diferenças de linguagem e corpus.

### 5.2.2 Análise qualitativa

Na Tabela 2 são apresentados alguns exemplos de textos traduzidos por cada modelo. No primeiro exemplo, o *baseline* MMT erra a tradução das palavras fonte "consultoria" e "Coral Gables", enquanto tanto o VTLM quanto o TLM as traduzem corretamente, obtendo um melhor desempenho (cerca de 80 pontos BLEU acima do *baseline*). Isso indica a eficácia do pré-treinamento no desempenho dos modelos.

No segundo exemplo, a diferença entre as pontuações de VTLM e TLM é maior. VTLM atinge uma pontuação BLEU de 100,0 e TLM obtém uma pontuação BLEU de 37,8, principalmente devido à tradução incorreta das palavras "brown" e "whole". Portanto, observamos que as *features* visuais de região podem ajudar o modelo a entender o contexto, resultando em uma tradução mais precisa.

No terceiro caso, a imagem possui objetos extras desassociados da frase, como a mão e outros componentes da tela. Como resultado, a imagem pode trazer informações irrelevantes ao texto, o que pode gerar ruídos e afetar a qualidade da tradução. A tradução do modelo VTLM ilustra essa possível desvantagem, pois o modelo apresenta desempenho inferior ao modelo TLM em cerca de 19 pontos BLEU.

Além disso, no quarto exemplo, nenhum dos modelos traduziu com precisão o texto-fonte. A razão para isso é que a palavra "grooming" na frase de referência aparece apenas algumas vezes no conjunto de treinamento (25 vezes em um conjunto de 3.304.534 *tokens*). No entanto, há uma grande diferença entre as traduções dos três modelos. Por exemplo, a tradução do modelo TLM mostra um grau maior de imprecisão em comparação à tradução do modelo VTLM, e o resultado do *baseline* é "What kind of treatment our horse likes this horse." ("Que tipo de tratamento nosso cavalo gosta deste cavalo."), que está ainda mais distante da resposta correta.

Tabela 2 – Exemplos de tradução de diferentes modelos MMT: *baseline Transformer* MMT, TLM e VTLM.

| | | |
|---|---|---|
|  | **Fonte:** Consultoria de imagem e etiqueta em Coral Gables, Flórida.<br>**Referência:** Image and Etiquette Consulting in Coral Gables, Florida. | |
| | ***Baseline*:** Image and Etiquette Etiquette ette: Florida, Florida. | 17,57 BLEU |
| | **TLM:** Image and Etiquette Consulting in Coral Gables, Florida. | 100,0 BLEU |
| | **VTLM:** Image and Etiquette Consulting in Coral Gables, Florida. | 100,0 BLEU |
|  | **Fonte:** E então algo como arroz integral ou pão de trigo integral.<br>**Referência:** And then something like brown rice or whole wheat bread. | |
| | ***Baseline*:** And then something like full rice or full trigger. | 29,38 BLEU |
| | **TLM:** And then something like full rice or full wheat bread. | 37,82 BLEU |
| | **VTLM:** And then something like brown rice or whole wheat bread. | 100,0 BLEU |
|  | **Fonte:** E você pode usar quantas dessas faixas de bateria dentro da sua janela de arranjos lógicos.<br>**Referência:** And you can use as many of these drum tracks within your logic arrange window. | |
| | ***Baseline*:** And you can use how many of these drum tracks inside your clock window. | 42,83 BLEU |
| | **TLM:** And you can use as many of these drum tracks within your logic arrange window. | 100,0 BLEU |
| | **VTLM:** And you can use as many of these drum tracks within your logic plugs. | 81,07 BLEU |
|  | **Fonte:** Que tipo de tratamento o vosso cavalo gosta.<br>**Referência:** The kind of grooming that your horse likes. | |
| | ***Baseline*:** What kind of treatment our horse likes this horse. | 06,67 BLEU |
| | **TLM:** What kind of treatment our horse likes. | 15,25 BLEU |
| | **VTLM:** What kind of treatment your horse likes. | 36,28 BLEU |

## 5.3 VTLM com estratégias de *masking* mais informadas

### 5.3.1 Adaptação do VTLM para identificar categorias de objetos durante o treinamento

Esta seção descreve os resultados obtidos com a adaptação do VTLM para identificar categorias de objetos durante o treinamento.

#### 5.3.1.1 Incorporação de detectores de objeto no VTLM

O desempenho das ferramentas foi analisado usando o corpus How2, uma vez que a análise foi feita durante o pré-treinamento do VTLM En-Pt. Assim, a análise dos resultados baseou-se na correta identificação e categorização dos elementos visuais presentes em cada *frame* de vídeo. Os resultados foram analisados comparando as categorias atribuídas a cada elemento visual com as categorias esperadas. Além disso, também foi levado em consideração o tempo gasto por cada ferramenta para realizar a previsão.

Os resultados mostraram que a categorização dos componentes visuais durante o treinamento ocorre corretamente para todas as ferramentas de detecção de objetos, indicando eficácia na alteração proposta.

Além disso, os resultados obtidos com cada ferramenta foram comparados em termos de velocidade e eficácia na identificação da categoria de todos os elementos nas imagens durante o treinamento, a fim de verificar qual ferramenta resultou em melhor desempenho para o VTLM. Os resultados mostraram que Faster R-CNN é muito mais eficaz em detectar uma grande quantidade de objetos em *frames* de vídeo do que as outras ferramentas, mostrando maior adequação para incorporação no VTLM para permitir que ele identifique a categoria de *tokens* visuais durante o treinamento para realizar um *masking* visual mais informado.

Entretanto, apesar do bom desempenho do Faster R-CNN na identificação de *tokens* visuais durante o treinamento, ele possui uma velocidade menor em comparação com os outros detectores. Por outro lado, SSD com MobileNet, SSD com Resnet e YOLO possuem alta velocidade, o que é um fator crucial para a identificação de componentes na imagem durante o treinamento.

#### 5.3.1.2 Adaptação do VTLM para identificar as categorias de objetos durante o treinamento usando features de objeto

Os resultados mostraram que a categorização dos componentes visuais ocorre de forma correta durante o treinamento, indicando efetividade na mudança proposta. Além disso,

foi observado que a eficácia na categorização de *tokens* visuais é a mesma em comparação com o experimento anterior usando a mesma ferramenta de detecção, como esperado.

No entando, apesar da eficácia na categorização de *tokens* visuais ser a mesma, a análise do tempo mostrou que usar as informações obtidas com extração de *features* resulta em uma detecção consideravelmente mais rápida (cerca de 17x mais rápida para o Faster R-CNN).

Portanto, a análise mostrou que esta última implementação leva ao melhor desempenho, pois permite identificar a categoria de *tokens* visuais de forma rápida e eficiente durante o treinamento, fatores cruciais para permitir o desenvolvimento de uma nova estratégia de *masking* visual mais informada.

Como resultado, tornou-se possível identificar a categoria de cada componente na imagem durante o treinamento do VTLM a fim de usá-lo para escolher seletivamente os *tokens* visuais que serão mascarados.

### 5.3.2 *Masking* visual mais informado

Esta seção está relacionada ao experimento descrito na Seção 4.3.2, que foi realizado com o intuito de explorar maneiras mais informadas de realizar o *masking* visual visando melhorar a eficácia do treinamento e, consequentemente, melhorar a performance final do modelo.

#### 5.3.2.1 Resultados

Os modelos treinados foram avaliados para a tarefa de tradução automática multimodal. A Tabela 3 mostra as pontuações BLEU e METEOR em conjuntos de validação e teste do How2.

Tabela 3 – Pontuações BLEU e METEOR para o VTLM pré-treinado e ajustado para a tarefa de MMT, mantendo a estratégia original de *masking* aleatório e utilizando a nova estratégia de *masking*.

| | | **Teste** | | **Validação** | |
|---|---|---|---|---|---|
| | | BLEU | METEOR | BLEU | METEOR |
| VTLM: *masking* aleatório | | 51,80 | 78,04 | 52,44 | 78,25 |
| VTLM: *masking* visual informado | 5% | **52,70** | **79,63** | **53,25** | **79,83** |
| | 7,5% | 51,92 | 79,10 | 52,51 | 79,41 |
| | 10% | 51,65 | 78,64 | 52,26 | 79,09 |

Os resultados obtidos mostram que o *masking* visual mais informado afeta o desempenho final do modelo. No primeiro experimento, a proporção de *tokens* relacionados a pessoas que foram selecionados para o *masking* foi de 5%, com os 10% restantes sendo relacionados a outras categorias e escolhidos de maneira aleatória. Neste caso, o VTLM de *masking* visual mais informado obteve 52,70 BLEU no conjunto de teste e 53,25 BLEU

no conjunto de validação, superando o VTLM de *masking* aleatório em aproximadamente 1 BLEU.

Já no segundo experimento, em que a proporção foi de 7,5%, ou seja, metade dos *tokens* mascarados, o VTLM de *masking* visual mais informado também superou o *baseline*, mas sua performance foi pior em relação ao primeiro experimento, obtendo 51,92 BLEU no conjunto de teste e 52,51 BLEU no conjunto de validação.

Por fim, no último experimento, a proporção de *tokens* relacionados a pessoas que foram selecionados para o *masking* foi de 10%, ou seja, em torno de 66,67% do total de *tokens* que foram mascarados. Neste caso, a performance do modelo foi inferior ao *baseline* em aproximadamente 0,20 BLEU, comportamento diferente do observado nos últimos dois experimentos.

Dessa forma, os resultados indicam que o *masking* visual mais informado beneficia o desempenho final do modelo até um certo ponto. Ao aumentar a proporção de *tokens* relacionados a pessoas sendo selecionados para o *masking*, observa-se uma melhoria no desempenho do modelo em relação ao *baseline*. No entanto, quando esta proporção se torna superior à metade do total de *tokens* sendo mascarados, esta melhoria na performance tende a diminuir.

Uma possível causa deste comportamento pode estar relacionada à diminuição do mascaramento de *tokens* relacionados a outras categorias, já que a proporção total de *tokens* selecionados para o *masking* não foi alterada, ou seja, continuou em 15%. Assim, aumentar a proporção de *tokens* relacionados a pessoas sendo mascarados significa diminuir a proporção de *tokens* relacionados a outras categorias sendo mascarados, o que pode prejudicar o aprendizado de elementos de outras categorias.

#### 5.3.2.2 Análise qualitativa

A seguir, são apresentados alguns exemplos de textos traduzidos pelo VTLM de *masking* aleatório (*baseline*) e pelo VTLM de *masking* visual mais informado, junto a suas referências e *frames* de vídeo correspondentes. Os exemplos mostrados ilustram situações que foram observadas nos resultados obtidos.

No primeiro exemplo, o VTLM de *masking* aleatório erra a tradução dos pronomes pessoais do caso reto “he” e “she”, traduzindo ambos para “it”, enquanto o VTLM de *masking* visual mais informado os traduz corretamente, obtendo um melhor desempenho.

**Referência:** So **he** or **she** is not carrying all the weight of the scba, in the shoulder area, or region around the neck.

***Baseline*:** So **it** or **it** doesn’t carry all the weight of the scba, in the shoulder area, or region around the neck.
**VTLM:** So **he** or **she** won’t carry all the weight of the scba, in the shoulder area, or region around the neck.

No segundo exemplo, o VTLM de *masking* aleatório erra a tradução do pronome pessoal do caso oblíquo “him” e o traduz erroneamente para “it”, enquanto o VTLM de *masking* visual mais informado o traduz corretamente.


**Referência:** So there’s a couple of different ways to take **him** out.

---

***Baseline*:** So there’s a couple of different ways to take **it** out.
**VTLM:** So there’s a couple of different ways to get **him** out.

O terceiro exemplo ilustra a tradução correta do pronome possessivo “her” (“dela”, em português) pelo VTLM de *masking* visual mais informado, enquanto o *baseline* o traduz erroneamente para “your” (“seu”, em português).

**Referência:** And we’re going to be cornrowing that into **her** hair today.

---

***Baseline*:** And we’re going to do that on **your** hair today.
**VTLM:** And we’re going to do that on **her** hair today.

Por fim, o VTLM de *masking* aleatório faz referência a um objeto utilizando o pronome pessoal “she” (“ela”, em português) ao invés do pronome “it”, que é um pronome usado na língua inglesa para se referir a algo que não seja humano, como um objeto, um animal, uma ação ou um sentimento. Em contraposição, o VTLM de *masking* visual mais informado não comete o mesmo erro e utiliza o pronome de forma correta.

**Referência:** **It** will catch snow and push it over to the side of the road or **it** will catch dirt out of a high spot and move it over to the side.

---

***Baseline*: She** will take snow and push it to the side of the road or **she** will take the dirt from a high point and move it to the side.
**VTLM: It** will take snow and push it to the side of the road or **it** will take the dirt from a high spot and move it to the side.

Dessa forma, as traduções indicam que ao realizar um mascaramento mais informado, aplicando o *masking* em *tokens* que se referem a pessoas nas imagens, o modelo aprende a relacionar melhor as informações textual e visual, obtendo uma melhor performance na tradução de pronomes pessoais, possessivos e outros. Como resultado, o *masking* visual mais informado auxilia na tradução de textos mais coerentes, em que é possível utilizar pronomes de forma correta para substituir substantivos, objetos ou frases nominais.

### 5.3.3 *Masking* textual mais informado

Esta seção está relacionada ao experimento descrito na Seção 4.3.3, que foi realizado com o intuito de explorar maneiras mais informadas de realizar o *masking* textual visando

melhorar a eficácia do treinamento e, consequentemente, melhorar a performance final do modelo.

#### 5.3.3.1 Resultados

Os modelos treinados foram avaliados para a tarefa de tradução automática multimodal. A Tabela 4 mostra as pontuações BLEU e METEOR em conjuntos de validação e teste do How2.

Tabela 4 – Pontuações BLEU e METEOR para o VTLM pré-treinado e ajustado para a tarefa de MMT, mantendo a estratégia original de *masking* aleatório e utilizando a nova estratégia de *masking*.

| | | **Teste** | | **Validação** | |
|---|---|---|---|---|---|
| | | BLEU | METEOR | BLEU | METEOR |
| VTLM: *masking* aleatório | | 51,80 | 78,04 | 52,44 | 78,25 |
| VTLM: *masking* textual informado | 5% | **52,64** | **79,45** | **52,96** | **79,53** |
| | 7,5% | 52,39 | 79,35 | 52,94 | 79,51 |
| | 10% | 52,21 | 79,27 | 52,82 | 79,42 |

Os resultados obtidos mostram que o *masking* textual mais informado afeta o desempenho final do modelo. No primeiro experimento, a proporção de *tokens* referentes a pronomes que foram selecionados para o *masking* foi de 5%. Neste caso, o VTLM de *masking* textual mais informado obteve 52,64 BLEU no conjunto de teste e 52,96 BLEU no conjunto de validação, superando o desempenho do VTLM de *masking* aleatório.

Já no segundo experimento, em que a proporção foi de 7,5%, ou seja, metade dos *tokens* mascarados, o VTLM de *masking* textual mais informado também superou o *baseline*, mas sua performance foi pior em relação ao primeiro experimento, obtendo 52,39 BLEU no conjunto de teste e 52,94 BLEU no conjunto de validação.

Por fim, no último experimento, a proporção de *tokens* referentes a pronomes que foram selecionados para o *masking* foi de 10%, ou seja, em torno de 66,67% do total de *tokens* que foram mascarados. Neste caso, obteve-se 52,21 BLEU no conjunto de teste e 52,82 BLEU no conjunto de validação, ou seja, a performance do modelo também foi superior ao *baseline*. No entanto, o desempenho foi inferior aos últimos dois experimentos, em que as proporções escolhidas foram de 5% e 7,5%.

Dessa forma, os resultados indicam que ao mascarar uma maior quantidade de *tokens* que revelam o gênero gramatical das palavras em uma determinada frase – neste caso, pronomes pessoais, possessivos e demonstrativos – observa-se uma melhoria no desempenho final do modelo. Apesar do VTLM de *masking* textual mais informado ter superado o *baseline* em todos os experimentos, esta melhoria de desempenho é limitada, já que o melhor desempenho foi observado com a proporção de 5%, seguido de 7,5% e 10%, respectivamente.

Este comportamento também foi observado ao aplicar o *masking* visual mais informado e, da mesma forma, pode estar relacionado à diminuição do mascaramento de *tokens* relacionados a outras categorias, pois aumentar a proporção de *tokens* referentes a pronomes sendo mascarados resulta na diminuição da proporção de *tokens* relacionados a outras categorias sendo mascarados, o que pode prejudicar o aprendizado de elementos de outras categorias.

#### 5.3.3.2 Análise qualitativa

A seguir, são apresentados alguns exemplos de textos traduzidos pelo VTLM de *masking* aleatório (*baseline*) e pelo VTLM de *masking* textual mais informado, junto a suas referências e *frames* de vídeo correspondentes. Estes exemplos ilustram situações que foram observadas nos resultados obtidos.

No primeiro exemplo, o VTLM de *masking* aleatório utiliza os pronomes "he" e "him" para se referir à palavra "dog" ("cachorro", em português) ao invés de utilizar o pronome "it", que deveria ter sido usado neste caso por se tratar de um animal. No entanto, o VTLM de *masking* textual mais informado não comete o mesmo erro e utiliza o pronome correto em todos os casos, obtendo um melhor desempenho na tradução.

**Referência:** If you walk your dog on your left side you want **it** to sit on the side because what **it** does is tighten up so if you're over here the dog should have it over here.

---

***Baseline*:** If you walk your dog on your left side you want **him** to sit on the side because what **he** does is squeeze, then if you're standing over here the dog should have him here.
**VTLM:** If you walk your dog on your left side you want **it** to sit on the side because what **it** does is tighten, then if you're over here the dog should have it here.

No segundo exemplo, o VTLM de *masking* aleatório erra a tradução do pronome pessoal do caso reto "she" e o traduz erroneamente para "it", o que se trata de um erro grave na tradução Português-Inglês, pois o pronome "it" não pode ser usado para se referir a uma pessoa. Em contraposição, o VTLM de *masking* textual mais informado utiliza o pronome correto ("she") e obtém uma melhor performance.


**Referência:** **She** walks into the scene after the scene begins between the police officer and Stanley.

---

***Baseline*:** **It** goes into scene after the scene starts between the police officer and Stanley.
**VTLM:** **She** goes into scene after the scene starts between the police officer and Stanley.

No terceiro caso, observa-se a tradução correta do pronome pessoal do caso oblíquo "him" pelo VTLM de *masking* textual mais informado, enquanto o *baseline* o traduz erroneamente para "it".

**Referência:** So we started talking about it and I said to **him**, well you know there are certain things that are required to have a really a healthy group of employees.

---

***Baseline*:** So we start talking about that and I said to **it**, and you know there are certain things that are needed to have a really healthy working group.
**VTLM:** So we started talking about that and I said to **him**, and you know there are certain things that are necessary to have a really healthy working group.

O próximo exemplo ilustra a tradução incorreta do pronome pessoal do caso oblíquo "her" pelo VTLM de *masking* aleatório, que novamente utiliza o pronome "it" para se referir a uma pessoa. Porém, este erro não é cometido pelo VTLM de *masking* textual mais informado, que faz o uso correto do pronome na tradução.

**Referência:** And I only worked one night with **her**.

---

***Baseline*:** And I just worked a night with **it**.
**VTLM:** And I just worked a night with **her**.

Os quatro exemplos anteriores ilustram situações semelhantes às observadas com a aplicação do *masking* visual mais informado. No entanto, o último exemplo mostra uma nova melhoria na tradução. Esta melhoria está relacionada ao uso do pronome "it" como objeto direto de um verbo. Enquanto o *baseline* omite este pronome na tradução, o VTLM de *masking* textual mais informado o utiliza corretamente após o verbo "try".


**Referência:** But, I'm going to try **it** anyway and you can get an idea of what you might want to do.

---

***Baseline*:** But, I'm going to try anyway and you might have an idea of what you might want to do.
**VTLM:** But, I'm going to try **it** anyway and you might get an idea of what you might want to do.

Portanto, as traduções indicam que ao realizar um mascaramento mais informado, aplicando o *masking* em *tokens* que se referem a pronomes nas frases, o modelo aprende melhor estes conceitos e obtém uma melhor performance na tradução. Assim, a aplicação do *masking* textual mais informado também contribui para a tradução de textos mais coerentes.

### 5.3.4 *Masking* visual e textual mais informados

Esta seção está relacionada ao experimento descrito na Seção 4.3.4, que foi realizado com o intuito de explorar maneiras mais informadas de realizar o *masking* visual e textual

visando melhorar a eficácia do treinamento e, consequentemente, melhorar a performance final do modelo.

#### 5.3.4.1 Resultados

Os modelos treinados foram avaliados para a tarefa de tradução automática multimodal. A Tabela 8 mostra as pontuações BLEU e METEOR em conjuntos de validação e teste do How2.

Tabela 5 – Pontuações BLEU e METEOR para o VTLM pré-treinado e ajustado para a tarefa de MMT, mantendo a estratégia original de *masking* aleatório e utilizando a nova estratégia de *masking*.

| | **Teste** | | **Validação** | |
|---|---|---|---|---|
| | BLEU | METEOR | BLEU | METEOR |
| VTLM: *masking* aleatório | 51,80 | 78,04 | 52,44 | 78,25 |
| VTLM: *masking* visual e textual informados | **52,34** | **78,77** | **53,28** | **79,44** |

Os resultados obtidos mostram que esta nova estratégia de *masking* também afeta o desempenho final do modelo. O VTLM de *masking* aleatório obteve 51,80 BLEU no conjunto de teste e 52,44 BLEU no conjunto de validação, enquanto o VTLM de *masking* visual e textual mais informados obteve 52,34 BLEU no conjunto de teste e 53,28 BLEU no conjunto de validação, superando a performance do *baseline*.

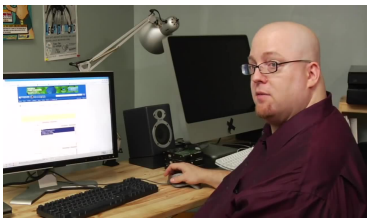
Dessa forma, apesar da melhoria do desempenho não ter sido muito alta, os resultados indicam que a aplicação do *masking* visual e textual mais informados beneficia o desempenho final do modelo. Assim, ao mascarar uma maior quantidade de *tokens* relacionados a pessoas, como objetos na imagem categorizados como *man*, *woman*, *boy* ou *girl*, assim como *tokens* que revelam o gênero gramatical das palavras em uma determinada frase, observa-se uma melhoria no desempenho do modelo em relação ao *baseline*.

#### 5.3.4.2 Análise qualitativa

A seguir, são apresentados alguns exemplos de textos traduzidos pelo VTLM de *masking* aleatório (*baseline*) e pelo VTLM de *masking* textual e visual mais informados, junto a suas referências e *frames* de vídeo correspondentes. Os exemplos mostrados ilustram situações que foram observadas nos resultados obtidos.

No primeiro exemplo, o VTLM de *masking* aleatório faz referência a palavra “website” utilizando o pronome pessoal do caso reto “he” ao invés do pronome “it”, que deveria ter sido usado neste caso. Em contraposição, o VTLM de *masking* visual e textual mais informados não comete o mesmo erro e utiliza o pronome de forma correta.

No segundo exemplo, o pronome pessoal do caso oblíquo “him” é utilizado incorretamente pelo VTLM de *masking* aleatório. Neste caso, o pronome “it” deveria ter sido



**Referência:** That is what gives my site the color options that **it** has.

---

***Baseline*:** That's what gives my website to the color options that **he** has.
**VTLM:** That's what gives my website to the color options that **it** has.

utilizado e o VTLM de *masking* visual e textual mais informados faz o uso correto deste pronome.

**Referência:** And I'm going to push **it** back down.

---

***Baseline*:** And I'm going to push **him** back.
**VTLM:** And I'm going to push **it** back down.

O terceiro caso ilustra a tradução correta do pronome possessivo "your" ("seus", em português) pelo VTLM de *masking* visual e textual mais informados, enquanto o *baseline* o traduz erroneamente para "their" ("deles", em português).

**Referência:** They keep **your** fingers kind of together and are good for a lot of activities.

---

***Baseline*:** They keep **their** fingers together and they're good for many activities.
**VTLM:** They keep **your** fingers together and they're good for a lot of activities.

No quarto exemplo, o VTLM de *masking* visual e textual mais informados utiliza o pronome "it" corretamente como objeto direto do verbo "take", enquanto o *baseline* omite este pronome na tradução.

**Referência:** Now take **it**, put the potter's mark in there.

---

***Baseline*:** Now take, put your potter's mark on there.
**VTLM:** Now take **it**, put the potter's mark in there.

Por fim, a última situação ilustra uma nova melhoria que não foi observada ao aplicar o *masking* visual mais informado ou o *masking* textual mais informado isoladamente. Apesar da informação visual melhorar o desempenho deste modelo multimodal, observou-se que ela pode levar ao uso incorreto de determinados pronomes. Por exemplo, quando o *frame* de vídeo associado ao texto possui um elemento categorizado como *man*, os pronomes usados na tradução tendem a ser "he" ou "him". Da mesma forma, quando há um elemento categorizado como *woman* no *frame* de vídeo, os pronomes tendem a ser "she" ou "her".

Neste exemplo, os dois elementos categorizados como *man* na imagem possivelmente influenciaram a escolha incorreta do pronome “him” após o verbo “bring”. No entanto, este erro não foi cometido pelo VTLM de *masking* visual e textual mais informados, que utilizou o pronome “it” de maneira correta.

**Referência:** He wants to bring **it** back naturally.

---

***Baseline*:** He wants to bring **him** back naturally.
**VTLM:** He wants to bring **it** back naturally.

Dessa forma, as traduções indicam que a aplicação do *masking* visual e textual mais informados beneficia o modelo ao melhorar sua performance na tradução de pronomes pessoais, possessivos e outros. Como resultado, esta estratégia de *masking* também auxilia na tradução de textos mais coerentes, em que é possível utilizar pronomes de forma correta para substituir substantivos, objetos ou frases nominais.

## 5.4 Exploração do *masking* com diferentes categorias semânticas

### 5.4.1 Estudo do corpus

Nesta seção, são descritos os resultados obtidos com o estudo realizado no conjunto de treinamento do corpus How2.

#### 5.4.1.1 *Frames* de vídeo

A Tabela 6 apresenta um panorama das categorias mais recorrentes identificadas na análise, juntamente com algumas subcategorias que demonstram maior complexidade hierárquica.

Tabela 6 – Hierarquia dos níveis de categorias presentes no corpus, acompanhada de suas frequências correspondentes, refletindo a quantidade de *frames* em que cada categoria aparece em relação ao conjunto total de *frames*.

| **Hierarquia** | | | | **Frequência** |
|---|---|---|---|---|
| Nível 0 | Nível 1 | Nível 2 | Nível 3 | (%) |
| Person | | | | 94,81 |
| | Man | | | 73,34 |
| | Woman | | | 51,76 |
| Clothing | | | | 92,51 |
| | Fashion accessory | | | 28,78 |
| | | Hat | | 3,88 |
| | | | Cowboy hat | 0,30 |
| | | | Sombrero | 0,03 |
| Furniture | | | | 40,65 |
| | Table | | | 35,40 |
| | | Coffee table | | 1,44 |
| | | Dining room table | | 0,14 |

#### 5.4.1.2 Texto

Ao contabilizar as palavras relacionadas a seres humanos, obtivemos um maior entendimento sobre o conteúdo do texto, pois identificar palavras que se referem a seres humanos pode ser relevante para determinar a frequência de referências a pessoas no corpus. Os resultados mostraram que o conjunto de texto é formado por 3.796.916 palavras, sendo 69.195 relacionadas à seres humanos (isto é, 1,8%) e 183.787 frases, sendo que 43.206 delas contêm alguma palavra relacionada à seres humanos (isto é, 23,5%).

No entanto, a análise minuciosa das palavras mais proeminentes, após realizar a classificação e remoção de elementos menos representativos, como artigos e preposições, revelou que o conjunto de texto é amplamente caracterizado por pronomes e palavras associadas à categoria semântica "parte do corpo", tais como "mão", "pé", "corpo", entre outras. Dada a predominância dessas palavras relacionadas a seres humanos e levando em consideração que já analisamos o impacto dessa categoria em experimentos anteriores, decidimos direcionar nossos esforços para uma exploração mais aprofundada do *masking* visual.

### 5.4.2 Aumento do *threshold*

Nesta seção, apresentamos os resultados obtidos ao aumentar o *threshold* para 60% na etapa de identificação das classes de objetos durante o processo de mascarar mais *tokens* relacionados a pessoas, com o objetivo de realizar um *masking* mais informado para essa categoria específica.

Ao aumentar o *threshold*, nossa expectativa era selecionar detecções mais precisas e confiáveis de objetos relacionados a pessoas, reduzindo a presença de falsos positivos na

identificação das classes de objetos. Essa abordagem visa garantir que apenas as regiões relevantes e corretamente associadas à categoria "person" fossem mascaradas durante o treinamento do modelo.

#### 5.4.2.1 Resultados

A tabela a seguir apresenta uma comparação dos resultados obtidos com os resultados do experimento anterior.

Tabela 7 – Pontuações BLEU para o VTLM pré-treinado e ajustado para a tarefa de MMT com diferentes níveis de *threshold*.

| | | **Threshold: 40%** | | **Threshold: 60%** | |
|---|---|---|---|---|---|
| | | Teste | Validação | Teste | Validação |
| VTLM: *masking* visual informado | 5% | 52,7 | 53,2 | **53,1** | **53,8** |
| | 7,5% | 51,9 | 52,5 | 52,7 | 52,9 |
| | 10% | 51,6 | 52,2 | 51,6 | 52,2 |

Assim, é possível observar uma melhoria na performance da tradução automática multimodal. Esses resultados indicam que, ao mascarar mais *tokens* que realmente se referem a pessoas e ao diminuir a presença de falsos positivos, o modelo apresenta um desempenho melhor na tradução. Essa melhoria na performance ressalta que a abordagem seletiva ao mascarar determinadas categorias semânticas pode de fato auxiliar e contribuir positivamente para o desempenho geral do modelo.

#### 5.4.2.2 Análise qualitativa

A seguir, é apresentado um exemplo de texto traduzido pelo modelo anterior (*threshold* de 40%), que chamaremos de *baseline*, e pelo modelo atual (*threshold* de 60%), junto a sua referência e *frame* de vídeo correspondente.

Em termos gerais, os desfechos corroboram as tendências identificadas previamente (subseção 5.3.2.2). Ou seja, o modelo continua a demonstrar uma melhora na interconexão entre os elementos textuais e visuais, resultando em um desempenho mais aprimorado na tradução de pronomes pessoais, possessivos e similares.

Além disso, houve uma melhoria em relação ao *baseline* na tradução do pronome "it", que muitas vezes era traduzido para "him" pelo modelo *baseline*. O exemplo a seguir ilustra esta situação, que foi observada nas traduções obtidas.

Neste exemplo, o novo modelo faz a tradução do pronome "it" corretamente, enquanto o *baseline* utiliza incorretamente o pronome "him" para se referir ao boneco.

**Referência:** Then I sometimes use a poppet, as I had mentioned before, and I wrap herbs around **it**.

***Baseline*:** So sometimes I use a puppet, as I mentioned before, and I wrap herbs around **him**.
**VTLM (60%):** So sometimes, I use a puppet, like I mentioned before, and I wrap herbs around **it**.

### 5.4.3 Experimentação com classes mais frequentes

Nesta seção, apresentamos os resultados obtidos a partir da experimentação com as classes mais frequentes do corpus How2, com foco nas categorias "person", "clothing" e "furniture". Nosso objetivo foi explorar o efeito do *masking* visual mais informado ao mascarar uma maior proporção de *tokens* relacionados a cada uma dessas categorias específicas, variando as proporções de mascaramento em 5%, 7,5% e 10%.

#### 5.4.3.1 Resultados

A tabela a seguir apresenta os resultados obtidos.

Tabela 8 – Pontuações BLEU para o VTLM pré-treinado e ajustado para a tarefa de MMT com foco em diferentes categorias semânticas.

| | | ***Person*** | | ***Clothing*** | | ***Furniture*** | |
|---|---|---|---|---|---|---|---|
| | | Teste | Validação | Teste | Validação | Teste | Validação |
| VTLM: *masking* visual informado | 5% | **53,1** | **53,8** | 52,0 | 51,9 | 52,2 | 52,9 |
| | 7,5% | 52,7 | 52,9 | 51,5 | 51,4 | 51,8 | 52,0 |
| | 10% | 51,6 | 52,2 | - | - | - | - |

| | Teste | Validação |
|---|---|---|
| VTLM: *masking* aleatório | 51,80 | 52,44 |

Ao analisar os resultados, constatamos que o *masking* mais informado, ao mascarar uma maior proporção de *tokens* da categoria "person", resultou em uma melhoria significativa na performance da tradução automática multimodal em comparação com o modelo original. Essa melhora indica que a categoria "person" desempenha um papel importante no contexto das informações visuais do corpus, sendo relevante para aprimorar a interpretação e a tradução adequada dessas informações.

Por outro lado, ao focar nos experimentos com as categorias "clothing" e "furniture", não observamos um impacto tão significativo na performance da tradução automática multimodal. O *masking* mais informado, mascarando uma maior proporção de *tokens* nessas categorias, não resultou em melhorias tão substanciais quanto às obtidas com a categoria "person". Isso sugere que, em relação às informações visuais presentes, as categorias

"clothing" e "furniture" podem não ser tão determinantes para a correta interpretação e tradução como a categoria "person".

Portanto, o estudo detalhado das classes mais frequentes revelou que a categoria "person" possui uma relevância significativa para o desempenho do modelo de tradução automática multimodal no corpus How2. O mascaramento seletivo de *tokens* relacionados à categoria "person" resultou em melhorias notáveis na performance, destacando a importância de considerar essa categoria específica ao aprimorar as técnicas de *masking* em abordagens multimodais.

#### 5.4.3.2 Análise qualitativa

A seguir, são apresentados alguns exemplos de textos traduzidos pelo VTLM de *masking* aleatório (baseline) e pelo VTLM de *masking* visual mais informado com foco nas categorias "person", "clothing" e "furniture", junto com suas referências e *frames* de vídeo correspondentes. Os exemplos mostrados ilustram situações que foram observadas nos resultados obtidos.

De forma geral, o VTLM de *masking* visual mais informado com foco na categoria "person" apresenta as mesmas melhorias na tradução que foram observadas anteriormente (subseção 5.3.2.2) com um *threshold* menor. Ou seja, o modelo aprende a relacionar melhor as informações textual e visual, obtendo uma melhor performance na tradução de pronomes pessoais, possessivos e outros.

O primeiro exemplo ilustra um dos casos observados. Neste exemplo, o VTLM de *masking* aleatório utiliza o pronome "him" na tradução ao invés de "it". Em contrapartida, o VTLM de *masking* visual mais informado com foco na categoria "person" traduz a frase corretamente.

**Referência:** Probably not going to catch a flush with a three to **it**.

---

***Baseline*:** It's probably not going to get a flush with a three for **him**.
**VTLM:** It's probably not going to get a flush with a three to **it**.

O VTLM com foco na categoria "clothing" também mostra uma melhoria em relação ao baseline, embora ainda não consiga uma tradução perfeita. Verificou-se que o modelo emprega substantivos associados ao vestuário de forma mais eficaz do que o VTLM de *masking* aleatório. Os dois exemplos a seguir ilustram algumas situações observadas. No primeiro exemplo, o modelo faz o uso correto do substantivo "gloves", ao contrário do modelo *baseline*. E o segundo exemplo ilustra uma situação similar, em que o modelo faz o uso correto do substantivo "jeans", enquanto o modelo *baseline* omite esta palavra da tradução.

**Referência:** These are a cotton polyester blend of cloth **gloves** and they're good for making snowballs.

---

***Baseline*:** These are a cloth blend and they are good for doing snow balls.
**VTLM:** These are a cotton mixture of cloth **gloves** and they're good to make snow balls.

**Referência:** Now the funny thing is a lot of times it's going to be cheaper at a retail site either because it's on special, they have coupons, they have some sort of all **jeans** twenty percent off sale, etc.

---

***Baseline*:** Now the funny thing is that a lot of times it's cheaper on a retail website because it's special they have coupons, they have some sort of twenty dollars a discount store, etc.
**VTLM:** Now the funny thing is that a lot of times it's cheaper on a retail site because it's in special, they have coupons, they have some type of **jeans** twenty percent discount, etc.

Por fim, o VTLM com foco na categoria "furniture" também apresenta uma melhoria em relação ao *baseline*. Observou-se que o modelo emprega substantivos associados à mobília de forma mais eficaz do que o VTLM de *masking* aleatório. Os dois exemplos a seguir ilustram algumas situações observadas. No primeiro exemplo, o modelo traduz corretamente o substantivo "furniture", enquanto o modelo *baseline* traduz este substantivo incorretamente para "moves". E, no segundo exemplo, o modelo faz o uso correto do substantivo "bed", ao contrário do modelo *baseline*, que utiliza a palavra "reading".

**Referência:** Some of the miscellaneous pieces of **furniture** that I have, are this stool that goes along with my couch, I have a footstool kind of thing.

---

***Baseline*:** Some of the fun **moves** that I have, are this stool that goes with my couch, I have kind of a stool.
**VTLM:** Some of the fun **furniture** that I have, are this stool that goes along with my couch, I have a type of stool.

**Referência:** Make it a little easier to walk and build up the **bed** at the same time.

---

***Baseline*:** Leaving a little bit easier to walk and raising the **reading** at the same time.
**VTLM:** Let it a little bit easier to walk around and increasing the **bed** at the same time.

Dessa forma, a análise dos exemplos de traduções geradas pelo VTLM de *masking* visual mais informado, com ênfase nas categorias "person", "clothing" e "furniture", revela uma melhoria em comparação com o VTLM de *masking* aleatório. O modelo aprimorado

demonstra a sua capacidade de compreender e integrar informações visuais, resultando em traduções mais coerentes e alinhadas com o contexto. Esse avanço é evidenciado nas melhorias na tradução de pronomes no contexto "person", na seleção mais precisa de termos relacionados a vestuário no âmbito "clothing", e na utilização mais acertada de palavras associadas a móveis sob a categoria "furniture". Embora a performance ainda não alcance a perfeição em todas as situações, esses exemplos indicam uma tendência positiva na evolução do modelo.

Contudo, apesar das melhorias observadas, é importante notar que a performance global do VTLM com enfoque nas categorias "clothing" e "furniture" ainda não supera de forma significativa a do modelo *baseline*. Isso sugere que, embora haja avanços promissores, ainda há espaço para aprimoramentos visando atingir traduções mais consistentes e precisas em todas as categorias abordadas.

# Capítulo 6

# Conclusão

Este trabalho abordou um estudo extensivo do VTLM, visando testar sua capacidade de generalização para outras linguagens e corpora, assim como melhorar a eficácia do pré-treinamento de modelos de visão e linguagem, através de novas abordagens de *masking*.

Para isso, o corpus multimodal e multilíngue How2 foi utilizado, proporcionando um novo idioma (português brasileiro) a ser aprendido e apresentando um cenário mais desafiador, onde o o texto associado a cada *frame* de vídeo não é uma simples descrição do mesmo, mas uma legenda que pode não estar relacionada ao seu *frame* correspondente.

Assim, etapas de pré-processamento necessárias foram aplicadas no corpus para que ele pudesse ser usado para investigar a eficácia do modelo para um novo par de idiomas. O desempenho dos modelos treinados também foi comparado com algumas *baselines* e uma análise qualitativa foi realizada para comparar as traduções obtidas por cada modelo, o que evidenciou o impacto de adicionar a modalidade visual junto aos pares de tradução. Os resultados também indicaram uma capacidade de generalização do modelo para outros idiomas e sua boa performance mostrou que ele é capaz de superar os desafios adicionais trazidos com o uso de um corpus que é uma coleção de vídeos (e não imagens).

Além disso, este trabalho mostrou que a predição de elementos mascarados específicos pode beneficiar o pré-treinamento visual e multilíngue, pois o modelo pré-treinado pode adquirir uma melhor compreensão de determinadas estruturas de linguagem, o que melhora tarefas de visão e linguagem, como tradução automática multimodal.

São apresentadas três estratégias de *masking* seletivo que se concentram em mascarar *tokens* linguísticos e visuais específicos que podem contribuir para a compreensão de alguns padrões de linguagem: *masking* visual mais informado, *masking* textual mais informado, *masking* visual e textual mais informado.

Os resultados obtidos evidenciaram acurácia estado-da-arte no conjunto de dados

How2, mostrando que as abordagens de *masking* propostas produzem melhorias significativas em relação à estratégia original de *masking* aleatório para o desempenho na tarefa de tradução automática multimodal.

Por fim, exploramos o uso do *masking* em diferentes categorias semânticas com o objetivo de melhorar a tradução automática multimodal. Os resultados mostraram que o *masking* mais informado em *tokens* relacionados à categoria "person" leva a uma melhoria significativa na performance, destacando a importância dessa categoria para a interpretação das informações visuais do corpus. Por outro lado, outras categorias que aparecem com frequência no corpus, como "clothing" e "furniture", não apresentam melhorias, sugerindo que podem ser menos determinantes na tradução automática neste corpus específico. Além disso, constatamos também que o aumento do *threshold* para identificação das classes de objetos durante o treinamento resulta na redução de falsos positivos nas detecções, o que leva a uma melhor eficácia do *masking* mais informado e melhora a performance da tradução automática multimodal.

Assim, este projeto destaca a relevância do *masking* seletivo em categorias semânticas específicas para aprimorar a tradução automática multimodal e proporciona *insights* sobre o comportamento do modelo em relação a diferentes tipos de informações semânticas. Essas descobertas podem contribuir para o desenvolvimento de técnicas mais eficazes de *masking* em abordagens multimodais e enriquecer a compreensão sobre a interação entre imagens e textos em sistemas de tradução automática.

Para trabalhos futuros, existem várias direções promissoras que podem ser exploradas para aprimorar a eficácia das estratégias de mascaramento propostas neste estudo. Uma abordagem interessante seria considerar mascarar verbos de ação de forma seletiva, uma vez que as legendas frequentemente descrevem procedimentos e ações realizadas nas imagens. Além disso, estender a avaliação das estratégias de *masking* para outros corpora texto-imagem poderia enriquecer a compreensão da robustez e generalização dessas estratégias em contextos variados.

# Referências

BANERJEE, S.; LAVIE, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 65–72.

BARRAULT, L. et al. Findings of the third shared task on multimodal machine translation. In: **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**. Belgium, Brussels: Association for Computational Linguistics, 2018. p. 304–323.

CAGLAYAN, O. et al. Cross-lingual Visual Pre-training for Multimodal Machine Translation. In: **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers**. online: Association for Computational Linguistics, 2021.

CALIXTO, I.; LIU, Q. Incorporating global visual features into attention-based neural machine translation. In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 992–1003.

CALIXTO, I.; RIOS, M.; AZIZ, W. Latent variable model for multi-modal translation. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, 2019. p. 6392–6405.

CHEN, Y.-C. et al. Uniter: Universal image-text representation learning. 2020.

CLARK, K. et al. Electra: Pre-training text encoders as discriminators rather than generators. In: **International Conference on Learning Representations**. [s.n.], 2020. Disponível em: <https://openreview.net/forum?id=r1xMH1BtvB>.

CONNEAU, A.; LAMPLE, G. Cross-lingual language model pretraining. Curran Associates, Inc., v. 32, 2019.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language**

**Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186.

ELLIOTT, D. et al. Findings of the second shared task on multimodal machine translation and multilingual image description. In: **Proceedings of the Second Conference on Machine Translation**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 215–233.

_____. Multi30K: Multilingual English-German image descriptions. In: **Proceedings of the 5th Workshop on Vision and Language**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 70–74.

ELLIOTT, D.; KÁDÁR, Á. Imagination improves multimodal translation. In: **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017. p. 130–141.

GIRSHICK, R. **Fast R-CNN**. arXiv, 2015. Disponível em: <https://arxiv.org/abs/1504.08083>.

GIRSHICK, R. et al. **Rich feature hierarchies for accurate object detection and semantic segmentation**. arXiv, 2013. Disponível em: <https://arxiv.org/abs/1311.2524>.

HE, K. et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. In: **Computer Vision – ECCV 2014**. Springer International Publishing, 2014. p. 346–361. Disponível em: <https://doi.org/10.1007%2F978-3-319-10578-9_23>.

_____. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778.

HUANG, J. et al. Speed/accuracy trade-offs for modern convolutional object detectors. 11 2016.

HUANG, P.-Y. et al. Attention-based multimodal neural machine translation. In: **Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 639–645.

_____. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. 2021.

IVE, J.; MADHYASTHA, P.; SPECIA, L. Distilling translations with visual awareness. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, 2019. p. 6525–6538.

JOSHI, M. et al. SpanBERT: Improving pre-training by representing and predicting spans. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 8, p. 64–77, 2020. Disponível em: <https://aclanthology.org/2020.tacl-1.5>.

KINGMA, D.; BA, J. Adam: A method for stochastic optimization. 12 2014.

KUZNETSOVA, A. et al. The open images dataset v4. **International Journal of Computer Vision**, Springer Science and Business Media LLC, v. 128, n. 7, p. 1956–1981, Mar 2020. ISSN 1573-1405.

LAN, Z. et al. Albert: A lite bert for self-supervised learning of language representations. 2020.

LEVINE, Y. et al. {PMI}-masking: Principled masking of correlated spans. In: **International Conference on Learning Representations**. [s.n.], 2021. Disponível em: <https://openreview.net/forum?id=3Aoft6NWFej>.

LI, L. H. et al. Visualbert: A simple and performant baseline for vision and language. In: **Arxiv**. [S.l.: s.n.], 2019.

LIN, J. et al. Interbert: Vision-and-language interaction for multi-modal pretraining. 2021.

LIU, P.; CAO, H.; ZHAO, T. Gumbel-attention for multi-modal machine translation. 2021.

LIU, W. et al. Ssd: Single shot multibox detector. In: LEIBE, B. et al. (Ed.). **ECCV (1)**. Springer, 2016. (Lecture Notes in Computer Science, v. 9905), p. 21–37. ISBN 978-3-319-46447-3. Disponível em: <http://dblp.uni-trier.de/db/conf/eccv/eccv2016-1.html#LiuAESRFB16>.

LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. 2019.

LONG, Q.; WANG, M.; LI, L. Generative imagination elevates machine translation. In: **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Online: Association for Computational Linguistics, 2021. p. 5738–5748.

LU, J. et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Curran Associates, Inc., v. 32, 2019.

NI, M. et al. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2021. p. 3977–3986.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318.

RADFORD, A. et al. Improving language understanding by generative pre-training. 2018.

REDMON, J. et al. **You Only Look Once: Unified, Real-Time Object Detection**. arXiv, 2015. Disponível em: <https://arxiv.org/abs/1506.02640>.

REDMON, J.; FARHADI, A. **YOLOv3: An Incremental Improvement**. 2018. Cite arxiv:1804.02767Comment: Tech Report. Disponível em: <http://arxiv.org/abs/1804.02767>.

REN, S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Curran Associates, Inc., v. 28, 2015.

_____. **Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks**. 2016.

ROTHE, S.; NARAYAN, S.; SEVERYN, A. Leveraging pre-trained checkpoints for sequence generation tasks. 2020.

SANABRIA, R. et al. How2: A large-scale dataset for multimodal language understanding. 2018.

SATO, J.; CASELI, H.; SPECIA, L. Multilingual and multimodal learning for brazilian portuguese. In: **Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)**. Marseille, France: European Language Resources Association (ELRA), 2022.

_____. Choosing what to mask: More informed masking for multimodal machine translation. In: **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 244–253. Disponível em: <https://aclanthology.org/2023.acl-srw.35>.

SENNRICH, R.; HADDOW, B.; BIRCH, A. Neural machine translation of rare words with subword units. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 1715–1725.

SHARMA, P. et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 2556–2565.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. [S.l.: s.n.], 2020.

SPECIA, L. et al. A shared task on multimodal machine translation and crosslingual image description. In: **Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 543–553.

SRIVASTAVA, N. et al. Dropout: A simple way to prevent neural networks from overfitting. v. 15, n. 56, p. 1929–1958, 2014. Disponível em: <http://jmlr.org/papers/v15/srivastava14a.html>.

SU, W. et al. Vl-bert: Pre-training of generic visual-linguistic representations. 2020.

SZEGEDY, C. et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In: **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**. [S.l.]: AAAI Press, 2017. (AAAI'17), p. 4278–4284.

TAN, H.; BANSAL, M. LXMERT: Learning cross-modality encoder representations from transformers. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 5100–5111.

VASWANI, A. et al. Attention is all you need. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

_____. Attention is all you need. Curran Associates, Inc., v. 30, 2017.

WU, Y. et al. **Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation**. 2016.

XIAO, D. et al. ERNIE-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding. In: **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Online: Association for Computational Linguistics, 2021. p. 1702–1715. Disponível em: <https://aclanthology.org/2021.naacl-main.136>.

YANG, Z. et al. **XLNet: Generalized Autoregressive Pretraining for Language Understanding**. 2020.

YAO, S.; WAN, X. Multimodal transformer for multimodal machine translation. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 4346–4350.

YIN, Y. et al. A novel graph-based multi-modal fusion encoder for neural machine translation. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 3025–3035.

ZHANG, Z. et al. ERNIE: Enhanced language representation with informative entities. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, 2019. p. 1441–1451. Disponível em: <https://aclanthology.org/P19-1139>.

ZHOU, M. et al. A visual attention grounding neural model for multimodal machine translation. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 3643–3653.

ZHUANG, L. et al. A robustly optimized BERT pre-training approach with post-training. In: **Proceedings of the 20th Chinese National Conference on Computational Linguistics**. Huhhot, China: Chinese Information Processing Society of China, 2021. p. 1218–1227. Disponível em: <https://aclanthology.org/2021.ccl-1.108>.