Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5792

A bilingual analysis of multi-head attention mechanism for image captioning based on morphosyntactic information

João Gondim* [**State University of Campinas / Federal University of Bahia** | ***joao.gondim@ic.unicamp.br***]
Daniela Barreiro Claro [**FORMAS Research Center - Institute of Computing - Federal University of Bahia** | ***dclaro@ufba.br***]
Marlo Souza [**FORMAS Research Center - Institute of Computing - Federal University of Bahia** | ***msouza1@ufba.br***]

FORMAS Research Center on Data and Natural Language, Institute of Computing, Federal University of Bahia, Av. Milton Santos, s/n, Campus de Ondina, Salvador - Bahia 40190-110, Brazil.

Received: 31 March 2025 • **Accepted:** 17 July 2025 • **Published:** 17 October 2025

Abstract Image Captioning is the task of describing the information conveyed by an image, i.e., its visual content in a natural language. Most of the current researches make use of the encoder-decoder architecture to create relations between images (the inputs) and text (the output). These relations are originated from the attention mechanisms, present on the Transformer model, and can be leveraged to understand how the image-text relationship is encoded during training and inference times. This work investigates the hypothesis that the attention mechanism behaves analogously for words that share morphosyntactic labels within texts. To this matter, the attention weights for each predicted word — posed as the "focus" given in the image at each step — are gathered, averaged and inspected; also, the analysis are performed taking into account one model trained with English captions and another trained with Portuguese captions, therefore comparing two languages with different morphological organization. Our results show that words with the same functioning in the sentence,e.g., being prone to similar inflections, usually have the same focal point in the image. Our work sheds light to the importance of linguistic studies for the vision-language area, reinforcing the benefits of including language-aware knowledge during training.

Keywords: Image Captioning, Attention, Morphosyntax

1 Introduction

Image Captioning is the task of describing the information conveyed by an image, i.e., its visual content in a natural language [Stefanini *et al*., 2023]. While it is an important task for Computer Vision and Natural Language Processing within multiple applications, such as description for visually impaired [Web Accessibility Initiative, 2022], multimodal information retrieval [Sharif *et al*., 2020; Gatt and Krahmer, 2018], this task presents many important difficulties, both from a technical and theoretical point of views.

Image Captioning requires knowledge and technology from distinct areas of Artificial Intelligence. From Computer Vision, systems can extract meaningful content from images. From Representation Learning, models encoding the visual information can be employed by other methods. From the Natural Language Generation, systems that construct the linguistic description of the image [Reale-Nosei *et al*., 2024].

On the other hand, from the state of the art, the concept of what constitutes a good linguistic description of an image is not yet evident in the community. Visual communication theorists, such as Ann Marie Barry Barry [1997], have long discussed how visual communication cannot be reduced to language, as these mechanisms operate radically differently. In fact, based on the semiotics of C. S. Pierce, authors such as Moriarty [2002] argue that visual communication may operate in different systems of signifying, such as *mimesis* or evidence, from that of verbal/linguistic communication, in which the relation between the signifier and the meaning is arbitrary.

For example, the same image can evoke different concepts - and thus descriptions - when presented to people from different cultural or linguistic backgrounds, as discussed by Liu *et al*. [2021]. As such, it remains unclear how to circumscribe the task of comparing different approaches in the literature in a meaningful manner. To account for the variability in possible descriptions, some image captioning datasets presented multiple captions for the same image, representing the variability in descriptions.

Recent image caption approaches have shown a good increase in their results by employing deep learning-based architectures [Vinyals *et al*., 2015; Hossain *et al*., 2019; Sharma and Padha, 2023], such as the encoder-decoder architecture (Cho *et al*. [2014]). In encoder-decoder image captioning, images are encoded into feature vectors later decoded when generating the captioning as natural language sentences [Stefanini *et al*., 2023]. Encoder-decoder layers with Transformers [Vaswani *et al*., 2017] increment multi-head attention mechanisms, employing the attention mechanism to create representations of input and output data and make relations between the expected outputs [Cornia *et al*., 2022].

Such relations between the inputs and outputs deserve special attention, as some works have investigated how the at-

*Work performed while at Federal University of Bahia

tention mechanism can be employed to understand how information is encoded within transformer-based models and how this information can be used to explain the decisions of the model [Vig, 2019; Clark *et al*., 2019].

Recent works focus on the visual sub-task of image captioning, enabling different pre-training strategies to explore relations between text and image [Li *et al*., 2020; Zhang *et al*., 2021b; Hu *et al*., 2021; Wang *et al*., 2022b]. Other researchers employed part-of-speech (POS) as an additional input to Image Captioning models, suggesting a correlation between the next word of a sentence and specific parts of the image that are being focused at each generation timestep [Lu *et al*., 2017; He *et al*., 2017]. Recently, approaches have argued that POS summarizes the image contents and can guide the caption generation [Deshpande *et al*., 2019]. Words with different POS have divergent roles in the sentence and correlate differently with visual information [Zhang *et al*., 2021a].

Hence, we hypothesized that text inputted in training may enhance explainability for models from images, enabling a more accurate relation between image and caption. We evaluate whether/how linguistic structures are represented within a transformer-based model following an encoder-decoder approach. In order to achieve this goal, we investigate how specific linguistic structures are represented in a transformer attention mechanism. In doing so, we evaluate what elements of an image are in evidence to determine which words and functions they must perform on the generated caption.

Our work inspects self-attention heads by assessing the average attention head weights for each word predicted. By clustering each head's weights, we expect common shared patterns will emerge. More specifically, we hypothesize that words that share morphosyntactic characteristics and functions, e.g., are subject to similar inflections, have their attention focus behaving similarly during the next word prediction. We consider such feature investigation to be of relevant use for other researchers and machine learning engineers during the creation of Image Captioning models, especially for rich flexional languages, such as Portuguese, and, possibly, for agglutinative languages, where words are formed through the combination of units called morphemes [Indurkhya and Damerau, 2010].

We employ the Flickr30k dataset [Young *et al*., 2014] to perform the experiments. Since multilingual Image Captioning has received little attention due to the lack of proper datasets [dos Santos *et al*., 2022; Gondim *et al*., 2022], our study uses an automatically translated version of the Flickr30k dataset, with its captions in the Portuguese language. Even though the employment of an automatic translator could elicit the introduction of translation artifacts, we went in this direction both because it is a common practice in the image captioning literature [Rosa *et al*., 2021; Chen *et al*., 2023; Santos *et al*., 2023] and also as a form of maintaining a pattern between the captions used to train the model and to evaluate our hypothesis. By employing translated sentences, our trained models will be adjusted with parallel sentences, which is crucial to the analysis. We argue that this could not be the case if we used a dataset crafted by Brazilian annotators, who might have different biases when creating the sentences compared to the original ones in the Flickr30k dataset.

Using a translated dataset, we also consider and analyze whether our findings are language-specific. We envisage our experimental results clarifying the relations learned within a Transformer model. Moreover, our results show that being aware of such significant issues in developing multilingual datasets for the Image Captioning task is important.

Our contributions are threefold:

- We provide an analysis of attention head weights of our Image Captioning model, focused on the morphosyntactic class of each word;
- We evaluate the use of attention heads for interpreting Image Captioning models, considering different languages: Portuguese (PT) and English (EN);
- We developed a new method[1] to analyze the general behavior of a model by averaging the attention distributions of the exact words between different predicted sentences.

This paper is organized as follows: Section 2 describes our Related Works; Section 3 introduces the Transformer and its usage in the Image Captioning area. Section 4 presents the model employed in the experiments and our methodology to assess attention heads. Section 5 presents the experimental setup of each experiment performed and the dataset details; Section 6 shows the results; Section 7 discusses our findings, followed by our conclusions and future work.

2 Related Work

This section explores relevant works on Image Captioning, tackling three aspects: image captioning models, the addition of linguistic information into the training pipeline, and efforts towards exploring attention weights to understand the model's biases and behavior.

Early approaches to Image Captioning were mainly rule-based, focusing on filling templates with inferences from the images [Yao *et al*., 2010; Aker and Gaizauskas, 2010; Yang *et al*., 2011]. Other strategies were concentrated on caption retrieval, employing neural networks trained to rank the best suited captions for one image, making use of a learned multimodal representation space for both text and images [Karpathy *et al*., 2014; Ordonez *et al*., 2011; Farhadi *et al*., 2010; Pan *et al*., 2004]. These strategies were eventually replaced with neural-based techniques, with the emergence and success of deep learning generative models [Stefanini *et al*., 2023], which explored evolutions in learned representation techniques, higher-performing neural architectures, and evolution of computer infrastructure that enabled a great performance increase for the task.

Recent work on the field employ the encoder-decoder model, introduced by Cho *et al*. [2014] for the machine translation task, which encapsulates a sentence to be translated into a representation (encoding) and than makes use of such representation to create the final translation (decoding). Xu *et al*. [2016] state that the usage of encoder-decoder architectures places the task of describing an image similarly to "translating an image into a sentence". To our knowledge, Vinyals

[1] We provide our code in `https://github.com/FORMAS/ImageCaptioningPT`.

et al. [2015] introduced the first encoder-decoder model for the Image Captioning task, employing a Convolutional Neural Network (CNN) as the encoder responsible to generate a representation of the image to be captioned, which is later employed as input into a Recurrent Neural Network (RNN) to create an image description. He *et al*. [2017] employ a similar architecture, adding the part-of-speech of words to determine if visual features should be used to predict the next word of the image description, a first attempt to employ morphosyntactic features to enhance the generated caption. Xu *et al*. [2016] were the first to add an attention mechanism into a encoder-decoder model. Beyond improving metrics, they show that the outputs of an attention mechanism resemble the idea of visualizing what the model "sees" or "focus" when predicting each word, an idea that our work leverages to inspect possible patterns on attention mechanisms' outputs. Lu *et al*. [2017] argue that the decoder needs less visual information when predicting words like "of", "in" or "at", their *Adaptive Attention Model* learns whether the prediction needs information only from the language model at the decoder or if visual features are needed. Though some of these early approaches leverage morphosyntactic features, they solely focus on the English language, and do not analyze word inflections.

Recent Image Captioning models are based on the Transformer architecture [Vaswani *et al*., 2017]. Cornia *et al*. [2020] modify the self-attention operator, adding extra memory vectors that learn multi-level relationships not present in the extracted image features, for example: from the presence of a man and the presence of a basketball, the extra space can infer the concept of "player" or "game". In Huang *et al*. [2019], the authors employ a second attention mechanism to refine the outputs from the regular attention heads, naming their approach as *Attention on Attention*. They argue that not all information retrieved from the encoder's attention mechanism is useful, leading the language model of the decoder to take wrong decisions when predicting the next word. Such work demonstrates the importance of inspecting attention weights. In Wang *et al*. [2022a], authors argue that current models treat inferences of visual and nonvisual words as the same, which generates generic captions. They implement an Image Captioning model that leverages Transformers with POS guidance to train different attention models with syntactic information. In a similar direction, Al-Qatf *et al*. [2024] claim that different attention contexts differ on their importance depending on the word being generated. As stated in their paper, visual words (such as nouns or verbs) leverage information from visual features and are often given more focus on current research, whereas non-visual words (for example, prepositions or determiners) are usually neglected, requiring semantic context which, in their work, is managed by the addition of different blocks for the generation of a global semantic context and modeling of its importance with the visual clues that will be used for the next word generation. Beyond the bilingual analysis (resulting in findings for both English and Portuguese), our work differentiates by proposing a methodology for investigating the attention heads relative to morphosyntactic attributes.

Moreover, recent advances and works on vision-language tasks[2] [Hu *et al*., 2021; Wang *et al*., 2022b; Li *et al*., 2022] explore pre-training with large datasets, connecting joint representations of multimodal knowledge [Stefanini *et al*., 2023]. These works lead the path to extensive training, enhancing models results on benchmark datasets, but without considering linguistic characteristics that can be explored for explaining the system's predictions and more effective training.

Different from these related works, our approach focus on the analysis of the outputs of Image Captioning models. Vig [Vig, 2019] developed a tool to facilitate visualizing attention mechanisms of Transformers, arguing that one advantage added by the use of attention is the augment in interpretability by inspecting the weight assignment depending on the inputs given. Clark et al. [Clark *et al*., 2019] investigate what linguistic knowledge are being learned by large pre-trained language models. Using BERT [Devlin *et al*., 2019] as base model, they found specific attention heads that attend to certain POS of input sentences. To our knowledge, our work is the first to assess attention mechanisms' weight assignment for multi-modal tasks, evaluating the outcomes of the attention mechanisms with images as input.

3 Background

Introduced by Vaswani *et al*. [2017], the Transformer architecture is an encoder-decoder neural network architecture that relies on attention mechanisms to build dependencies and representations from the inputs and outputs during training. The authors' work avoids the usage of recurrent neural networks to decrease training time while maintaining tasks' metrics [Vaswani *et al*., 2017]. This architecture was developed for Machine Translation and Natural Language Understanding, and it has had significant influence in many NLP fields, later growing in usage in Vision Language (e.g., visual question answering, image-text retrieval, image captioning) tasks as well [Stefanini *et al*., 2023].

At the core of a Transformer network, attention mechanisms implement scaled dot-products to measure similarity between the input tokens, while linear transformations create queries, keys and values. Taking as an example: given the input sequences X and $\dot{X}$, formed by *n* tokens $x_1, x_2 ... x_n$ and $\dot{x}_1, \dot{x}_2 ... \dot{x}_n$, linear transformations are applied in order to create key and value vectors, respectively $k_i = W_k x_i$ and $v_i = W_v x_i$, where W_k and W_v are defined as key and value transformation matrices. The operation is the same to query vectors: $q_i = W_q \dot{x}_i$, with W_q having the same size as W_k and W_v. Query and key vectors have their similarity computed, generating weights distributed over the value vectors [Cornia *et al*., 2022]. When employed between the same inputs (i.e. $X = \dot{X}$), the attention mechanism is named "self-attention" [Vaswani *et al*., 2017]. In their work, Vaswani *et al*. [2017] state that there are benefits to project the value, key and query vectors to different linear subspace, learned during training, and then perform the attention mechanism, with their resulting outputs concatenated as final values. This method allows parallel processing and produces better representations, since the model is capable of jointly learning

[2] Image Captioning, Vision Question Answering, Image Generation, Vision Entailment

to different positions at different subspaces, they call this "Multi-head attention".

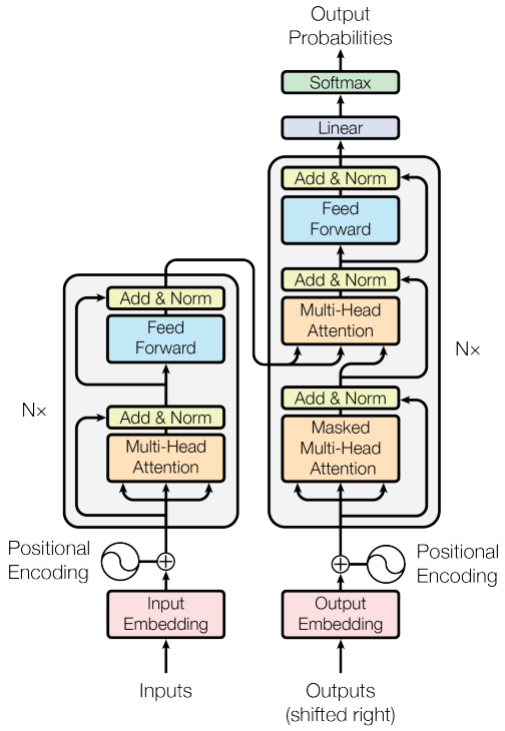


Figure 1. Transformer's architecture model. Source Vaswani *et al*. [2017]

The Transformer's encoder module, left side in Figure 1, is organized with a multi-head self-attention block followed by a feed-forward (FF) layer, after each of them there is a residual connection ("Add") and a normalization ("Norm") operations. The encoder module is responsible for generating representations of the input sequence, that are then passed to the decoder. The decoder is organized with a masked multi-head self-attention block — ensuring that the generation of the next word takes into consideration only the current and the previous context —, followed by a multi-head attention performed on the values the encoder yields (also called "cross-attention" by Cornia *et al*., 2022) and then another feed-forward layer, each attention block and FF layer is followed by an Add & Norm operations. The first attention block of the decoder has its inputs masked, so no representation from future positions are used when predicting the current word. Both the encoder and the decoder are stacked in N identical layers (represented with N_x in Figure 1). After the stacked decoder layers, a linear transformation is performed followed by a Softmax operation that outputs words probabilities [Vaswani *et al*., 2017].

In the context of Image Captioning, modern architectures are composed of two modules: a visual encoder and a language model as a decoder [Salgotra *et al*., 2024]. The visual encoder is responsible for processing image representations returned by visual feature extractors, and the language model generates the final caption describing the input image. Hence, the language model performs cross-attention between visual elements (used as keys and values) and words (used as queries). This multi-modal cross-attention outputs relations between the word to be predicted and visual features from the visual encoder, hence being the center of the analysis performed in this work. Xu *et al*. [2016] have shown that the attention mechanism operations between image features and words relate well with human perception when predicting words of a description, Vig [2019] argue that attention mechanisms aid on model interpretation and Clark *et al*. [2019] showed that attention heads can learn linguistic information, therefore, in our work, we analyze these outputs, inspecting the relations learned on different attention heads of the Multi-Head attention layer.

4 Model

Our model consists of a Transformer adapted from the combination of three different models available [Nain, 2021; Gautam, 2021; Tensorflow, 2022]. To extract visual features from the images, we employed EfficientNetV2 [Tan and Le, 2021] due to its combination of accuracy a small number of parameters[3].

The hyperparameters were 2 sublayers for each Encoder and Decoder layers, 6 Multi-Head Attention heads, the embedding dimension was set to a size of 512, and the fully-connected networks to 2048. The model was trained using Adam optimizer and a learning rate schedule with 20,000 warm-up steps and a learning rate of 10^{-5} after warm-up, we set a stopping criteria of 5 epochs without a decrease in the validation loss. Also, the model was trained with a batch size of 64 and a fixed sentence length of 25 words. The same architecture was employed to create and train two Image Captioning models, one trained with Portuguese captions (Model-PT) and another with English sentences (Model-EN). Furthermore, to be able to inspect the outcomes of the Multi-Head Attention mechanism that performs cross-attention between image features and the contextualized sentence vector, we retrieve the output weights of the responsible attention module, as illustrated by the highlighted part in Figure 2. The figure shows a scheme of the architecture employed in this work, emphasizing the output weights given by the attention mechanism.

Since we intend to analyze the self-attention mechanism's behavior, we decided not to employ any other known strategies to enhance captions, such as beam search, large data pre-training, or Self-critical Sequence Training [Rennie *et al*., 2016]. In other words, we used a naïve transformer architecture in our experiments.

4.1 Methodology

After training the model, captions from each test set image are generated. As shown in Figure 3, we collect the attention weights returned for each word of the model's outputs during this process. These collected weights are gathered regarding the word predicted; for example, among all the images in the test set, the word "grupo" (translation: "group") was predicted 105 times by Model-PT. Therefore, 105 attention weights were collected for the word "grupo". We summarize these outcomes, using their mean to make a final representation of the heads for each word to be analyzed. We represent each

[3] https://keras.io/api/applications/

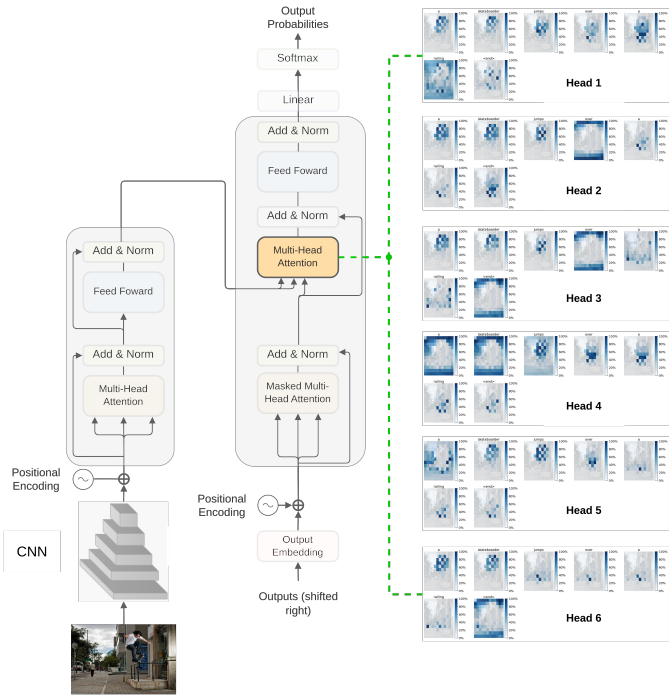


Figure 2. Diagram of the employed architecture. The highlighted part shows the retrieval of the weights distributed by each attention head.

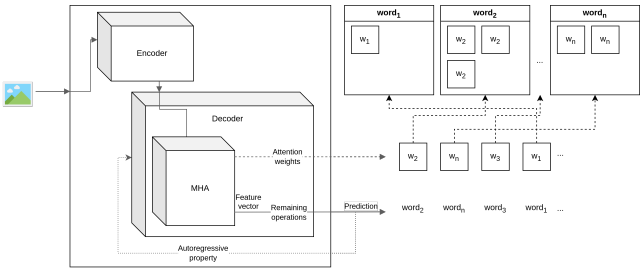


Figure 3. Pipeline of the experiment.

word by their attention head weights' mean to observe the general behavior of the Multi-Head Attention mechanisms during inference.

When using the trained models to describe Figure 4, our evaluation mechanism considers each head's attention as described in Figures 5 and 6. A generated caption with the attention weights of *Head 1*[4] distributed over the image for each word.

Figure 4. Image depicted by the models trained and with attention heads shown in Figures 5 and 6.

The image shows that the attention weights returned by the Multi-Head Attention heads are distributed over a 12x12 grid (determined by the size of the image input and the outcome of the feature extractor), where each cell is given a value of the distributed attention weight, in other words the "focus" given on a certain portion of the image. The attention has different values distributed over each tile of the grid, with all the values summing up to 1. Hence, to facilitate the evaluation, maximum values (where the "focus" is higher) are normalized to 100%. For example, in Figure 5 the word "um" (translation: "a") has a different distribution of weights compared with the word "de" (translation: "of"). In this case, we observe that the words related to the group of people and the scenario — "um" ("a"), "grupo" ("group"), "pessoas" ("people"), "rua" ("street") — had completely different weights when compared with prepositions or verbs — "de" ("of"), "está" ("are"), "andando" ("walking"), "pela" ("down the").

In Figure 6, the same image is shown with the captioning generated by Model-EN and the attention weights for Head 1. Although the caption is wrong, as the model describes every person as a woman, what breaks our attention is the difference in the weight distribution measured between two different languages: Portuguese and English.

Thus, our work takes this gap and investigates Multi-Head Attention mechanisms when applied in an Image Captioning model. In order to achieve this, we summarize the attention given to each word by taking the mean distribution for each time they are generated during a sentence prediction. Figure 7 shows the mean attention distribution for the word "andando" (translation: "walking") for each head in Model-PT. The image illustrates the average focus of each head when predicting the word "andando", showing different behaviors depending on which head is assessed. In the following section, we show that, for the Portuguese generated sentences, inflections in certain word classes, such as gender inflection of nouns and person inflection on verb conjugations, may influence the behavior of attention heads, showing that attention heads may explore specialized morphosyntactic information in generating captions.

Figure 8 shows the mean attention weight distribution of Model-EN when predicting the word "walking". In a first glance on both Figures 7 and 8, even though both words have the same sense when employed in a sentence, there are notable divergences in Heads 1 and 5. We hypothesize that the weights distributed on the corner on the Portuguese example are due to the two different forms the verb "walking" was translated when creating Flickr30k-PT: "a andar" — the infinitive form — and "andando" (as in the example above) — the present continuous form.

To ease the analysis of differing and alike heads, we present the difference of each attention head tile on the 12x12 grid (after normalization over the maximum value) in an intention to emphasize behaviors that are common on word gender variation or verbal inflection. This time, instead of normalizing the maximum value to 100%, the absolute difference is presented, meaning that values close to 1 indicate higher difference between analyzed words and values close to 0 going in the opposite direction. We also make use of Maximum Mean Discrepancy (MMD) test [Gretton *et al.*, 2012], a multivariate two-sample test, to test the null hypothesis that the matrices representing each attention head from both inspected words came from the same distribution. We hypothesize that,

[4]We use "Head *n*" to indicate the n-th head from the Multi-Head Attention mechanism.

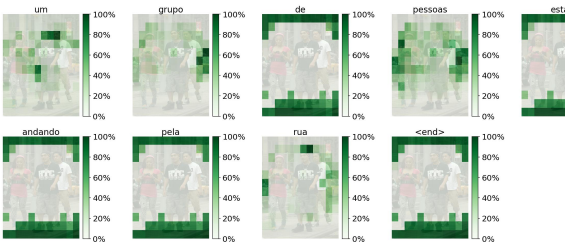


Figure 5. Portuguese sentence generated with Model-PT and the attention weights of Head 1 for each word (translation: "a group of people are walking down the street").

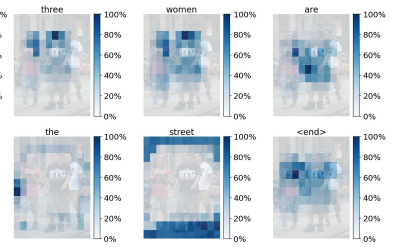
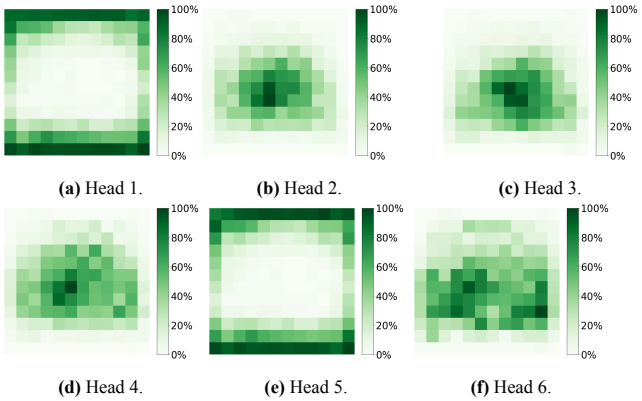


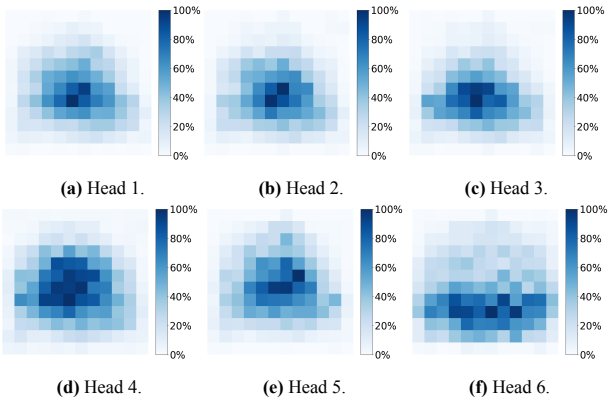
Figure 6. English sentence generated with Model-EN and the attention weights of Head 1 for each word (sentence: "three women are walking down the street").



(a) Head 1. **(b)** Head 2. **(c)** Head 3.

(d) Head 4. **(e)** Head 5. **(f)** Head 6.

Figure 7. Means of each attention heads' weights for the word "andando" (translation: "walking") produced by Model-PT.



(a) Head 1. **(b)** Head 2. **(c)** Head 3.

(d) Head 4. **(e)** Head 5. **(f)** Head 6.

Figure 8. Means of each attention heads' weights for the word "walking" produced by Model-EN.

in cases where words share morphosyntactic behavior, the MDD test will not reject the null hypothesis, i.e., groups of attention heads' matrices have no substantial evidence to come from different distributions.

Figure 7 presents attention heads' behavior for the word "andando", which is demonstrated to be different from what is shown for the same word in English. One possible reason for this might be the two forms of the verb that appear in the dataset ("andando" and "andar"); we explore this possibility with the difference between both of these words average attention heads' weights, as shown in Figure 9. Such Figure also exhibits the mean and the standard deviations of the values in each tile, calling for attention that the most differing Heads (also the ones with the highest mean and standard deviation) are the same ones standing out when comparing Figures 7 and 8, indicating that the verbal inflections "-r" and "-ndo" might be influencing an organization among attention heads to represent such these word endings.

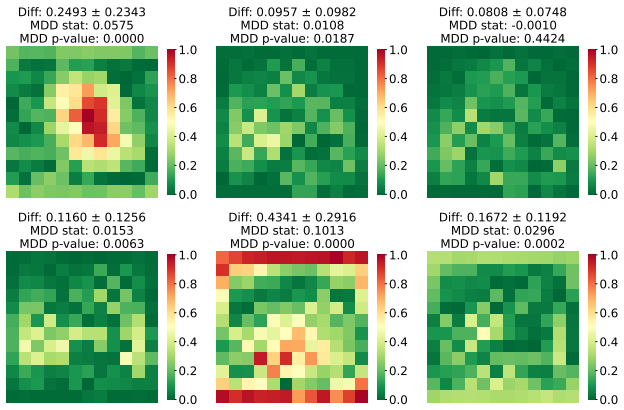


Figure 9. Difference between Attention Heads weights' averages when predicting the words "andando" and "andar".

In the next section we introduce the experiments performed after summarizing attention weights of similar words by their means, which lead to the emergence of observable behaviors, exemplified in Figures 7 and 9 .We make use of these patterns, found during the evaluation step, to sustain that linguistic knowledge is learned by attention heads during training — even if the inputs are images.

5 Experimental setup

In this section, we describe the experiments performed and the dataset employed. Our work evaluates two similar models, each trained with a different dataset, one with English captions and the other with Portuguese captions, to analyze the attention heads' weights. Bilingual captions are comparable on training Image Captioning models and to assess if the arrangements found in the attention heads are language dependent or not. The model used was a naïve Transformer (as depicted in Fig. 2), without using techniques to improve sentence generation.

The first experiment is a comparison of metrics obtained by each trained model; this is done to evaluate if the impacts of training with a translated dataset are an impediment when inspecting the attention weights of the Portuguese model. Secondly, we introduce the analysis concerning the morphosyntactic labels of predicted words and their influence on the cross-modal Multi-Head Attention mechanism.

5.1 Dataset

Our dataset was the Flickr30k [Young *et al*., 2014], which was developed to the task of image description and retrieval. It consists of 31,014 images with a varied number of daily activities, each containing 5 different descriptions to them. To train and evaluate our model, we make use of the train, validation and test splits defined in Karpathy and Fei-Fei [2015], thus using 29,000 images for training, 1,014 for validation and 1,000 for testing the model.

The original dataset (hereinafter named Flickr30k-EN) has an average sentence length of 12.34 tokens with a standard deviation of 5.21. To train the model in this dataset, we maintain in the vocabulary only words that appeared more than 4 times throughout the captions on the training dataset (due to memory constraints on the cluster), resulting in a final vocabulary of 7400 words.

To evaluate the behavior of attention weights in the Portuguese language, each of the 155,070 captions were automatically translated from English with LibreTranslate[5] software. Google Translated or other paid APIs were not considered to ease experiments reproduction and due to cost restrictions. The Portuguese version of the dataset (hereinafter named Flickr30k-PT) has an average sentence length of 12.68 tokens with a standard deviation of 5.45. Similar to the English version, in the Portuguese dataset, we added to the final vocabulary words that appeared at least 5 times, resulting in an 8056-length vocabulary.

Figure 10 shows an example of image from the dataset. The original English captions for this image are:

- A carefully balanced male stands on one foot near a clean ocean beach area.
- A young man performs yoga, inspired by a beautiful day on the beach.
- A man in printed board shorts is doing a yoga pose on the beach.
- A man in swim trunks poses in a yoga stance on the beach.
- A man posing on the beach by the water.

Figure 10. An example from the Flickr30k dataset. Source Young *et al*. [2014].

And the automatic translated captions are shown below:

- Um macho cuidadosamente equilibrado está em um pé perto de uma área de praia de oceano limpo.
- Um jovem executa ioga, inspirado por um belo dia na praia.
- Um homem em calções impressos é fazer uma pose de ioga na praia.
- Um homem em troncos de natação posa em uma postura de ioga na praia.
- Um homem posando na praia junto à água.

By employing automatic translation methods our newly created dataset is susceptible to translation errors. Some translation errors can be seen on the captions listed above. At first, the sentence "A man in printed board shorts is doing a yoga pose on the beach" is translated to "Um homem em calções impressos é fazer uma pose de ioga na praia", here the term "calções impressos" is not the proper way to translate "printed shorts" as it is too literal. Secondly, in the term "is doing" the verbs "is" and "doing" are translated as if they are independent of each other, resulting in "é fazer" which is syntactically wrong. There are also translation errors of nouns, as seen in "swim trunks" being translated to "troncos de natação", this time the word "trunks" is wrongly transcribed to "troncos".

5.2 Experiment 1 - Metrics Analysis

We employed common metrics used for the Image Captioning task to evaluate our bilingual models: BLEU 1-4 [Papineni *et al*., 2002] and METEOR [Banerjee and Lavie, 2005]. The BLEU metric calculates the precision of n-grams (a sequence of n words), penalizing sentences that are too short. Meanwhile, the METEOR metric performs a unigram (1-gram) matching of words, synonyms, or stems of words, favoring the recall of a given model. Even though this work assesses the comparison of attention heads, the use of automatic metrics remains important so we can evaluate how well-trained both models are and, more importantly, if both models can be compared fairly.

5.3 Experiment 2 - Morphosyntactic Analysis

This experiment aims to assess the general behavior of trained Multi-Head Attention heads. The head weights returned by the cross-attention between image features and sentence dependencies are aggregated into a single representation for each word by the centroid of the set, which is evaluated with groups of words from the same parts of speech. This was done to check if words with similar morphosyntactic classes would show similar behaviors according to the attention mechanisms after training. By doing this, we aim to show that attention mechanisms leverage morphosyntactic information during training, similar to what authors have shown in Clark *et al*. [2019], but organizing amongst model's heads to gather information from image features.

In the first part, we inspect the shifts in attention weights when words vary in their gender, either with differing words for each gender (e.g., "man" and "woman"), with words that have grammatical gender assigned in the Portuguese language, such as "homem" (word with male gender in Portuguese, and translated to "man") and "criança" (word with female gender in Portuguese, and translated to "child"), and by inflectional morphemes, e.g., the morpheme "-a" in the Portuguese adjective "amarel**a**" (translation "yellow"). Furthermore, we inspect behavioral differences of attention verbs for divergent

[5] https://github.com/LibreTranslate/LibreTranslate

verb tenses, e.g., “walks” and “walking” or “jogando” (translation “playing” or “throwing”) and “jogar” (translation “to play” or “to throw”).

5.3.1 Gender Variation

According to Young *et al*. [2014], the annotations of the captions from the Flickr30k dataset follow the same instructions from Flickr8k, a smaller predecessor. In the original Flickr8k paper, the authors state that “images in this data set focus on people or animals (mainly dogs) performing some action” [Hodosh *et al*., 2013] and that the annotators were asked to describe the images depicting the scene, situations, events and entities. Given that, words like “woman”, “dog”, “man” or “person” are often the subject and main character of the scenes.

In the Portuguese language, as words are assigned grammatical genders, due to nominal and verbal concordance, all modifiers or lexical units in a syntactical dependence with a word must agree in gender and number to its root, e.g.: in Portuguese, sentences as “the yellow ball” and “the yellow car” are respectively translated to “**a** bol**a** amarel**a**” and “**o** carr**o** amarel**o**”, where the gender of the noun influences the gender of the words that modify it. We argue that the influence of gender assignment might affect the weights distribution of the attention heads for Portuguese words.

5.3.2 Verb Inflection

We also identified common behavior of attention heads weights’ distributions when verbs were predicted by the trained Image Captioning models. The analysis of verbs came to our attention after noticing the different forms a verb might appear in the sentences of the dataset. As seen by the original captions from Figure 10, the verb “to pose” can appear in different forms: “a man posing” and “a man in swim trunks poses”. The same applies to Portuguese with two forms been shown: “um homem em troncos de natação posa” and “um homem posando”. There is also the form “um homem está a posar” (translation: “a man is posing”), also appearing in Portuguese and present in the dataset after automatic translation. Such differences in verbal inflection might also have an impact on attention heads conducts, as we show in the following section.

6 Experimental Results

This section describes our results analyzing both models’ outputs. First, we present the results regarding the evaluation metrics discussed in Section 5 for the models trained on our datasets. Secondly, we start a qualitative analysis with the differences in average weights for gender variations and verb inflection.

We evaluate the results of the trained models with BLEU and METEOR automatic metrics. Since we are using two different datasets, with the Portuguese one created with automatic translation, we considered it an important step to compare the models’ performance on these metrics to check if the following comparisons will not be biased by their capabilities. Table 1 shows the final metrics.

Table 1. BLEU1-4 and METEOR metrics for each trained model.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Model-EN	60.96	43.35	30.07	20.89	19.56
Model-PT	57.80	39.52	26.70	17.97	18.16

Table 1 shows that Model-EN had better values in all metrics, which can be a result influenced by the poor quality captions generated by the automatic translation to the Portuguese language employed in Flickr30k-PT, as discussed in subsection 5.1. However, we observe that the values remain close within all metrics, and changing similarly as the n-grams counting increases for the BLEU metric; we use these behaviors as proxies to enable going further with our analysis and comparisons.

6.1 Gender variation

We start with Nouns, and the first observation concerns the words “homem” (translation “man”) and “mulher” (translation “woman”). This pair is an example of words of opposite grammar categories (gender) that do not differ in their endings but in their root instead. Figure 11 shows the difference between the means of the average head weights of the two words. A first glance at the image shows that the tiles are colored differently when compared to Figure 9, this time with Head 5 having a mean value close to 0. We also observe that the p-values are all smaller than 0.0009, way below a threshold of 0.05.

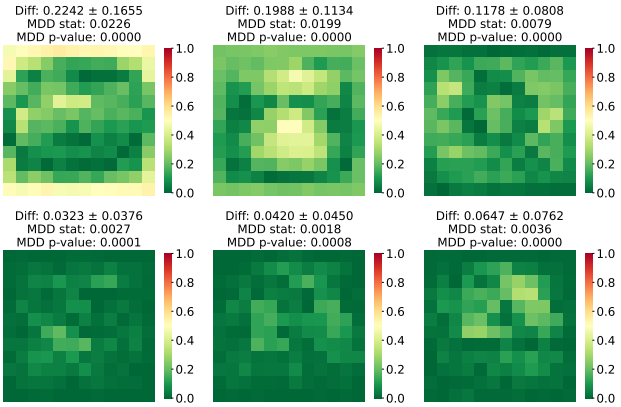


Figure 11. Difference between Attention Heads weights’ averages when predicting the words “homem” (man) and “mulher”(woman).

This functioning is not observed with their English counterparts. As seen in Figure 12, not only do the attention heads show average values closer to 0 for all cases,

Such different behavior is also observed when comparing the differences between “menina” (translation: “girl”) and “rapaz”, Figure 13, (even though “menino” (translation: “boy”) would be a more appropriate gender counterpart to “menina”, Model-PT only predicted it 2 times among 1000 test images, making the average attention weights’ distribution not well suited for analysis, we therefore chose “rapaz” as it can also be translated to “boy”) and “girl” and “boy”, Figure 14. English words, from Figures 12 and 14, are seen not to change much in attention when predicting these words, hence their mean differences give values close to 0. Whereas Portuguese words, from Figures 11 and 13, show differences

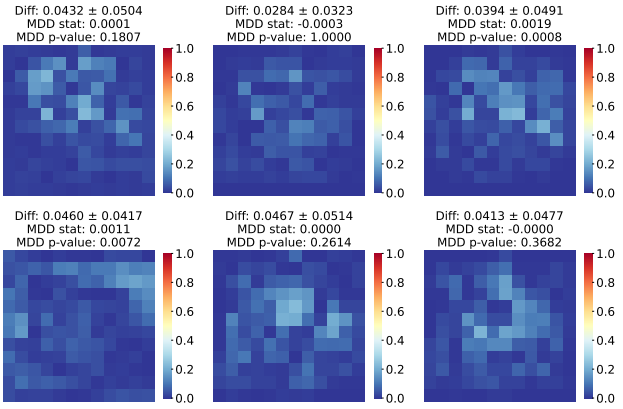


Figure 12. Difference between Attention Heads weights' averages when predicting the words "man" and "woman".

in similar heads, with Head 1 standing out.

Head 1 not only showed a p-value much smaller than 0.05 — as all the others —, it also had the greatest statistics value (0.0398), enforcing how far apart the groups of matrices are. It is also important to notice that none of the p-values were smaller than 0.05 within the English heads' matrices.

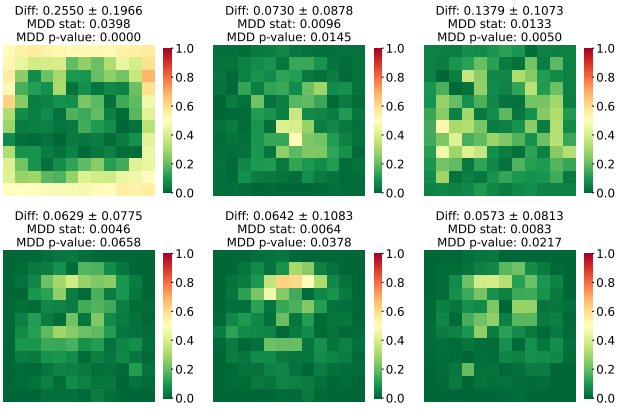


Figure 13. Difference between Attention Heads weights' averages when predicting the words "rapaz" and "menina".

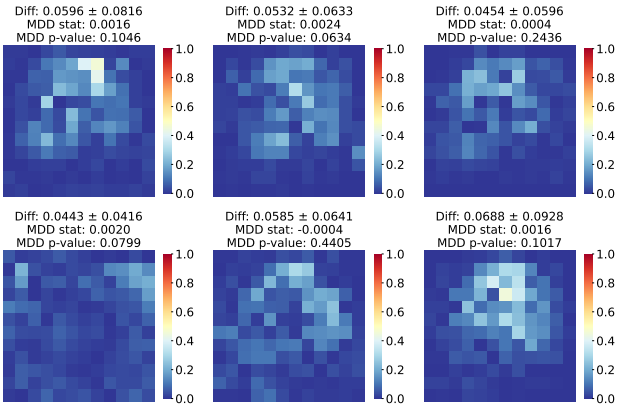


Figure 14. Difference between Attention Heads weights' averages when predicting the words "boy" and "girl".

Since the attention mechanism gives information on the "focus" of the model when predicting the current word, the higher values of differences on the corners of images from Model-PT is odd, as it indicates a common behavior of attention Head 1 when predicting only one of the two words being compared (since the absolute difference is higher in those parts). This singular functioning is present only in the Portuguese words used as Nouns on Flickr30k, and where there is a morphological flexibility concerning the word gender. The gender flexibility is relevant since the same average distribution difference does not show up when comparing "homem" and "cão" (translation "dog"), given that "cão" is a masculine gender word in Portuguese, but it does appear when comparing "homem" and "criança" (translation "child"), where, this time, the word "criança" has a feminine gender in Portuguese. Both comparisons are respectively shown in Figures 15 and 16. We point out that, for words with different meanings, this relevance is not supported by hypothesis test, although they do become present again when analyzing adjectives with nominal inflections in the following paragraphs.

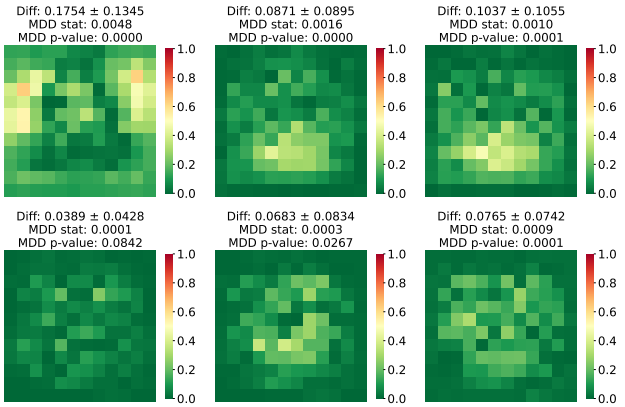


Figure 15. Difference between Attention Heads weights' averages when predicting the words "homem" and "cão".

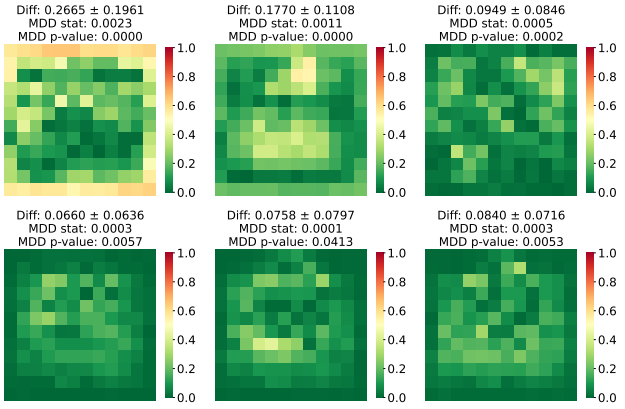


Figure 16. Difference between Attention Heads weights' averages when predicting the words "homem" and "criança".

Following gender variation analysis, authors in Plummer *et al*. [2015] inspected the Flickr30k dataset, presenting the most common words used by the annotators during the dataset creation. According to authors, adjectives relative to colors are frequent. After the most repeated adjective, "young", the most commons are, respectively, "white", "black", "blue"

and “red”. These color adjectives’ Portuguese counterparts, respectively: “branco”, “preto”, “azul” and “vermelho”, have nominal inflection to determine their gender, hence “branc**o**” and “vermelh**o**” respectively refer to white and red male nouns, whereas “branc**a**” and “vermelh**a**” respectively refer to white and red female nouns. Due to this recurring appearance and presence of inflection for gender variation, we inspect the attention heads of words related to colors.

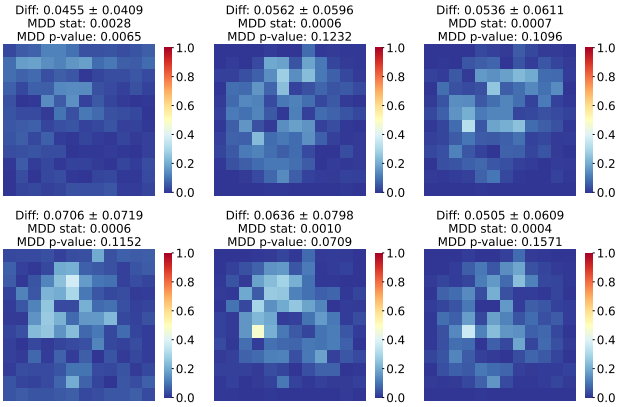


Figure 17. Difference between Attention Heads weights’ averages when predicting the words “black” and “white”.

Figure 17 shows the difference of average head weights for “black” and “white”, exemplifying how similar attention heads usually behave when predicting colors with Model-EN. The related average organization of weights distribution follows the ones presented in Figures 12 and 14, which was expected, as we argued before that the gender variation of nouns might affect attention heads organization, and such inflection is more present in the Portuguese language.

When inspecting the differences of means in Portuguese, there is a need to separate the analysis into feminine and masculine words. The difference between “preta” and “branca” (respectively the feminine translations for “black” and “white”) is shown in Figure 18. The image shows low values for the difference of means, similar to what is discussed in Figure 15, where words with the same gender show analogous attention heads’ organization. However, this time the p-values of the tests were greater than 0.05, five out of six times, reinforcing the similarities on attention heads between these words.

This is sustained with Figures 19 and 20. First the differences between “preto” and “branco” (respectively the masculine translations for “black” and “white”) are presented and, even though the differences is higher for Head 4, the second image with the differences between “branca” and “branco” have higher mean for all heads but the fifth (which has lower difference values). In Figure 19, though the p-values are all below 0.05, we highlight that not only one of them is close to the threshold, but all the statistical values are much higher than each head’s counterparts in Figure 17.

These results indicate that, for Portuguese, grammatical gender information is relevant for attention mechanisms to learn from images, as they may influence the behavior of the model throughout the generation process.

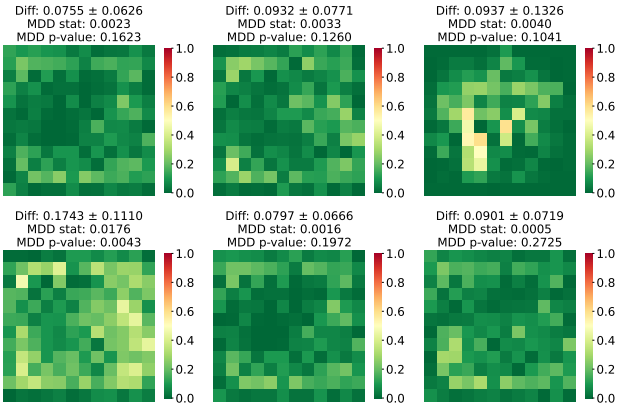


Figure 18. Difference between Attention Heads weights’ averages when predicting the words “preta” and “branca”.

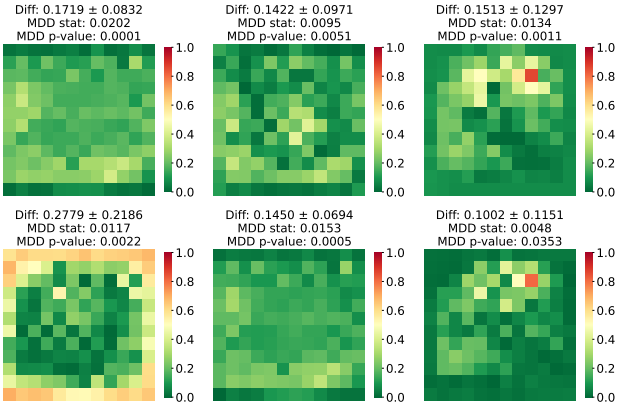


Figure 19. Difference between Attention Heads weights’ averages when predicting the words “preto” and “branco”.

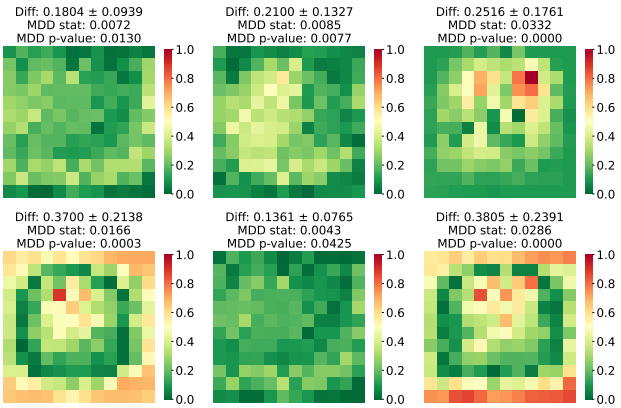
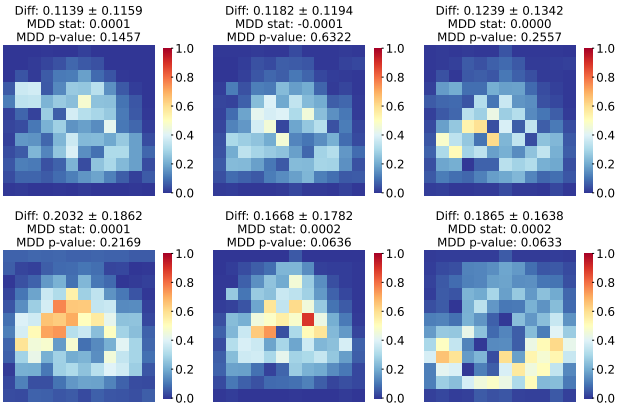


Figure 20. Difference between Attention Heads weights’ averages when predicting the words “branca” and “branco”.

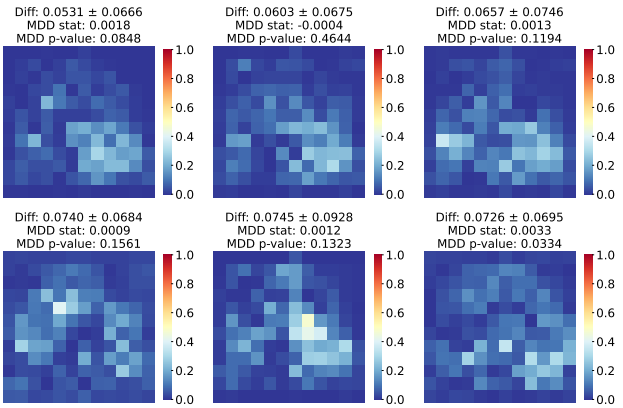
6.2 Verb inflection

Analysis of verbal inflections’ influence on attention heads for Model-EN reveals that, even though being in a lower magnitude, such changes are represented within the organization of attention heads as well. In Figures 21a and 21b it is shown how different and similar verb endings equally have different and similar attention heads.

This behavior corroborates the discussions of gender variations influencing inflections of surrounding words. This time,

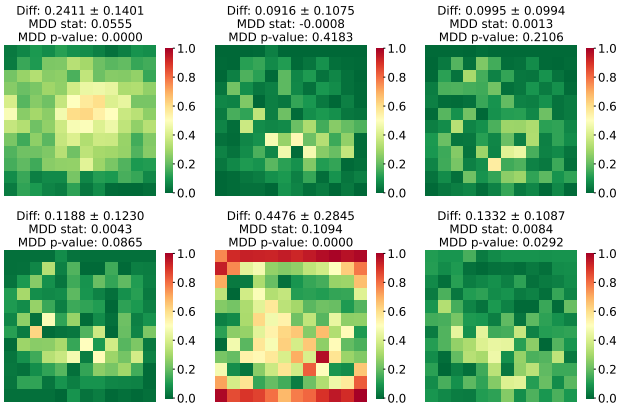


(a) Difference between Attention Heads weights' averages when predicting the words "walking" and "walks".

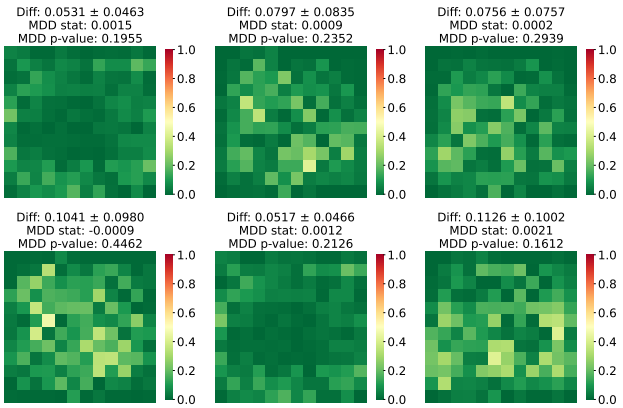


(b) Difference between Attention Heads weights' averages when predicting the words "walking" and "working".

Figure 21. Comparison of attention heads' organization for Model-EN. First, the same verb with different inflections, secondly, two different verbs, but with the same endings.



(a) Difference between Attention Heads weights' averages when predicting the words "jogando" and "jogar".



(b) Difference between Attention Heads weights' averages when predicting the words "jogando" and "andando".

Figure 22. Comparison of attention heads' organization in Portuguese. First, the same verb with different inflections, secondly, two different verbs, but with the same endings.

the inflections might be caused by a variability introduced in the dataset due to different forms of conjugation employed during its annotation.

We highlight that, for English verbs, these claims are not validated by the hypothesis tests made. We argue for the importance of further analysis of attention heads' functioning for English verbs.

Verbal inflections in Portuguese follow the effects on attention heads. In section 5, Figure 9 exemplified how different attention heads can organize when predicting the same verb, but with distinct inflections. Figure 22a shows a very similar behavior for the same setup of inflections: "jogando" and "jogar", respectively "playing" and "(to) play" (the infinitive form).

In this case, Head 5 appears to inform the existence of a verb inflection, repeating itself as observed in Figure 9 both in small p-value (followed by two other heads) and in high statistics value. In Figure 22b, the lower values of differences are present when both words, "andando" and "jogando" (translation: playing or throwing), have the same inflection — indicating gerunds, also corroborated by the p-values obtained with the tests.

7 Discussion

The pattern matching of different words with the same gender and inflections in sentences indicates that, during the training process, attention heads specialize on morphological knowledge, which has already been shown by authors in Clark *et al*. [2019], but not explored for image-to-text models. The anchoring of attention heads on parts of images to indicate linguistic organization raises questions on how learning happens during training and how well the image captioning datasets are being employed. During training, the Transformer is expected to learn the highest probabilities of the next word given the image and the previous predicted words [Cornia *et al*., 2022]; however, morphological attributes are also being learned.

Also, a relevant observation that calls for more attention is that the analysis performed in this work was possible, in Portuguese, due to the automatic translation of Flickr30k. There are significant differences in the learning of Multi-Head Attention Heads for both languages analyzed, and the lack of proper datasets for languages other than English imposes difficulties on probing multilingual Image Captioning models.

The experiments suggest that attention heads are learning inflections and gender variations during training. If so, the learning of highly inflected languages can be directly affected by what the optimal number of attention heads could be, since

the difficulties and amount of knowledge to be learned may vary greatly from English, the default language that most common datasets are created with. In the opposite direction, if a language has a poor morphological structure, adding too many attention heads can cause overfitting. Thus, we argue for the need of multilingual datasets for better understanding the behavior of Multi-Head Attention mechanisms, since the morphosyntactic knowledge learned might be language-specific.

8 Carbon Footprint

To calculate the energy costs of our experiments, we employed the "ML CO2 Impact"[6] tool. The usage of a NVIDIA A100 PCIe 40/80GB, for 10 hours of training, with Google Cloud Platform (through Google Colab) in the "southamerica-east1" Region of Compute, resulted in 0.5 kg of CO_2 emitted.

9 Conclusion and Future Works

In this work, we proposed a methodology for analyzing attention heads for the Image Captioning task and applied it to probe our own trained Transformer. Our experiments suggest that, not only the Multi-Head Attention mechanisms are capable of learning morphological knowledge, but that they also adapt accordingly to the language used during training phase.

Results indicate that the attention heads specialize to learn how to associate morphosyntactic information of words to be predicted with visual information from the input, as shown with the case of gender and verb inflection. Such findings can be a key indicative that more linguistic knowledge from the captions could be explored for better learning of Image Caption models. For example: leveraging the knowledge already present by extracting POS information can be a strategy to facilitate the training of Image Captioning models, as suggested in some works in the literature [He *et al*., 2017; Lu *et al*., 2017; Wang *et al*., 2022a].

Following the work presented in this paper, we intend to go further with the analysis presented, extending it by assessing inflections depending on the previous word predicted or adding POS to the analysis as well. We also intend to expand our model, adding POS tags and other morphosyntactic knowledge present on the captions as metadata input during training, as shown by other works presented above, and verify their influences on the learning of morphological knowledge.

Declarations

Acknowledgements

The authors would like to thank FAPESB TIC 002/2015 - CSIS Project.

Authors' Contributions

João Gondim: Conceptualization, methodology, software, formal analysis, evaluation, writing (original draft, review and editing).

Daniela Claro and Marlo Souza: Conceptualization, evaluation, data curation, visualization, writing and reviewing (original draft, review and editing).

Competing interests

The authors declare no competing interests.

Availability of data and materials

The code materials used in this study are openly available at `https://github.com/FORMAS/ImageCaptioningPT`.

[6] `https://mlco2.github.io/impact/`

References

Aker, A. and Gaizauskas, R. (2010). Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, page 1250–1258, USA. Association for Computational Linguistics. Available at:`https://aclanthology.org/P10-1127.pdf`.

Al-Qatf, M., Hawbani, A., Wang, X., Abdusallam, A., Zhao, L., Alsamhi, S. H., and Curry, E. (2024). NPoSC-A3: A novel part of speech clues-aware adaptive attention mechanism for image captioning. *Engineering Applications of Artificial Intelligence*, 131:107732. DOI: 10.1016/j.engappai.2023.107732.

Banerjee, S. and Lavie, A. (2005). 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics. Available at:`https://aclanthology.org/W05-0909.pdf`.

Barry, A. M. S. (1997). *Visual intelligence: Perception, image, and manipulation in visual communication*. State University of New York Press. Book.

Chen, G., Hou, L., Chen, Y., Dai, W., Shang, L., Jiang, X., Liu, Q., Pan, J., and Wang, W. (2023). mCLIP: Multilingual CLIP via Cross-lingual Transfer. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/2023.acl-long.728.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv. DOI: 10.48550/ARXIV.1406.1078.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/W19-4828.

Cornia, M., Baraldi, L., and Cucchiara, R. (2022). Explaining transformer-based image captioning models: An empirical

analysis. *AI Commun.*, 35(2):111–129. DOI: 10.3233/AIC-210172.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/cvpr42600.2020.01059.

Deshpande, A., Aneja, J., Wang, L., Schwing, A. G., and Forsyth, D. (2019). Fast, Diverse and Accurate Image Captioning Guided by Part-Of-Speech . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10696, Los Alamitos, CA, USA. IEEE Computer Society. DOI: 10.1109/CVPR.2019.01095.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. DOI: 10.48550/arxiv.1810.04805.

dos Santos, G. O., Colombini, E. L., and Avila, S. (2022). #pracegover: A large dataset for image captioning in portuguese. *Data*, 7(2). DOI: 10.3390/data7020013.

Farhadi, A., Hejrati, M., Sadeghi, A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision (ECCV'10)*, volume 6314, pages 15–29. DOI: 10.1007/978-3-642-15561-1_2.

Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170. DOI: 10.1613/jair.5477.

Gautam, T. (2021). Implementation of attention mechanism for caption generation on transformers using tensorflow. Website. Available at:`https://www.analyticsvidhya.com/blog/2021/01/implementation-of-attention-mechanism-for-caption-generation-on-transformers-using-tensorflow/`.

Gondim, J., Claro, D. B., and Souza, M. (2022). Towards image captioning for the portuguese language: Evaluation on a translated dataset. In *Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 384–393. INSTICC, SciTePress. DOI: 10.5220/0011080000003179.

Gretton, A., Borgwardt, K. M., Rasch, M. J., SchÃ¶lkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773. Available at:`https://dl.acm.org/doi/abs/10.5555/2188385.2188410`.

He, X., Shi, B., Bai, X., Xia, G.-S., Zhang, Z., and Dong, W. (2017). Image caption generation with part of speech guidance. *Pattern Recognition Letters*, 119. DOI: 10.1016/j.patrec.2017.10.018.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899. DOI: 10.1613/jair.3994.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36. DOI: 10.1145/3295748.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L. (2021). Scaling up vision-language pre-training for image captioning. arXiv. DOI: 10.48550/ARXIV.2111.12233.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). Attention on attention for image captioning. arXiv. DOI: 10.48550/ARXIV.1908.06954.

Indurkhya, N. and Damerau, F. J. (2010). *Handbook of Natural Language Processing*. Chapman & Hall/CRC, 2nd edition. DOI: 10.1201/9781420085938.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. arXiv. DOI: 10.48550/ARXIV.1412.2306.

Karpathy, A., Joulin, A., and Li, F.-F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. arXiv. DOI: 10.48550/ARXIV.1406.5679.

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., *et al*. (2022). mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*. DOI: 10.18653/v1/2022.emnlp-main.488.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. arXiv. DOI: 10.48550/ARXIV.2004.06165.

Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., and Elliott, D. (2021). Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485. DOI: 10.18653/v1/2021.emnlp-main.818.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. DOI: 10.1109/cvpr.2017.345.

Moriarty, S. E. (2002). The symbiotics of semiotics and visual communication. *Journal of Visual Literacy*, 22(1):19–28. DOI: 10.1080/23796529.2002.11674579.

Nain, A. K. (2021). Image captioning. Website. Available at:`https://keras.io/examples/vision/image_captioning/`.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*. Available at:`https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html`.

Pan, J.-Y., Yang, H.-J., Duygulu, P., and Faloutsos, C. (2004). Automatic image captioning. volume 3, pages 1987–1990 Vol.3. DOI: 10.1109/ICME.2004.1394652.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k en-

tities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649. DOI: 10.1109/ICCV.2015.303.

Reale-Nosei, G., Amador-DomÃnguez, E., and Serrano, E. (2024). From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. *Medical Image Analysis*, 97:103264. DOI: 10.1016/j.media.2024.103264.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2016). Self-critical sequence training for image captioning. arXiv. DOI: 10.48550/ARXIV.1612.00563.

Rosa, G. M., Bonifacio, L. H., Souza, L. R. d., Lotufo, R., and Nogueira, R. (2021). A cost-benefit analysis of cross-lingual transfer methods. arXiv. DOI: 10.48550/arXiv.2105.06813.

Salgotra, G., Abrol, P., and Selwal, A. (2024). A Survey on Automatic Image Captioning Approaches: Contemporary Trends and Future Perspectives. *Archives of Computational Methods in Engineering*. DOI: 10.1007/s11831-024-10190-8.

Santos, G. O. d., Moreira, D. A. B., Ferreira, A. I., Silva, J., Pereira, L., Bueno, P., Sousa, T., Maia, H., Silva, N. D., Colombini, E., Pedrini, H., and Avila, S. (2023). CAPIVARA: Cost-Efficient Approach for Improving Multilingual CLIP Performance on Low-Resource Languages. arXiv. DOI: 10.48550/arXiv.2310.13683.

Sharif, N., Nadeem, U., Shah, S., Bennamoun, M., and Liu, W. (2020). *Vision to Language: Methods, Metrics and Datasets*, pages 9–62. Springer International Publishing. DOI: 10.1007/978-3-030-49724-8_2.

Sharma, H. and Padha, D. (2023). A comprehensive survey on image captioning: From handcrafted to deep learning-based techniques, a taxonomy and open research issues. *Artificial Intelligence Review*, 56(11):13619–13661. DOI: 10.1007/s10462-023-10488-2.

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., and Cucchiara, R. (2023). From Show to Tell: A Survey on Deep Learning-Based Image Captioning . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(01):539–559. DOI: 10.1109/TPAMI.2022.3148210.

Tan, M. and Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. DOI: 10.48550/arxiv.2104.00298.

Tensorflow (2022). Image captioning with visual attention. Website. Available at:`https://www.tensorflow.org/tutorials/text/image_captioning`.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arxiv.1706.03762.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. arXiv. DOI: 10.48550/ARXIV.1906.05714.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. DOI: 10.1109/cvpr.2015.7298935.

Wang, D., Liu, B., Zhou, Y., Liu, M., Liu, P., and Yao, R. (2022a). Separate syntax and semantics: Part-of-speech-guided transformer for image captioning. *Applied Sciences*, 12(23). DOI: 10.3390/app122311875.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022b). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052. Available at:`https://proceedings.mlr.press/v162/wang22al.html`.

Web Accessibility Initiative (2022). Introduction to web accessibility. Available at:`https://www.w3.org/WAI/fundamentals/accessibility-intro/`.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2016). Show, attend and tell: Neural image caption generation with visual attention. Available at:`https://proceedings.mlr.press/v37/xuc15.html`.

Yang, Y., Teo, C., Daumé III, H., and Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Edinburgh, Scotland, UK. Association for Computational Linguistics. Available at:`https://aclanthology.org/D11-1041.pdf`.

Yao, B. Z., Yang, X., Lin, L., Lee, M. W., and Zhu, S.-C. (2010). I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508. DOI: 10.1109/JPROC.2010.2050411.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78. DOI: 10.1162/tacl_a_00166.

Zhang, J., Mei, K., Zheng, Y., and Fan, J. (2021a). Integrating part of speech guidance for image captioning. *IEEE Transactions on Multimedia*, 23:92–104. DOI: 10.1109/TMM.2020.2976552.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021b). Vinvl: Revisiting visual representations in vision-language models. DOI: 10.48550/ARXIV.2101.00529.