How Vision Affects Language: Comparing Masked Self-Attention in Uni-Modal and Multi-Modal Transformer

Nikolai Ilinykh **Simon Dobnik**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg, Sweden
{nikolai.ilinykh,simon.dobnik}@gu.se

Abstract

The problem of interpretation of knowledge learned by multi-head self-attention in transformers has been one of the central questions in NLP. However, a lot of work mainly focused on models trained for uni-modal tasks, e.g. machine translation. In this paper, we examine masked self-attention in a multi-modal transformer trained for the task of image captioning. In particular, we test whether the multi-modality of the task objective affects the learned attention patterns. Our visualisations of masked self-attention demonstrate that (i) it can learn general linguistic knowledge of the textual input, and (ii) its attention patterns incorporate artefacts from visual modality even though it has never accessed it directly. We compare our transformer's attention patterns with masked attention in distilgpt-2 tested for uni-modal text generation of image captions. Based on the maps of extracted attention weights, we argue that masked self-attention in image captioning transformer seems to be enhanced with semantic knowledge from images, exemplifying joint language-and-vision information in its attention patterns.

1 Introduction

Recently, we have seen a surge of interest in explainability research for large-scale neural networks, e.g. transformers (Vaswani et al., 2017). A lot of the existing literature focuses on the analysis of attention (Bahdanau et al., 2015) in terms of linguistic knowledge it encodes (Belinkov and Glass, 2019). Clark et al. (2019) show that attention heads' patterns in BERT (Devlin et al., 2019) resemble syntactic dependencies present in the text. They also use a probing classifier to identify how knowledge of syntax is distributed between attention heads. Vig and Belinkov (2019); Hoover et al. (2020) have shown that visualising the structure of attention in transformer models can help us see which parts of the model capture specific syntactic knowledge. Voita et al. (2019) demonstrate that not all attention heads are equally suitable for learning syntactic information. Thus, pruning such heads can be an option to reduce the model's complexity. While attention is not always an explanation (Jain and Wallace, 2019), some work (Ravishankar et al., 2021) has shown that extra fine-tuning on a syntax-related task can guide the model's attention to truly resemble syntactic information about the text. Other approaches to the model's interpretability include, for example, a work by Rethmeier et al. (2020), which inspects how knowledge is transferred on the neuron level rather than attention level.

While most of the existing research has placed the problem of model's explainability in the context of **uni-modal** text-based tasks, e.g. machine translation, the field of language-and-vision is somewhat lacking similar analysis for models trained to solve **multi-modal** tasks. This becomes especially important with the increasing interest in adopting transformers for learning better cross-modal representations (Tan and Bansal, 2019). In addition, using large-scale models to improve grounding between language and vision representations (Lu et al., 2019) requires vigilance regarding how information is learned in different parts of such densely structured models. Multi-modal transformers are required to not only learn to perform *symbol grounding*, e.g. mapping natural language symbols into visual representations as defined by Harnad (1990) and a language model, but also learn *to fuse information* from two modalities, the nature of which has been an open question in the field (Lu et al., 2017; Caglayan et al., 2019; Ilinykh and Dobnik, 2020). The effect that such multi-modal representations have on the attention in large-scale models has not been addressed a lot in the language-and-vision literature. More specifically, we need a

Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR), pages 45–55
June 16, 2021. ©2021 Association for Computational Linguistics

better understanding of how self-attention in transformer processes the multi-modal information.

In this paper, we analyse the masked self-attention part of the image captioning transformer, which performs a standard language masking task based on the textual input, and compare its attention patterns with masked attention in distilgpt-2, a text-only transformer. Our goal is to identify what kind of knowledge is captured in representations learned by this part of the model and whether it is affected in any way by the visual modality, which is not directly accessible for this particular self-attention. We aim to answer the following questions:

- Does masked self-attention show patterns which resemble any syntactic knowledge of the input text?
- What are the differences in attention on previous words when generating the next word in either the uni-modal or multi-modal task set-up?
- What is the task's effect (uni-modal vs. multi-modal) on the semantics of words captured by masked-self attention in image captioning transformer?

In addressing these questions, we believe that we show novel insights into how the information is transferred between inner self-attentions of complex architectures such as a transformer and how representations from specific components of such models are affected by the training objective and multi-modality.

2 Model

Fig. 1 shows the architecture of the image captioning transformer that we use for our experiments, first introduced by Herdade et al. (2019) and built on top of the basic image captioning transformer (Luo et al., 2018). This architecture resembles many parts of the classic transformer (Vaswani et al., 2017), which was initially introduced for machine translation, consisting of three multi-head self-attention mechanisms. The standard transformer's encoder learns representations of the input text by passing it through two sub-layers: multi-head self-attention and feed-forward network. Each sub-layer has a residual connection around itself, followed by layer-normalisation operation. The decoder contains masked self-attention, which is used to learn linguistic knowledge of the ground-truth target translation. In a uni-directional task, it masks the words in the future so that the model learns to attend to the previously generated words only. The third self-attention is performing a cross-modelling task, using information from both encoder and decoder. This cross self-attention identifies correlations between the source text and currently generated target text in a machine translation context.

Once we reformulate the model's task from machine translation to image captioning (Fig. 1), we naturally change the encoder's inputs. Instead of

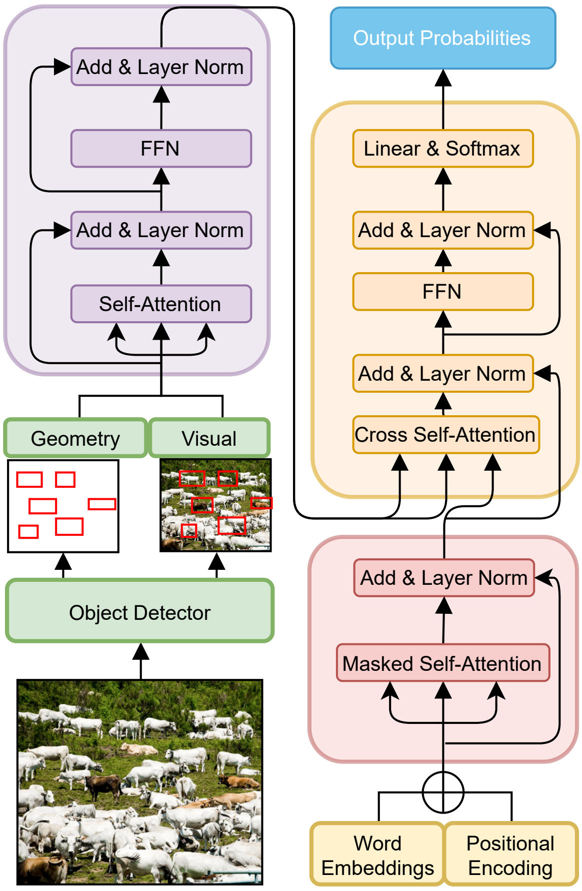


Figure 1: Object relation image captioning transformer. The image is first passed through a pre-trained object detector to extract visual and geometric features. The left side self-attention (image encoder) consists of attention heads, where each of them utilises both visual and geometry information. On the right side, the masked self-attention (text encoder) is given the embeddings of the caption words and their positional information. The words are fed to the text encoder in an auto-regressive manner, e.g. one word at a time plus all the preceding words. The cross self-attention uses keys K and values V from the visual encoder, while queries Q are coming from the textual encoder and finally predicts the output probabilities of the next word.

the source sentence, the encoder uses representations of objects from the image as its input. On the decoder's side, the ground-truth captions that the model learns to generate are used as inputs during training. To prepare inputs to our encoder, we first extract visual features of the detected objects $X = \{x_1, ..., x_N\}$, where $x_n \in \mathbb{R}^{1\times D}$ with $N = 36$ and $D = 2048$. We use a bottom-up feature extractor (Anderson et al., 2018), which is based on Faster-RCNN (Ren et al., 2015) and pretrained on Visual Genome (Krishna et al., 2016) with the ResNet-101 as its backbone (He et al., 2016). For each detected object we also extract geometry features $G = (x, y, w, h)$ (centre coordinates, width, height). In the next step, queries $Q = W^Q X$ and keys $K = W^K X$ are used to get scaled dot product Ω^V:

$$\Omega^V = \frac{QK^T}{\sqrt{d_k}} \quad (1)$$

Then, Ω^V and geometric features G are combined, taking into account the displacement between the objects and producing a fused representation Ω.[1] Finally, each attention head h from each encoder layer l outputs a combination of values V and geometry-aware visual features Ω:

$$\text{head}_{l,h} = \text{self-attention}(Q, K, V) = \Omega V \quad (2)$$

Masked self-attention in the decoder The idea of self-attention is that each token from the input text learns to attend to the other tokens from the same sequence. However, this is not feasible for the caption generation task since attending to the future tokens is unfair and it cannot be used when generating text. Therefore, the self-attention in the decoder is using masking of future tokens to keep the auto-regressive nature of the model. In particular, the token w_t and the future tokens $w_{t+1}, ..., w_W$ are replaced with $[MASK]$. Then, w_t is predicted using the previous context in the standard left-to-right fashion: $W_{\setminus t} := (w_1, ..., w_{t-1})$.

We have specifically focused on the analysis of the attention weights in the **decoder's masked self-attention** of the image captioning transformer. We extract the attention weights for each head h in each layer l of this self-attention and use them for our visualisations and analysis. These weights are calculated similarly to the attended visual features (Eq. 1). Our masked self-attention has six layers, consisting of eight heads in each of them.

For the model checkpoint, we use the best model released by the authors of the architecture[2]. This checkpoint has been chosen on the basis of automatic evaluation scores: the model uses bottom-up representation of images, geometry features and self-critical training (Rennie et al., 2017). The captions are generated using beam search with beam width $bw = 5$ in the standard auto-regressive manner.

3 Learning syntactic knowledge

In our first experiment we investigate whether the attention weights of the masked self-attention are able to capture any general syntactic knowledge about the input text. It has been shown that the multi-head attention patterns in the transformer trained for the task of machine translation resembles syntactic properties of language at the level of part-of-speech tags and syntactic dependencies (Mareček and Rosa, 2019; Ravishankar et al., 2021). Since the self-attention that we are focused on is trained in a very similar task (masked language modelling), we first explore if particular layers and heads attend to specific part-of-speech tags the most. Then, we continue with the analysis of how information about syntactic dependencies is reflected in the learned attention patterns.

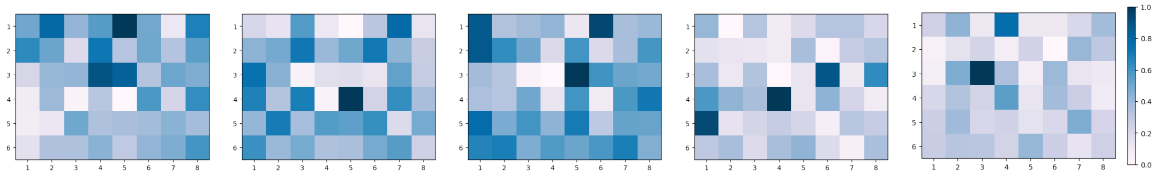
Attention on Part-of-Speech We follow Vig and Belinkov (2019) who compute the proportion of attention from each head that this head pays to tokens of a particular part-of-speech tag and accumulate the results over our test set:

$$P(\alpha|tag) = \frac{\sum_{s\in S}\sum_{i=1}^{|s|}\sum_{j=1}^{i}\alpha(s_i, s_{j,pos(j)=tag})}{\sum_{s\in S}\sum_{i=1}^{|s|}\sum_{j=1}^{i}\alpha(s_i, s_j)} \quad (3)$$

where S is the corpus of generated captions, tag is the part-of-speech tag of the attended word, and $\alpha(s_i, s_j)$ is the attention from i^{th} word to j^{th} word for the given head. We use Spacy (Honnibal et al., 2020) to get part-of-speech tags of words and syntactic dependencies between them for all our experiments. We also perform normalisation (linear scaling) on the values of the calculated attention proportion to place all values in a single scale from 0 to 1. The masked self-attention is always given

[1] For more details on how geometric information is combined with visual features in this model, we refer the reader to Herdade et al. (2019).

[2] Available at: `https://github.com/yahoo/object_relation_transformer`



(a) Determiner (b) Noun (c) Adjective (d) Verb (e) Adposition

Figure 2: Each heat-map demonstrates the proportion of attention targeted towards a word of a specific part of speech. Vertical and horizontal axes indicate layers and heads respectively.

the START token at the start of the generation. We consider attention on this token non-informative (as it is over-attended) and ignore the corresponding attention weights for better visualisations. The heads pay only ∼26% of their attention to the START token on average per caption. We use BertViz tool (Vig, 2019) to produce our visualisations.

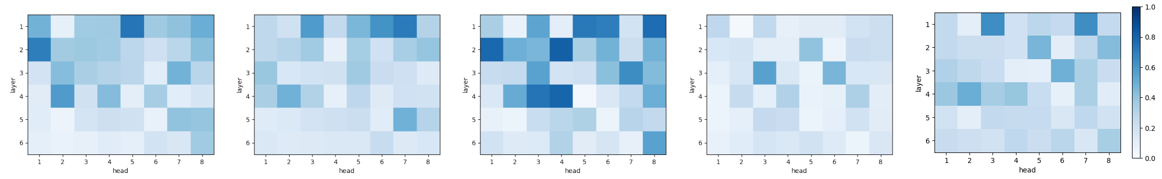
The results for the five most frequently occurring part-of-speech tags (more than 1000 individual instances) are shown in Fig. 2. Words of such part-of-speech tags, which can be grounded in visual signals (nouns for objects, adjectives for attributes), receive attention from a large number of attention heads. On the other hand, only specific heads focus on words describing relations (verbs, adpositions). Specifically, seventeen heads (out of forty-eight) put more than 40% of their attention to the nouns, while only three heads give more than 30% of their attention to the verbs.

We also find supporting evidence for the previous studies (Belinkov, 2018; Vig and Belinkov, 2019), showing that deeper layers focus on more complex properties, e.g. relational part-of-speech tags (verbs), which require knowledge of objects learned from earlier layers (nouns). For example, the top 3 attention heads that attend to basic parts-of-speech such as determiners are all located in the model's first three layers. For adjectives, the top 3 heads are similarly located in the first three layers of the model, with the maximum value of the attention head being 0.25. However, attention on adjectives is more spread across many heads in different layers, with the attention value being 0.14 for more than half of the heads, which is also a mean value for attention on this part-of-speech tag. A less clustered pattern is observed for nouns: its top 3 heads are located in layers 1, 3, and 4, with thirty-three heads paying more than 30% of their attention to nouns. We argue that the reason why the attention on nouns is scattered over many heads, with most of them paying nearly one-third of their attention to the nouns, is because nouns are continuously required for caption generation: the model needs to take them into account when generating either a relation or an attribute.

Somewhat differently, verbs are attended mostly in the model's deeper layers: the top 3 most attentive heads are located in layers 3, 4, and 5 with values higher than 0.3. The vast majority of the heads (forty-three) have smaller attention values (less than 0.2), indicating that the model needs verbs only for specific situations, for example when a relationship needs to be generated. Overall, our visualisations demonstrate that masked self-attention weights resemble task-specific syntactic information about part-of-speech tags. For example, nouns are similarly attended across all heads since they are required for the captioning task the most (to describe, refer to, use in phrases, etc.). In contrast, more function-dependent parts of speech (verbs, adpositions) are attended to by fewer heads in the deeper layers of the model.

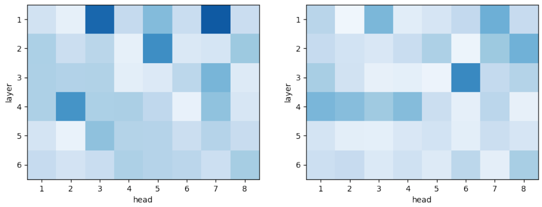
Attention on Syntactic Dependencies Fig. 3 shows the proportion of attention from the heads in masked self-attention for the most frequently occurring syntactic dependency relations. The proportions are calculated similarly to Eq. 3. In particular, we used the attention weights from root to the non-root part of the dependency phrase or vice versa, extracting dependencies in advance. This choice was affected by the auto-regressive nature of the generation task: for each word, we could only inspect attention focus on previous words. The attention on different dependencies seems to be distributed similarly to the attention on part-of-speech tags. More specifically, attention heads from the surface layers (1:5 and 2:1[3]) seem to be focused on the determiner in the det relation. Comparing heat-maps of attention distribution on part-of-speech and syntactic dependency may give us intuition

[3] We use layer: head notation.



(a) det (b) compound (c) nummod (d) pobj (e) prep

Figure 3: Attention distribution on different constituents of the specific syntactic dependencies. For det, compound, nummod we visualise which heads look the most on the non-root element of the dependency (e.g. "man" → "a" in "a man"). For pobj and prep we show attention in a different direction (e.g. "table" → "on" in "on table", "with" → "bathroom" in "bathroom with").



(a) Head = Noun (b) Head = Verb

Figure 4: Attention distribution for the prep syntactic dependency. The left-side heat-map is computed for phrases where noun is a head in the phrase ("kitchen with", while for the right-side heat-map it is the verb ("sitting at").

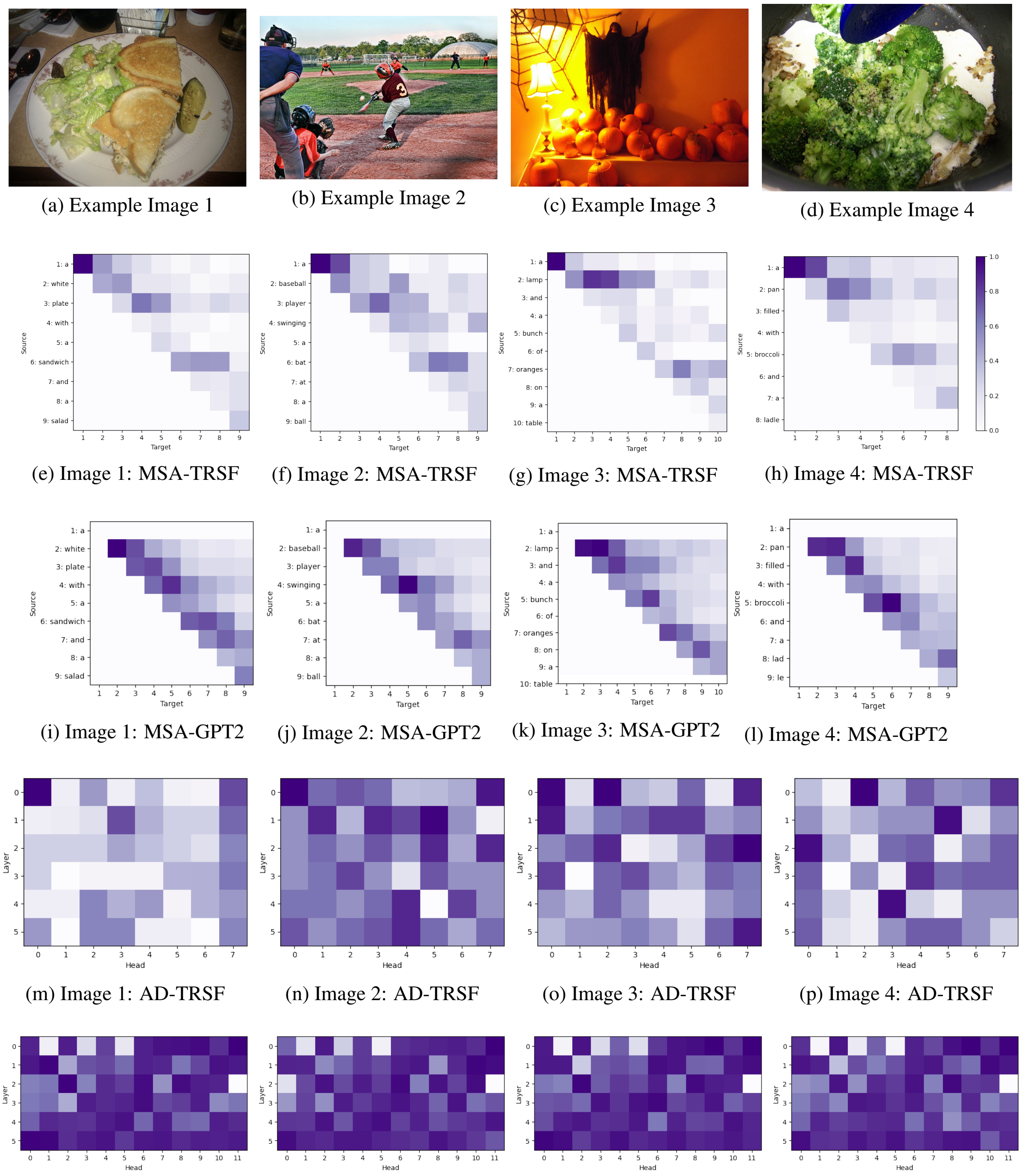
about the specific heads' role. For example, the heads 3:4 or 3:5 are not intensely active for the det relation, although they are among the most active heads when attending to the determiners. This indicates that these heads 1:5 and 2:1 may be more responsible for focusing on determiners when the phrase in the det relation is generated. Interestingly, many heads strongly attend to the numeral in the nummod dependency compared to all other relations. This could be related to the importance of learning about the number of objects in the scene, while other, simpler noun-based dependencies (det, compound) do not have to be attended so strongly.

Only a few heads specialise in dependencies that capture more complex properties (e.g. relations between different objects), with heads 3:3 and 3:6 being the most attending heads for pobj. The root of the prep phrase is often attended in the first layer, with only a few more heads in the later layers being activated. *Could this pattern be mapped with the fact that roots in these phrases are often nouns and verbs?* Fig. 4 shows that heads 1:3 and 1:7 are the most active heads when a noun is a root in the phrase of prep dependency. Same heads in the first layer are also active the most when looking at the nouns, according to Fig. 2b. This indicates that the model acquires basic knowledge of language syntax (dependencies, part-of-speech information) in its first layers. Similarly, as Fig. 4b demonstrates, the head 3:6 is the single most active head for the prep dependency. At the same time, according to Fig. 2d, this particular head is one of the few most active heads when the attention focus is on verbs. This might be interpreted as if this head is better at learning information about syntactic dependencies than other activated heads. We argue that it is helpful to look at the correspondence between attention on parts-of-speech and syntactic dependency since it is informative when determining specific heads' roles and how important they are for different language tasks, e.g., part-of-speech tagging and syntactic dependency identification.

4 Multi-modality and masked self-attention

In this section, we look at how a multi-modal task of image captioning affects attention on the previous words when a masked self-attention model predicts the next word. We also compare our model's attention patterns with patterns from an auto-regressive model, distilgpt-2 (Radford et al., 2019), which has been pre-trained on OpenWebTextCorpus. This model has 6 layers with 12 heads in each layer, which makes it more comparable to our captioning transformer than the standard GPT-2 model with 12 heads in each of the 12 layers.

Semantics of Attention Patterns Here, we compare the text-only uni-modal language model and its attention patterns with our multi-modal transformer's masked self-attention. We do this because we want to investigate to what extent the attention patterns produced by the language model in the



(a) Example Image 1 (b) Example Image 2 (c) Example Image 3 (d) Example Image 4

(e) Image 1: MSA-TRSF (f) Image 2: MSA-TRSF (g) Image 3: MSA-TRSF (h) Image 4: MSA-TRSF

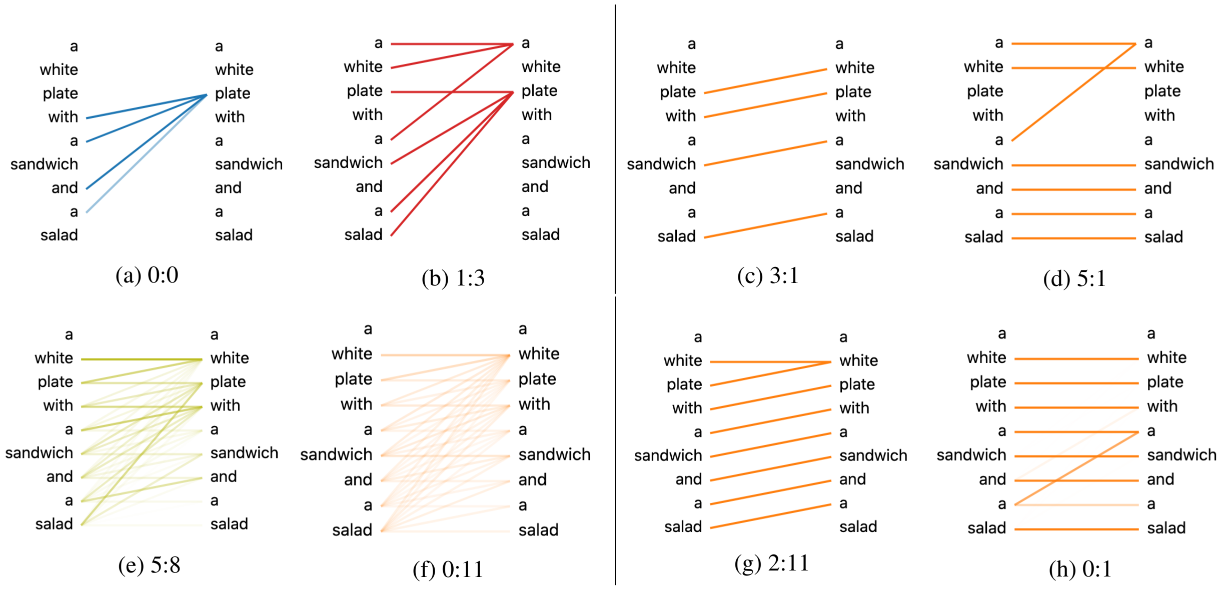
(i) Image 1: MSA-GPT2 (j) Image 2: MSA-GPT2 (k) Image 3: MSA-GPT2 (l) Image 4: MSA-GPT2

(m) Image 1: AD-TRSF (n) Image 2: AD-TRSF (o) Image 3: AD-TRSF (p) Image 4: AD-TRSF

(q) Image 1: AD-GPT2 (r) Image 2: AD-GPT2 (s) Image 3: AD-GPT2 (t) Image 4: AD-GPT2

Figure 5: Here are several examples of different attention visualisations for masked-self attention (**MSA**) from our image captioning transformer (**TRSF**) and distilgpt-2 (**GPT2**). **The top row** shows example images for which we generate a caption. **The second and third rows** show attention on the available context (indicated by the *Source* axis) when generating the next word (the *Target* axis). Word of the generated caption are displayed on the Source axis. To get more fine-grained visualisations in the third row, we exclude attention on the first token of each sentence for distilgpt-2 attention patterns since, based on our experiments and literature (Vig and Belinkov, 2019), attention on the first token is always very strong and not relevant. **The fourth and the fifth rows** show attention dispersion (**AD**) for each head in each layer. The colour bar in the second row indicates the range of values in all visualisations in this figure.

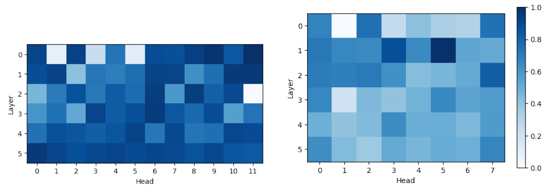
multi-modal setting differ from patterns where the task is uni-modal. For this, we run `distilgpt-2` (Radford et al., 2019) on the captions generated by our image captioning transformer, where both


a white plate with a sandwich and a salad
a white plate with a sandwich and a salad

(a) 0:0 (b) 1:3 (c) 3:1 (d) 5:1

(e) 5:8 (f) 0:11 (g) 2:11 (h) 0:1

Figure 6: Visualisation of attention for example attention heads. The first row shows heads from the masked self-attention in our transformer; the second row depicts the head's attention from distilgpt-2. The side to the left of the vertical line in the middle includes heads with **high entropy** in either of the models, while the right side contains heads with **low entropy**. The heads are denoted by a layer:head notation; they can be traced back to the more general attention concentration in Fig. 5m and Fig. 5q. Each figure displays attention from **target (left)** to **source (right)**.



(a) distilgpt-2 (b) Image Captioning Transformer

Figure 7: Mean normalised entropy of attention per head / layer calculated for the set of generated captions.

input and target are the same image descriptions. This way, we receive two sets of masked self-attention weights for the same texts from two models trained for different tasks. Both our decoder and the distilgpt-2 model are trained for the masked language modelling task; therefore, these models' attention is comparable with each other. We save the model's attention weights similar to how we did it for our captioning transformer's masked self-attention. The first three rows in Fig. 5 show visualisations of both models' attention for several captions and if applicable the corresponding images. Attention in our masked self-attention tends to focus on nouns much more than on other parts of the source (context). In comparison, distilgpt-2 patterns are more diagonal: every next word is focused on its surroundings the most, and the attention does not generally look at a single word for too long.

We believe that this is an artefact of the training for image captioning task: our masked self-attention learns to focus on nouns because they ground objects, and most of the time, the following words form a single phrase referring to these objects. For example, attention on "lamp" for the third image is very strong throughout the generation of the whole phrase "lamp and a bunch of". Once a new object is introduced ("oranges"), the attention shifts to this object for a different phrase ("oranges on a table"). The visualisations show that captioning transformer's masked self-attention learns global, phrase-based semantic features of sentences. In contrast, in the text-only setting, the model learns about local relations between words in a sentence. For example, distilgpt-2 continuously shifts its maximum attention after every 2-3 words are generated, indicating that it learns to capture local relations between words ("bunch of", "oranges on").

Attention Focus As demonstrated by Fig. 5, attention can constantly focus on particular words (e.g., nouns) while the caption is generated. We seek to identify which attention heads are responsible for such observed patterns in the masked self-attention of the image captioning transformer. This is potentially important for reducing the model's complexity by pruning non-important heads, which do not have an interpretable role defined by the measure of choice. Therefore, we calculate the entropy of attention distribution (Ghader and Monz, 2017) and use it as the measure of dispersion between attention weights:

$$Ent_{\alpha}(s_j) = -\sum_{i=1}^{|s|} \alpha(s_i, s_j)\, log(\alpha(s_i, s_j)) \quad (4)$$

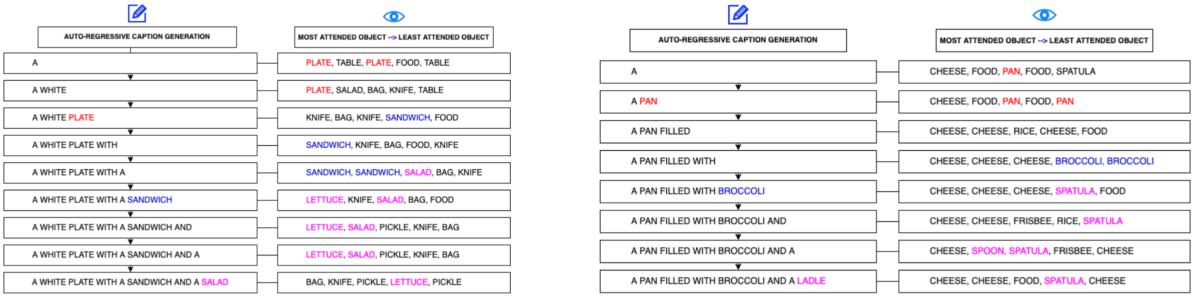
As Fig. 7a demonstrates, many heads in distilgpt-2 have high entropy scores which means that attention here is highly dispersed. The entropy increases in the deeper layers of the model. This correlates with the fact that deeper layers capture more distant syntactic relations and, therefore, lead to higher entropy scores (Vig and Belinkov, 2019). Fig. 7b shows the entropy scores for attention heads in captioning transformer's masked self-attention. Here, most heads have a relatively low entropy, with only some of them with higher entropy in the model's first layers.

Do heads have high/low entropy? Based on the examples of attention heads from Fig. 6, we can conclude that high entropy reflects a stronger concentration of attention from target words on *particular* source words to learn *specific* information. Such pattern can be observed, for example for captioning transformer's masked self-attention heads in Figures 6a and 6b. Note that these heads heavily link several words with nouns (e.g. "plate"), which increases the head's entropy - many words in the target sentence attend to a single word from the context. Another important observation is that the attention distribution from target to source is not always strong: not every word on the left side has a connecting line to the right side, indicating that attention is used to learn only specific properties. For example, as Fig. 6b demonstrates, focusing on "plate" when other objects ("sandwich", "salad") are mentioned may indicate that the model learns the notion of scene structure reflected in the text. At the same time, Fig. 6a shows that focusing on "plate" can be required when generating relations between objects, e.g. "plate *with* a sandwich *and* a salad". However, as figures 6e and 6f demonstrate, distilgpt2 learns somewhat different attention between the source and the target words. While these patterns demonstrate that many words in the target sequence tend to focus on the specific words from context, each attention connection is not as strong as for the heads of the captioning transformer's masked self-attention. The distilgpt-2 model does not focus on the caption's specific relations or properties. Instead, it learns weak attention between all words. The heads' entropy is high as the attention is dispersed, but each attention connection's is also *not as strong as* it is in the captioning transformer's masked self-attention. The examples of heads with low entropy (the right side of the Fig. 6) indicate that there is a word in the context that will be attended for each generated word.

5 Attention Alignment

It may be the case that the observed differences in attention patterns discussed in the previous section are simply due to different frequencies of words (in particular nouns) in the dataset on which the models are trained. For example, the multi-modal decoder also attends on the closest syntactic relations in the same way as a uni-modal decoder, but these happen to be nouns simply because there are more nouns in image captions. To test this hypothesis we calculated the Pearson correlation coefficient between the frequency of the nouns in our captions and attention distribution on the context words attended by heads when the next word is produced. The test has not shown a statistically significant correlation between the frequency of the nouns versus attention distribution on the context words in multi-modal decoder's self-attention ($r = 0.49, p = 0.056$). However, we observed a moderate positive correlation between the frequency of the nouns versus attention distribution in the uni-modal decoder's attention ($r = 0.60, p = 0.014$). These differences in correlations show that the uni-modal architecture is more biased to frequencies, whereas in a multi-modal setting, the effect of noun frequency is diminished. This provides support to our hypothesis, namely that this bias towards nouns is coming from somewhere else, e.g. **the multi-modal representations that the language model is grounded in**.

Since the model's parameters are jointly updated with an end-to-end training through back-propagation, representations learned by different self-attention mechanisms are expected to be *aligned* with each other. We present a small preliminary analysis of whether the attention weights in


AUTO-REGRESSIVE CAPTION GENERATION
MOST ATTENDED OBJECT --> LEAST ATTENDED OBJECT
A
PLATE, TABLE, PLATE, FOOD, TABLE
A WHITE
PLATE, SALAD, BAG, KNIFE, TABLE
A WHITE PLATE
KNIFE, BAG, KNIFE, SANDWICH, FOOD
A WHITE PLATE WITH
SANDWICH, KNIFE, BAG, FOOD, KNIFE
A WHITE PLATE WITH A
SANDWICH, SANDWICH, SALAD, BAG, KNIFE
A WHITE PLATE WITH A SANDWICH
LETTUCE, KNIFE, SALAD, BAG, FOOD
A WHITE PLATE WITH A SANDWICH AND
LETTUCE, SALAD, PICKLE, KNIFE, BAG
A WHITE PLATE WITH A SANDWICH AND A
LETTUCE, SALAD, PICKLE, KNIFE, BAG
A WHITE PLATE WITH A SANDWICH AND A SALAD
BAG, KNIFE, PICKLE, LETTUCE, PICKLE
AUTO-REGRESSIVE CAPTION GENERATION
MOST ATTENDED OBJECT --> LEAST ATTENDED OBJECT
A
CHEESE, FOOD, PAN, FOOD, SPATULA
A PAN
CHEESE, FOOD, PAN, FOOD, PAN
A PAN FILLED
CHEESE, CHEESE, RICE, CHEESE, FOOD
A PAN FILLED WITH
CHEESE, CHEESE, CHEESE, BROCCOLI, BROCCOLI
A PAN FILLED WITH BROCCOLI
CHEESE, CHEESE, CHEESE, SPATULA, FOOD
A PAN FILLED WITH BROCCOLI AND
CHEESE, CHEESE, FRISBEE, RICE, SPATULA
A PAN FILLED WITH BROCCOLI AND A
CHEESE, SPOON, SPATULA, FRISBEE, CHEESE
A PAN FILLED WITH BROCCOLI AND A LADLE
CHEESE, CHEESE, FOOD, SPATULA, CHEESE

(a) Cross-modal attention on objects for Fig. 5a.

(b) Cross-modal attention on objects for Fig. 5d.

Figure 8: Attention shifts in cross-modal attention. The left-side column of each sub-figure shows the generated caption one word at a time. The right-side column depicts the labels of the 5 most attend objects in images when generating each word.

the cross-modal self-attention (cross self-attention from Fig. 1) are responsible for information fusion between image encoder and text decoder. Our hypothesis is as follows: if cross-modal self-attention pays a significant portion of attention to the objects, which are generated as nouns in the caption as content words, we can conclude that due to the learning objective and nature of the information flow within the model's components, decoder's self-attention *aligns* with a higher-level cross-modal self-attention. In this case, we also expect that for every non-content word (e.g., determiner, preposition), the cross-attention keeps its attention on the most recent content word similar to what we observe for decoder's self-attention in Figs. 5e–5h. We use two example images and examine the differences among the top 5 most-attended objects for every word generated in image captions. We use the predicted labels from the feature extractor (Anderson et al., 2018) to refer to the detected objects. Fig. 8 shows changes in cross-modal attention on objects during generation of descriptions. From Fig. 8a we can see that every time a new content word is generated ("plate", "sandwich", "salad"), the cross-modal attention tends to focus on objects with labels that are similar to the generated content words. For example, "lettuce" and "salad" are among the most attended objects when the transformer is preparing to generate the content word "salad". Also, the same objects are continued to be attended when other non-content words are generated. This example provides initial evidence how text generation of nouns as exemplified by the decoder's attention is linked to multi-modal representations as exemplified by cross-modal attention on objects. The results suggest that in multi-modal settings models learn representations that are fused and aligned with each other. Since the self-attention in the uni-modal architecture only needs to generate the text one word at a time by taking into account only previously generated words, it learns a pattern over local syntactic dependencies. In our future work, we would like to provide a more detailed analysis of the cross-modal attention and the uni-modal visual attention and therefore further strengthen the arguments how multi-modality affects knowledge that different parts in the large scale transformer models learn.

6 Conclusion

We have shown that attention patterns learned by a sentence decoder module of a multi-modal transformer are highly affected by the task that the model is optimised for. We focused on the masked self-attention in a sentence decoder in an image captioning transformer, demonstrating that its attention weights resemble linguistic knowledge, which is affected by the task of image captioning. This indicates that such language model acquired important aspects of grounded semantics. Simultaneously, we show that that it is important to be cautious when applying large-scale pre-trained models on specific tasks to different semantic tasks as the original task does have an impact on the semantic representations learned. Our future work will focus on further examination of self-attention in the other two components of the multi-modal models which will give us an even clearer picture on what representations are learned by them.

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council

(VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. We thank anonymous reviewers whose comments helped improve this manuscript.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yonatan Belinkov. 2018. *On internal language representations in deep learning: an analysis of machine translation and speech recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. 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)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2020. When an image tells a story: The role of visual and semantic information for generating paragraph descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*.

J. Lu, C. Xiong, D. Parikh, and R. Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

R. Luo, Brian L. Price, S. Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

David Mareček and Rudolf Rosa. 2019. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. 2021. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.

Nils Rethmeier, Vageesh Kumar Saxena, and Isabelle Augenstein. 2020. Tx-ray: Quantifying and explaining model-knowledge transfer in (un-)supervised nlp. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 440–449. PMLR.

Hao Tan and Mohit Bansal. 2019. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.