Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer

Anonymous ACL submission

Abstract

We explore how a *multi-modal transformer* trained for generation of longer image descriptions learns syntactic and semantic representations about entities and relations grounded in objects at the level of masked self-attention (text generation) and cross-modal attention (information fusion). We observe that cross-attention learns the visual grounding of noun phrases into objects and high-level semantic information about spatial relations, while text-to-text attention captures low-level syntactic knowledge between words. This concludes that language models in a multi-modal task learn different semantic information about objects and relations cross-modally and uni-modally (text-only). Our code is available here: [the GitHub link placeholder].

1 Introduction

In this paper, we examine what kind of knowledge is encoded in the multi-modal transformer. Existing work has mostly looked at the knowledge captured in models that operate with **a single modality** (text). For instance, previous research has shown that the attention weights in large-scale models, e.g. BERT (Devlin et al., 2019), implicitly encode knowledge of sentence structure (Raganato and Tiedemann, 2018; Ravishankar et al., 2021), part-of-speech tags, syntactic dependencies (Clark et al., 2019; Vig and Belinkov, 2019), subject-verb agreement between words (Goldberg, 2019), and even information about textual co-reference (Tenney et al., 2019). Only a few papers have inspected what is captured by **multi-modal architectures**. Cao et al. (2020) demonstrate that the attention heads in image-and-text transformers effectively encode linguistic and cross-modal knowledge. Ilinykh and Dobnik (2021) provide the analysis of how language representations are indirectly affected by visual information in language-and-vision model.

Here we inspect what the model learns about two types of words in the multi-modal setting: (i) words denoting objects in the scene (e.g. "a red chair"), (ii) words depicting spatial relations between objects (e.g. "a chair *next to* the table"). While it is relatively simple to associate nouns with specific image regions, words describing relations are much harder to ground (Lu et al., 2017), possibly because visual representations are typically designed to capture objects without any explicit knowledge of relations. Such mismatch between visual features and relations could also be a valid reason to focus on language modality when generating relations (Ghanimifard and Dobnik, 2019). Ideally, each word type should be grounded in both modalities, but to a different degree. *The main challenge* is to develop such architectures that appropriately balance this information, and, therefore, we investigate grounding of different semantic types and test the following hypotheses:

- **H1**: Does attention across two modalities learn visually grounded semantics of nouns?
- **H2**: What is syntactic knowledge encoded in attention on text in the multi-modal set-up?
- **H3**: What does cross-modal attention learn about grounded semantics of spatial relations?

We use a two-stream multi-modal transformer (Herdade et al., 2019), which first attends to each modality independently and then learns to attend cross-modally. This architecture uses rich relative geometry between objects, while many other two-stream models (Tan and Bansal, 2019; Lu et al., 2019) simply use either coordinates of bounding boxes or their spatial location. We train the model for *image paragraph generation* (Ilinykh et al., 2019; Krause et al., 2017), allowing examination of the knowledge of semantic types in extensive contexts. We believe that our experiments show how language and vision are bridged in the multi-modal transformer. In addition, our work provides insights into the shortcomings of how multi-modal representations are learned for different word types.

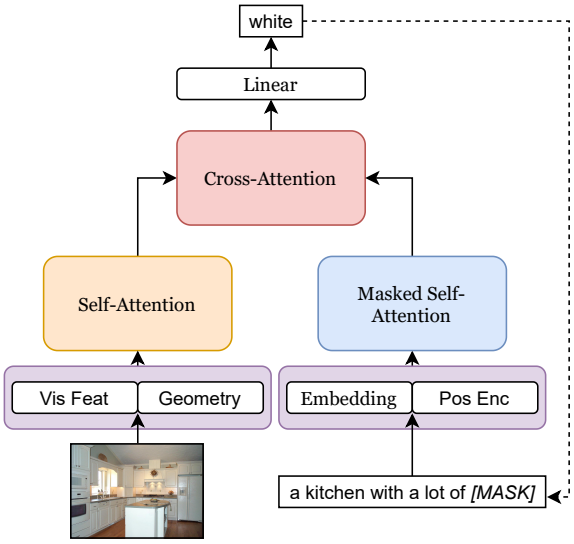


Figure 1: Multi-modal image description transformer. Every next generated word is concatenated with the previously generated words and passed to the model to output the next word prediction.

Figure 2: **Ground-truth description of the image**: It's a room with a bar on the side. There is a pink couch in the center. There's a coffee table in front of the couch. It has a light purple rug. There are three chairs at the bar.
Generated description of the image: This appears to be a bonus room that is red and white. There is a wooden table in the center of the room. There is a red couch. There is a large plant in the corner.

2 Experimental Set-Up

Model We train a multi-modal transformer for image paragraph generation. The model is based on the image captioning transformer proposed by Herdade et al. (2019)[1]. We use the object detector provided by Anderson et al. (2018)[2]. This model comes pre-trained on object annotations from Visual Genome (Krishna et al., 2016). We extract features of N objects per image, resulting in the set $\mathbf{V} = \{v_1, \ldots, v_N\}$ with $v_n \in \mathbb{R}^{1\times D}$. We set $N = 36$ and $D = 2048$. The object extractor also provides us with labels ("table") and attributes ("round") for the objects, which will be used in our experiments. Following Herdade et al. (2019), we also extract geometry information about each object $\mathbf{G} = \langle x, y, w, h\rangle$ (centre coordinates, width, height) and use it as an additional input along with visual features. Figure 1 describes the architecture of the model. In this model, each attention mechanism consists of six layers with eight attention heads in them. The *image encoder* (orange box) learns to combine visual and geometric features[3] and passes them through the standard self-attention block, consisting of multi-head self-attention, feed-forward network, residual connections and layer-normalisation.

Due to uni-directional nature of description generation, the *text decoder* (blue box) produces representation of the current token w_i, based on perviously generated tokens $(w_1, \ldots, w_{i-1})$, while $(w_{i+1}, \ldots, w_W)$ are replaced with $[MASK]$. Finally, the *cross-attention* (red box) uses information from both textual and visual streams to output a probability of the next word in the sequence.

Dataset We train our model on Tell-Me-More (Ilinykh et al., 2019), the dataset of multi-sentence descriptions of real-world images of rooms in the house setting (Zhou et al., 2017). Figure 2 shows an example of the ground-truth text and generated paragraph. For training, we use train and extra splits, providing us with 4820 image-sequence pairs, while for validation and testing we use 441 and 441 pairs respectively. We use beam search to generate sequences with beam width $bw = 2$. The model is trained with standard cross-entropy loss. The best model's checkpoint is chosen based on the highest CIDEr score (Vedantam et al., 2015) for the test set after training for 100 epochs. As Table 1 shows, our model achieves higher scores across most of the standard automatic metrics compared to the baseline architecture (CNN + LSTM + LSTM). Although our transformer performs slightly worse in terms of CIDEr score, note that different from previous work on multi-sentence image description generation (Krause et al., 2017; Chatterjee

[1] https://github.com/yahoo/object_relation_transformer
[2] https://github.com/peteanderson80/bottom-up-attention
[3] For more information on how image encoder employs both visual and geometric information, we refer the reader to the original implementation by Herdade et al. (2019).

Model Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	WMD
CNN+LSTM+LSTM (Ilinykh and Dobnik, 2020)	25.10	13.88	8.11	4.61	11.30	**26.38**	7.61
Multi-Modal Transformer (this paper)	**39.68**	**24.12**	**14.71**	**8.33**	**14.97**	17.54	**8.66**

Table 1: Automatic evaluation of image paragraphs generated by two different model architectures.

and Schwing, 2018; Ilinykh and Dobnik, 2020), we do not restrict the model to generate a specific number of sentences, instead stopping the generation when either the END token is encountered or the maximum number of words has been generated ($W = 150$). In addition, our dataset is much smaller than the Stanford image paragraph dataset (Krause et al., 2017), that the first model has been trained on.

3 Methods and Metrics

We extract the attention weights from both cross-modal attention and masked self-attention. We are not focused on description generation but on examining the attention that particular words and objects would receive. Therefore, we use the ground-truth descriptions where the words are known and test the model on these descriptions in the teacher-forcing setting. Both input and target sequence of tokens that the model sees are identical, which allows us to minimize any noise in the attention weights and inspect them as if our model is an expert in generating coherent texts. For every generated word w_i, the attention weight α per head h in each layer ℓ is extracted. In transformers the attention weights are computed as the scaled dot-product of the query matrix Q with all the keys in K followed by a $\mathrm{softmax}$ operation. These weights are focusing on either previously generated words (masked self-attention *MSA*, Equation 1) or image objects (cross-attention *CA*, Equation 2).

$$\alpha_{\ell,h}(w_i \mid w_1, \ldots, w_{i-1}) = \mathrm{softmax}(\frac{Q_{MSA} K_{MSA}^T}{\sqrt{d_k}}) \quad (1)$$

$$\alpha_{\ell,h}(w_i \mid v_1, \ldots, v_N) = \mathrm{softmax}(\frac{Q_{CA} K_{CA}^T}{\sqrt{d_k}}) \quad (2)$$

We inspect how much attention is focused on specific parts of the input sequence when particular parts of the target sequence are generated. We refer to this measure as the **attention focus** or **attention proportion**. In our experiments, we calculate the proportion of total attention from a specific head that is focused on specific parts of the source sequence, e.g. previously generated words or image objects. Attention proportions are generally calculated as follows:

$$\boldsymbol{P}_{\ell,h}(\alpha \mid S, T) = \frac{\sum_{u \in \mathbf{U}} \sum_{i=1}^{|S|} \sum_{t=1}^{|T|} \alpha(s_i, S | t_j, T)}{\sum_{u \in \mathbf{U}} \sum_{i=1}^{|S|} \sum_{t=1}^{|T|} \alpha(s_i, t_j, T)}, \quad (3)$$

where $\boldsymbol{P}_{\ell,h}$ is the attention proportion for a specific head, S and T are the specific conditions imposed on the source and target sequences unique for every experiment (described below), $\mathbf{U}$ is the set of image descriptions sequences, t_j is the text span for either a noun phrase or relation from the target (generated) sequence T, s_i is the particular object or a text span from the source sequence S.

Conditions on $\boldsymbol{P}$ for H1 For our experiments on visual grounding in cross-modal attention, T limits the target sequence to the text span of a noun phrase, while S defines the ground-truth object that this noun phrase depicts. The attention proportion is calculated by computing the accumulated attention weight from the words in the noun phrase towards the corresponding object and then divided by the overall attention on all objects attended when this noun phrase is generated. We use spaCy (Honnibal et al., 2020) to extract noun phrases from image paragraphs. We skip any phrases which contain at least one word from the list specified in Appendix B. We keep determiners and adjectives in the noun phrases and any numerals if they occur. Some of the paragraphs might contain noun phrases that cannot be grounded in the bounding boxes in the image; either because the bounding boxes are not identified or because the noun phrases refer to abstract concepts. These phrases typically contain words such as "room", "image" or "photo" and are generally placed at the beginning of the description (e.g., "the image is of a kitchen with ..."). In future experiments, we plan to investigate how the model grounds general descriptions of the scene ("the nursery room").

Conditions on $\boldsymbol{P}$ for H2 For the experiments on word-to-word attention, T is set to the generated word at the specific timestep t_j, while S accumulates attention on words of specified part-of-speech tags when the target word t_j is generated. We use

two sets of part-of-speech tags, which reflect the semantic difference between words. The first set contains determiners, adjectives and nouns, while the second set restricts attention to verbs and adpositions.

Conditions on P for H3 To examine grounding of spatial relations, both S and T are determined based on the set of static spatial relations extracted from the texts. We extracted $target - relation - landmark$ relations, based on the annotation schema described in Kolomiyets et al. (2013) and publicly available tool[4]. We obtained 1015 relations of *region* type ("clothes on hangers"), 239 relations of *direction* type ("a gold chandelier above the table"), and 6 relations of *distance* type ("a large vase in the middle of the table"). Each of these relations consists of three spatial elements: a target (*a cup*), a landmark (*a table*) and a relation (*on*) in "a cup on the table". Given that the word order describing relations is typically a $target - relation - landmark$ sequence, the attention proportion for masked self-attention can be extracted only in following directions: $relation \rightarrow target$, $landmark \rightarrow relation$, $landmark \rightarrow target$, and $landmark \rightarrow target + relation$. For example, a possible T could restrict currently generated word to $relation$ (typically expressed with adposition), while S could limit the calculation of the attention focus to $target$ (expressed as a noun phrase) in case of $relation \rightarrow target$ experiment.

4 Linking Nouns and Objects

To inspect attention heads for visual grounding, we require ground-truth annotations of correct linking between image objects and noun phrases. We construct such links automatically using semantic similarity between noun phrases and object labels provided by the object feature extractor. First, we use spaCy (Honnibal et al., 2020) and extract noun phrases on different levels of nesting. For example, a noun chunk "a window with white lace curtains" and the nested chunk "white lace curtains" are identified as two different noun phrases. Potentially, this design choice allows for more accurate linking between noun phrases focusing on different objects ("window" and "curtains") and corresponding fine-grained object detections. In addition, noun phrases with specific details potentially disambiguate linking when multiple objects of the same type are in the image, e.g., several windows. As for object labels, for every detected object in every image, we take the predicted label and its attribute if the extractor's confidence for this attribute is higher than 0.1.

[4]https://github.com/mmxgn/sprl-spacy

Combination Method	*Measure*	*mAP@K*	*Acc*
GloVe Multiply	cos	0.095	13.78
GloVe Add	cos	0.276	41.84
BERTScore	F_1	0.232	41.84
Sentence Transformer	cos	**0.313**	**44.39**

Table 2: Results of the search for the optimal method of linking noun phrases and object descriptions.

Noun phrases and object descriptions typically include multiple words, Therefore, we compute semantic similarity between phrases. We examine several methods for linking noun phrases and object descriptions and compare them against the small subset of image paragraphs with manually annotated linking. Specifically, we randomly sample ten image-text pairs, consisting of 196 detected noun phrases. Then, 158 noun phrases were manually linked with image objects by the first author. The subset of the remaining 38 noun phrases included pronouns and abstract descriptions, too ambiguous to be linked with the specific object in the scene. In addition, we found that some noun phrases describe either a non-detected object or were extracted by mistake. A fraction of noun phrases that were not linked with any object is shown in Appendix A.

Table 2 shows the results of our search for the best linking method. We use GloVe embeddings (Pennington et al., 2014) to represent each word in a phrase and combine them by either element-wise multiplication (*GloVe Multiply*) or addition (*GloVe Add*), inspired by methods for phrase meaning representation (Mitchell and Lapata, 2008). The resulting vectors for a noun phrase and object description were compared based on cosine similarity. For *BERTScore* we follow Zhang et al. (2020) and use contextual word embeddings (Devlin et al., 2019) to represent every word. Words in a noun phrase and object description are then matched against each other by cosine similarity, and the F1 score can be used to examine the similarity. Finally, for *Sentence Transformer* we represent each word with the embedding from Sentence Transformer (Reimers and Gurevych, 2019). This model fine-tunes BERT embeddings for numerous NLI tasks and applies a mean pooling operation to get the fixed-size vector representing embedding of

a whole phrase. We report accuracy *Acc* against manual annotations of ten image-text pairs. We also compute mean average precision *mAP@K*, a metric that allows us to see whether a particular combination method generally rates relevant object descriptions more similar to a noun phrase:

$$AP@K = \sum_{k=1}^{m} P_k(R_k - R_{k-1}), \qquad (4)$$

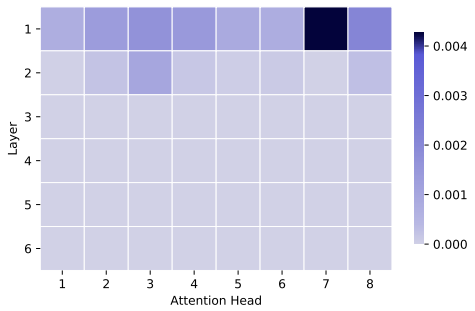
where P_k and R_k are the precision and recall at cut-off k, m is the number of noun phrases detected in an image paragraph. K is set to the number of objects (36) since we inspect the linking of noun phrases with the whole set of objects. The final *mAP@K* score is the mean of average precisions for noun phrases in descriptions of images. Our search results for the linking method demonstrate that using embeddings from Sentence Transformer and comparing them for cosine similarity performs the best in terms of both metrics. Interestingly, simply using BERT embeddings and match them for similarity (*BERTScore*) is not enough to achieve a high *mAP@K* score, and this method also performs worse than a simple addition of non-contextualised embeddings (*GloVe Add*). A more complex fusion of information from different words is required to represent a phrase. When examining attention heads for visual grounding of nouns and relations, we thus use the best performing linking method (*Sentence Transformer*). Noun phrases might describe a group of objects in the scene ("six chairs"), corresponding to multiple object detections (several chairs). Labels of such objects are often identical, which makes their cosine similarity scores also identical. Therefore, we link a noun phrase with multiple objects on top of the similarity ranking if they have the same cosine score. Otherwise, a noun phrase is linked with the object that is ranked the highest.

5 Experiments and Results

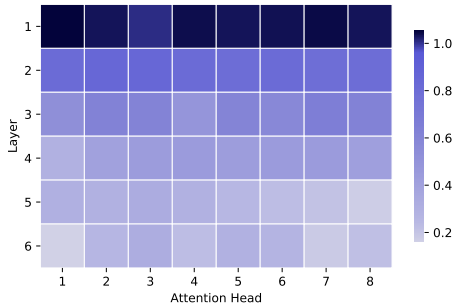
Attention Entropy We compute entropy of the attention weights in both modules for each attention head. Specifically, the entropy E of an attention head h in layer ℓ is defined as follows:

$$E_{\ell,h}(t_j) = -\sum_{i=1}^{|S|} \alpha(s_i, t_j) \log(\alpha(s_i, t_j)) \qquad (5)$$

where s_i and t_j are specific source and target sequence items and α is the attention weight between



(a) MSA entropy



(b) CA entropy

Figure 3: Normalised entropy of attention heads in different layers for masked-self attention (MSA) and cross-attention (CA). The darker the colour, the higher the entropy. All values were normalised by the maximum achievable entropy $-log_2(O)$.

them. As Figure 3 shows, the entropy pattern is similar across both attention modules. Attention heads have lower entropy in deeper layers, focusing more on specific parts of the source sequence. In contrast, earlier layers scatter attention across many items (either objects or previously generated words). Intuitively, such progressive increase of attention focus from lower to higher levels indicate that both modules first learn to generalise over low-level features, gradually moving to capture more specialised, high-level conceptual knowledge (Ullman, 1984). Here, a fair question to ask is *what kind of low-level and high-level knowledge do masked and cross-modal attention learn in different layers with different entropy?*

As Ghader and Monz (2017) show for the task of machine translation, lower attention entropy is mainly observed when looking at nouns and adjectives, while higher entropy is witnessed when attending to adpositions and verbs. This finding demonstrates that attending to nouns in purely textual syntactic dependencies is less complex than focusing on verbs. In the context of our task, ad-

positions and verbs would be used when generating spatial relations, while objects are described with nouns and adjectives. Learning nouns in a multi-modal setting implies their visual grounding, a more complex task that requires knowledge of the scene. Similarly, in general, understanding spatial relations is a much more sophisticated task for the multi-modal transformer. It requires higher-level semantic knowledge and identification of objects and relations, compared to simple attention on verbs and adpositions as part-of-speech tags in a uni-modal setting. It has also been shown that attention on highly complex phenomena (named entities) would happen in deeper layers of the model, while low-level constructs (determiners) are attended much earlier in the layers of both uni-modal (Vig and Belinkov, 2019), and multi-modal (Ilinykh and Dobnik, 2021) architectures. Therefore, in our experiments, we examine how attention heads in different layers of masked and cross-modal attention capture either **syntactic knowledge** (nouns and relation phrases as words) or **semantic information** (visually grounded nouns and spatial relations).

Visual Grounding in Cross-Attention Here we investigate whether the high focus of cross-attention heads in deeper layers can be attributed to their specialisation in visual grounding of nouns. Specifically, *based on the linking method*, we compute the proportion of attention that radiates from words in a noun phrase towards **corresponding objects** described by this noun phrase. Figure 4 shows the results. We can see that attention heads in deeper layers concentrate on linking bounding boxes of detected objects with noun phrases that describe them when these phrases are generated. Specifically, while in the first layer, attention heads pay on average 16% of their attention to the linked objects, in the deeper layers, the average attention focus reaches 29%. The best attention head is the second head in the sixth layer, which places 33% of its attention on connecting noun phrases with the bounding boxes of objects linked with this phrase. These findings show that the model captures complex visually grounded semantics of nouns in deeper layers of cross-attention. In addition, lower entropy observed in these layers (Figure 3b) also indicates that deeper heads are strongly focused and specialised in grounding of nouns.

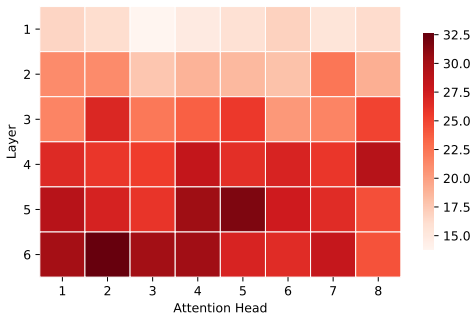
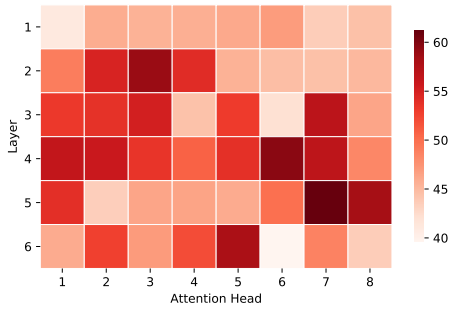
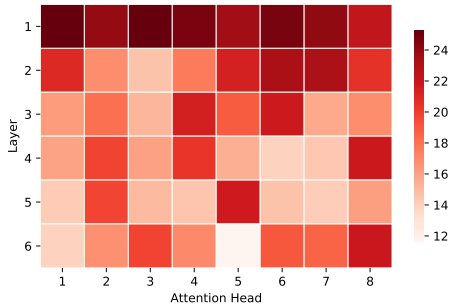


Figure 4: Attention proportions $\boldsymbol{P}$ on correct noun-object pairs (as determined by linking) for each attention head in the cross-modal attention. The darker the colour, the **bigger** the proportion. The proportions are averaged over the noun phrases in descriptions.



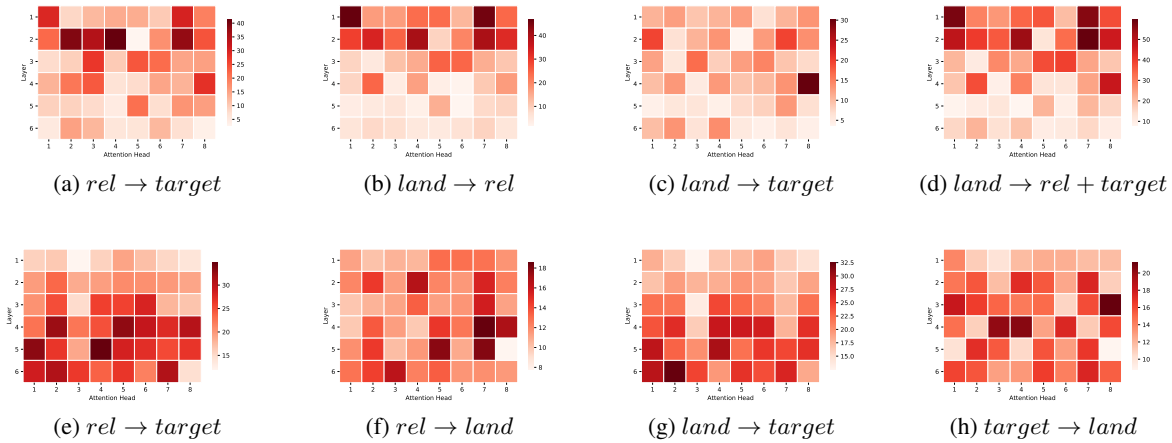
(a) $\boldsymbol{P}$ on DET, ADJ, NOUN



(b) $\boldsymbol{P}$ on VERB, ADP

Figure 5: Attention proportions on words of specific part-of-speech tags for every head in the masked self-attention module. The proportions are averaged over the samples in the test set.

Masked Self-Attention on Specific Part-of-Speech Tags Figure 5 demonstrates the attention focus on previously generated words of specific POS tags. We separate between tags which either describe objects ⟨DET, ADJ, NOUN⟩ or relations ⟨VERB, ADP⟩. Based on the heatmaps, we can see that previously generated determiners, adjectives and nouns are more attended in all layers



(a) $rel \rightarrow target$ (b) $land \rightarrow rel$ (c) $land \rightarrow target$ (d) $land \rightarrow rel + target$

(e) $rel \rightarrow target$ (f) $rel \rightarrow land$ (g) $land \rightarrow target$ (h) $target \rightarrow land$

Figure 6: Heatmap visualisations of $\boldsymbol{P}$ for masked self-attention (**the top row**) and cross-attention (**the bottom row**) for different possible configurations of attention between words constituting spatial relations. All attention proportions are normalised by the number of spatial relations in the test set.

except the first one, in which the focus is on relation part-of-speech tags. At the same time, according to Figure 3a, the attention in the first layer is more dispersed, which means that when attending to verbs and adpositions, attention is also looking at other words to a lesser degree, possibly such words which are involved in the action described by the verb. We calculated the Pearson correlation coefficient between both heatmaps in Figure 3. The test has shown a significant negative correlation ($r = -0.71, p = 1.7e-08$), indicating that there is a clear separation in attention focus on two types of words in masked self-attention. Overall, text-to-text attention is able to capture local and non-grounded syntactic knowledge of objects and relations between them.

Masked Self-Attention on Spatial Relations Figures 6a–6d show the attention focus in masked self-attention for several possible directions between parts of the phrase describing spatial relation. For example, $rel \rightarrow target$ shows the attention on the noun phrase describing the target object when a phrase describing relation is generated. Note that in masked self-attention, we are not able to look into the future; thus, we cannot inspect attention on $rel \rightarrow landmark$ or $target \rightarrow landmark$. The first important observation is a clear difference between attention on the word depicting the target object depending on where this attention is coming from. Numerous attention heads in the first layers focus on the target when relation is generated (Figure 6a), while only a few heads are looking at the target when landmark is generated. According to Figure 6b, relation is more important for landmark since it is widely attended by many heads, compared to only a few heads in Figure 6c and only a single head (head 8, layer 4) being highly active. In addition, there are three attention heads in the second layer (2, 3, 4) in Figure 6a, which are also highly activated in Figure 5a. This might indicate that these heads do not simply look at the words depicting objects but specialise in such words, which are playing the part of the "target" object in spatial relations. Therefore, we can identify particular heads that learn knowledge of syntactic dependencies between words describing spatial relations in the textual encoder. Also, based on Figure 6b, we can see that the focus on relation phrases is mostly captured in earlier layers, which supports our statement that the model first needs to learn general knowledge about existing relations in the scene, later starting to exploit it for better focus on correct target and landmark nouns.

Cross-Attention on Spatial Relations Figures 6e–6h show how much each head looks at the specific object that corresponds to a target or a landmark in spatial relations. Similar to our experiment on visual grounding, we linked every noun phrase describing either a target or a landmark with a bounding box of the detected object by computing semantic similarity between the noun phrase and the label of every object. Note that here we look at how words of *semantic categories* describing relations between objects are grounded in *visual representations* (objects) rather than other words, as in the case of the masked self-attention. One

noticeable difference between the top and bottom rows in the Figure 6 is that the attention focus in the cross-modal part of the architecture is much more distributed across many heads.

Given that, according to Figure 4, multi-modal grounding of nouns into objects is clearly observed in later parts of the model, grounding of relations in objects becomes much less interpretable. First, relations cannot be directly linked to the objects in the scene since relations do not describe them. When grounding relations, the system needs to rely on several sources of knowledge: it tends to rely on linguistic knowledge much more than on visual information (Ghanimifard and Dobnik, 2019). Learning is further complicated by the fusion of information in cross-attention. For example, the model needs to simultaneously rely on the semantic information from the language representations and identify objects that are targets and landmarks in spatial relations. Therefore, cross-modal attention activates many attention heads when trying to learn about spatial relations, which require attention on multiple sources of knowledge.

Interestingly, as Figure 6f shows, a lot of attention on landmark is distributed across multiple layers. Specifically, surface layers, which also have higher entropy (Figure 3b), are activated much more compared to, for example, $target \rightarrow land$ relation (Figure 6h). This can be attributed to the fact that the model learns to attend to targets with high confidence in deeper layers because targets constitute the central part of spatial relations and require more complex reasoning. At the same time, landmarks are intuitively semantically closer to the relation and general information about the target object, which can be captured in surface layers. Dobnik et al. (2018) have shown that there is a strong asymmetry between knowledge about targets and landmarks: landmarks are generally much easier to predict, and they contribute less to the perplexity of the model than targets. Intuitively, a speaker would like to describe the target, and they need to find a suitable contextually salient landmark to produce such a description. Therefore, it might happen that the model first distributes its attention between heads in earlier and later layers to identify landmarks in the context of particular relation, and then learns to strongly map this relation-landmark context with the specific target in deeper layers. This idea is also supported by strongly localised and focused attention on the target object in deeper layers when either a relation or a landmark are generated (Figure 6g and Figure 6e).

Note the differences between attention patterns in Figure 6a and Figure 6e for the $relation \rightarrow target$ direction. Lower layers in masked self-attention, as we have shown, seem to learn local syntactic dependencies between words in the source input (text). This is different from the multi-modal scenario, where deeper layers are much more activated for visual and language inputs. This indicates that spatial relations are much more sophisticated in the language-and-vision context: they need to capture semantic dependencies between words and objects in the scene. Also, the complexity of information might be the reason why $rel \rightarrow target$ attention is much more scattered across many heads in deeper layers in cross-modal attention, compared to more focused attention in specific heads in earlier layers for masked self-attention.

6 Conclusion

We have shown that the language model in a multi-modal task captures linguistic phenomena of different kind depending on the source knowledge (text or objects) and semantic type of the output words (noun phrases or spatial relations). In particular, while text-only attention learns low-level linguistic phenomena, e.g. local syntactic dependencies. On the other hand, cross-modal attention visually grounds objects and, therefore, semantic dependencies in its deeper layers. We has also shown that learning spatial relations cross-modally is challenging. Overall, our work demonstrates that attention on vision and language captures considerably more diverse linguistic knowledge, *both syntactic and semantic* which might not be *linearly aligned*, compared to uni-modal (language only) architectures. One possible follow-up experiment is to use attention as input to the probing classifier and identify a specific knowledge encoded by the weights. However, the performance of the probing model does not tell us whether the original model utilises acquired knowledge since it is *detached* from the original architecture (Belinkov, 2021). In contrast, inferring linguistic properties from attention weights makes our analysis "weightless", with no requirement of learning a new set of parameters. We believe that our work provides one possible explanation of how a complex, large-scale language model learns knowledge about the world.

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.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision – ECCV 2020*, pages 565–580, Cham. Springer International Publishing.

Moitreya Chatterjee and Alexander G. Schwing. 2018. Diverse and coherent paragraph generation from images. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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.

Simon Dobnik, Mehdi Ghanimifard, and John Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 1–11, New Orleans. 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.

Mehdi Ghanimifard and Simon Dobnik. 2019. What goes into a word: generating image descriptions with top-down spatial knowledge. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 540–551, Tokyo, Japan. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

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.

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.

Nikolai Ilinykh and Simon Dobnik. 2021. How vision affects language: Comparing masked self-attention in uni-modal and multi-modal transformer. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 45–55, Groningen, Netherlands (Online). Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie-Francine Moens, and Steven Bethard. 2013. SemEval-2013 task 3: Spatial role labeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 255–262, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Patterm Recognition (CVPR)*.

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.

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.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Shimon Ullman. 1984. Visual routines. *Cognition*, 18(1-3):97–159.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Appendix A

Pronouns, e.g. `it`, `his` were not linked with any object in the scene. Also, several noun phrases depicting spatial descriptions or locations were also ignored, e.g. `the right`, `the background`, `the corner`. Some noun phrases were describing properties of objects in the scene (e.g., `color`, `the overall color of the room`) or positional arrangement (`a straight line` in `three paintings hang in a straight line`). Other noun phrases described a general understanding of the image, and not a single bounding box could cover it (`a beachside hotel` in `a room that looks like inside a beachside hotel`). Some noun phrases looked incorrect due to either an error by spaCy or human who produced the original description, e.g. `the walls floor sofa`.

B Appendix B

When extracting noun phrases for the experiment on visual grounding, we ignore any pronouns and spatial phrases, which are found in the following list: `right`, `a right`, `the right`, `left`, `a left`, `the left`, `top`, `the top`, `bottom`, `the bottom`, `back`, `the back`, `front`, `the front`, `far`, `the far`, `close`, `the close`, `side`, `each side`, `background`, `the background`, `foreground`, `the foreground`, `middle`, `the middle`, `corner`, `a corner`, `the corner`.