Attending Self-Attention: A Case Study of Visually Grounded Supervision in Vision-and-Language Transformers

Jules Samaran[1] **Noa Garcia**[2]
Mayu Otani[3] **Chenhui Chu**[4] **Yuta Nakashima**[2]
[1]PSL Research University [2]Osaka University [3]CyberAgent, Inc. [4]Kyoto University
jules.samaran@mines-paristech.fr
{noagarcia, n-yuta}@ids.osaka-u.ac.jp
otani_mayu@cyberagent.co.jp chu@i.kyoto-u.ac.jp

Abstract

The impressive performances of pre-trained visually grounded language models have motivated a growing body of research investigating what has been learned during the pre-training. As a lot of these models are based on Transformers, several studies on the attention mechanisms used by the models to learn to associate phrases with their visual grounding in the image have been conducted. In this work, we investigate how supervising attention directly to learn visual grounding can affect the behavior of such models. We compare three different methods on attention supervision and their impact on the performances of a state-of-the-art visually grounded language model on two popular vision-and-language tasks.

1 Introduction

The introduction of Transformers (Vaswani et al., 2017) has been a major component of the success of pre-trained language models (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Lan et al., 2020) which achieved new records in many natural language processing tasks. The same mechanism has been adapted to create models (Su et al., 2020; Chen et al., 2020; Li et al., 2019; Lu et al., 2019, 2020; LXM) that can now tackle vision-and-language tasks with impressive performances.

A large body of research (Clark et al., 2019; Kovaleva et al., 2019) has been dedicated to understanding what attention heads learn during the pre-training of language models. Liu et al. (2016) have even shown how providing attention heads with guidance can improve performance on neural machine translation.

On the other hand, the internal behaviors of vision-and-language models have attracted less interest from the research community. Li et al. (2020) have shown some attention heads in vision-and-language models are able to map entities to image regions while others even detect syntactic relations between non-entity words and image regions. Nevertheless, no initiative has been taken towards supervising directly the attention modules.

In this paper, we study how different methods on attention supervision can affect vision-and-language models. We propose a fine-tuning method aimed at using the visual grounding of entities to provide guidance to attention heads. We compare three different methods by evaluating their performance on popular downstream tasks and visualize the different attention modules obtained. We observe that an indirect method which uses a module appended to the final output of the Transformer obtains worse results than methods which focus on supervising every attention head directly. The codes are available at https://github.com/jules-samaran/VL-BERT.

2 Our Method

We use a state-of-the-art pre-trained vision-and-language model on which we propose multi-task fine-tuning methods focusing on attention supervision. After this proposed fine-tuning, we judge the success of our approach by further fine-tuning the model on downstream tasks of visual question answering and referring expressions, and evaluating it. We propose a fine-tuning approach after an initial pretraining step on a large unlabelled dataset because we believe the model would benefit from learning first from scratch freely about text and images without any supervision on its attention heads, and that our fine-tuning would allow it to then refine the representations it provided using visual grounding labels.

2.1 Backbone Model

We choose as our basic architecture VL-BERT (Su et al., 2020), a state-of-the-art vision-and-language pre-trained model that revisits BERT (Devlin et al.,

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 81–86
August 5–6, 2021.

2019) to take both visual and linguistic inputs. Based on a multi-layer bidirectional multi-modal Transformer encoder (Vaswani et al., 2017), VL-BERT learns during its pre-training a generic feature representation mainly on the conceptual captions dataset consisting of 3.3M image-caption pairs (Sharma et al., 2018). Note that many other choices are possible for this backbone model (see the Section 4).[1] The reason we chose VL-BERT is that it was available; it achieved state-of-the-art performances (better than ViLBERT (Lu et al., 2019, 2020) for example) on several classical vision-and-language downstream tasks; the way it handles both visual and textual tokens in a single stream (whereas ViLBERT processes them in two separate streams) made it very adapted for our approach of supervising the attention between textual and visual elements.

2.2 VGP Fine-tuning

To provide guided attention supervision in vision-and-language models, we devise a multi-task fine-tuning method that aims to improve the model's ability to understand complex semantic relations (e.g. paraphrases) and align visual with linguistic elements. Li et al. (2020) hinted the importance of attention-based vision-and-language model's ability to map entity-words to corresponding image regions. Following this direction and to improve a model's reasoning abilities, we propose to further fine-tune a pre-trained model with the aim of learning visually grounded paraphrases (VGPs) (Chu et al., 2018; Otani et al., 2020).

VGPs are two phrasal expressions that describe the same visual concept in an image. As shown in Figure 1, we fine-tune a model based on VGPs with three different tasks simultaneously as a multi-task learning problem: an image description identification task (§2.2.1), a VGP classification task (§2.2.2), and an attention supervision task (§2.2.3). The first two tasks are inspired by Arase and Tsujii (2019), who showed that injecting semantic relations between a sentence pair can improve a BERT model's performance on several downstream tasks. We adapted them to make the model learn from both visual and linguistic elements.

Input The input of the fine-tuning process is composed of 1) an image, and 2) a pair of captions,

[1]It is unclear how well the results we obtained would generalize to other vision-and-language models, especially since our approach is designed for VL-BERT's architecture, but we leave it as future work.

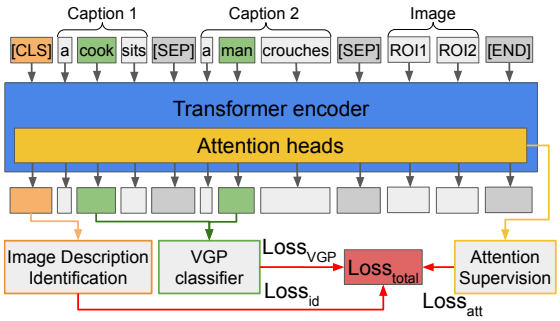

Figure 1: Overview of the VGP fine-tuning. The Transformer Encoder is the VL-BERT model, our contributions are the modules on the bottom.

c_1 and c_2, where at least one of them corresponds to the image. Hard negative captions are chosen offline following Lu et al. (2019). The input sequence to the Transformer model is constructed as: $[\texttt{[CLS]}, c_1, \texttt{[SEP]}, c_2, \texttt{[SEP]}, img, \texttt{[END]}]$, where `[CLS]` is the start of sequence token, `[SEP]` separates different elements, img is the image regional features extracted as Girshick (2015) and `[END]` is the end of sequence token.

2.2.1 Image Description Identification

In this task, a Softmax classifier f takes the output x_0 of the Transformer corresponding to `[CLS]` and predicts which of the captions corresponds to the image (c_1, c_2, or both). The loss is given by

$$Loss_{\text{id}} = L_{\text{CE}}(y, \log f(x_0)), \quad (1)$$

where y indicates which caption corresponds to the image and L_{CE} is the cross-entropy loss function.

2.2.2 VGP Classification

The second task is to classify VGPs according to the semantic relationship between the two phrases, e.g. entailment, equivalence, etc. More details on the semantic classes are provided in the supplementary material. Let $p = \{x_i\}$ denote the set of the final encodings x_i, where i ranges from the first to the last words of the sampled phrase. Phrase embedding e of the phrase is given by

$$e = \text{MaxPooling}(p). \quad (2)$$

We obtain phrase embeddings e_1 and e_2 from c_1 and c_2, combining them as Arase and Tsujii (2019)

$$e_c = [e_1, e_2, e_1 * e_2, |e_1 - e_2|], \quad (3)$$

where $*$ denotes the element-wise multiplication and $|\cdot|$ gives element-wise absolute values. For

VGP classification, we input e_c to a Softmax classifier g to predict the paraphrase's semantic class. The loss for this task is computed as

$$Loss_{\text{VGP}} = L_{\text{CE}}(t, \log g(e_c)), \quad (4)$$

where t is the label of the VGP semantic class.

2.2.3 Attention Supervision Task

Since VGPs align directly with image regions, we can have the model learn the visual grounding. We explore three methods of supervising attention.

Indirect attention supervision We learn the visual grounding using the final representations of grounded phrases and their aligned regions of interest. We train a binary classifier d that takes as input phrase embedding e as well as the representation x_i of one of the regions in img, and predicts whether they align or not. We repeat this classification for every grounded phrase with every region. The loss is computed by

$$Loss_{\text{att}} = \frac{1}{n_r} \sum_i L_{\text{CE}}(z_i, \log d([e, x_i])), \quad (5)$$

where z_i is the indicator of whether the phrase refers to the i-th region, n_r is the number of regions, and the summation is computed over i corresponding to regions in img.

Direct attention supervision Similarly to Liu et al. (2016), we view every attention head in every layer as a classifier that, given an input token, outputs probabilities distributed over all the other tokens. Our motivation is that the attention between two elements should reflect how much they are relevant to each other, hence grounded word entities should pay attention the most to their visual grounding, and vice versa. We re-normalize the attention so that this supervision has only a limited impact on text-to-text and region-to-region attention. Specifically, let W denote the set of indices i of all text tokens, and R corresponding to regions in img. The attention α_{ij}^{lh} for layer l, head h, from the i-th token to j-th region can be normalized by

$$\hat{\alpha}_{ij}^{lh} = \frac{\alpha_{ij}^{lh}}{\sum_{j' \in R} \alpha_{ij'}^{lh}}, \quad \tilde{\alpha}_{ij}^{lh} = \frac{\alpha_{ij}^{lh}}{\sum_{j' \in W} \alpha_{ij'}^{lh}}. \quad (6)$$

The phrase grounding gives pairs (i^*, j^*) in both $W \times R$ and $R \times W$ (we have multiple pairs since a phrase has multiple tokens). We use the average of cross-entropy losses for supervision, i.e.,

$$Loss_{\text{txt}\to\text{img}} = \frac{1}{n_l n_h} \sum_{l,h} L_{\text{CE}}(s_{i^*}, \log \hat{\alpha}_{i^* j}^{lh}), \quad (7)$$

where $j \in R$; s_{i^*} is the indicator whether respective region $j \in R$ forms pair (i^*, j^*); n_l and n_h are the numbers of layers and attention heads, respectively. We do the same for the loss $Loss_{\text{img}\to\text{txt}}$ for region-to-text pairs using $\tilde{\alpha}_{ij}^{lh}$, where $j \in W$ and s'_{i^*} is the corresponding indicator. The loss for direct attention supervision loss is given by

$$Loss_{att} = Loss_{\text{txt}\to\text{img}} + Loss_{\text{img}\to\text{txt}}. \quad (8)$$

Semi-direct attention supervision Abnar and Zuidema (2020) introduced a transformation of raw attention called attention rollout and showed that it gives a more accurate quantification of how much information one token contains about another token than raw attention does. Therefore, we propose to replace the raw attention with the attention rollout in the direct supervision method. In other words, if we denote $f_{\text{rollout}(\cdot)}$ as the function that transforms raw attention vectors into attention rollout then our semi-direct attention supervision approach consists in replacing α_{ij}^{lh} with $f_{\text{rollout}}(\alpha)_{ij}^{lh}$ in Equations (refeq6), (7) and (8).

The final loss for the VGP fine-tuning is

$$Loss_{\text{total}} = Loss_{\text{id}} + Loss_{\text{VGP}} + Loss_{\text{att}}. \quad (9)$$

3 Experiments

3.1 Dataset

For our fine-tuning, we used the VGP dataset (Chu et al., 2018), which was created from the Flickr30k-entities dataset's captions (Plummer et al., 2017). As it is based on Flickr30k-entities, those phrases come with the id of the image region that corresponds to their grounding. The dataset contains $54,313$ VGPs distributed across $31,784$ images.

3.2 Fine-tuning on Downstream Tasks

To evaluate how our fine-tuning methods can improve the generic representations generated by the model, we further fine-tune it on downstream vision-and-language tasks and compare their performances. Results are reported in Table 1, including the performance of the original VL-BERT model. We also include a model fine-tuned on VGPs without the attention supervision task, with only the image description identification and VGP classification in order to estimate the impact of forcing the model learn visual grounding.

	VQAv2.0	**Refcoco+** (Detected)			**Refcoco+** (Ground-truth)		
Fine-tuning Method	val	val	testA	testB	val	testA	testB
w/o Attention	66.73	67.10	74.36	57.07	77.38	81.28	71.53
Indirect Attention	66.71	65.95	72.72	54.49	77.20	80.61	70.81
Direct Attention	67.09	69.99	76.25	58.99	77.07	80.86	70.96
Semi-direct Attention	67.41	69.63	75.93	58.72	78.12	80.96	71.75
Original (Su et al., 2020)	67.73	71.60	77.72	60.99	79.88	82.40	75.01

Table 1: Comparison of our different fine-tuning methods on the VQAv2.0 and the Refcoco+ datasets. For each column, the best fine-tuning method is underlined. Original VL-BERT results added as reference.

3.2.1 Visual Question Answering

In this task, every input is an image coupled with a question expressed in natural language. We used the VQAv2.0 dataset (Goyal et al., 2017). The model is expected to answer the question with the correct answer picked from a shared set consisting of $3,129$ answers according to Anderson et al. (2018). We trained the models on the train split ($83k$ images and $444k$ questions) and report results on the validation split ($41k$ images and $214k$ questions). We used the same experimental protocol for prediction and evaluation as in Su et al. (2020).

Results in Table 1 indicate that the original model is the best performing method, showing that forcing VL-BERT to learn paraphrases before training the VQAv2.0 dataset does not contribute to the task. However, it is still relevant to compare the performances of different attention supervision methods. The two worse performance are attained by the method without attention and the Indirect, which does not seem to improve the model's ability to answer the question. Both the Direct and Semi-direct methods, which use the attention heads as classifiers, fare better with a slight advantage for the Semi-direct method.

3.2.2 Referring Expression Comprehension

The objective of this task is to locate the object in the image that is designated by the input phrase. The input is constituted of a referring expression and an image that contains the object that is being referred to. We used the RefCOCO+ dataset (Kazemzadeh et al., 2014) ($141k$ expressions for $50k$ referred objects in $20k$ images). The dataset contains two test sets, where testA contains images with multiple persons and testB with multiple objects. We report results both with ground-truth RoIs and with the bounding boxes detected by Yu et al. (2018). We also used the same experimental protocol for prediction and evaluation as in Su et al. (2020).

As shown in Table 1, despite having been designed with the referring expression task in mind, the Indirect attention supervision method is the worse one, even behind the method without attention. The original VL-BERT model is still the leading performance followed by the Direct and Semi-direct attention supervision methods. Direct attentions works better on detected bounding boxes. We think the reason is that direct attention tries to link tokens with image regions similarly to how the region detector would do it.

VL-BERT is pre-trained on 3.3M image-caption pairs, while VGP fine-tuning is conducted on 30k image-caption pairs only. Therefore, for both tasks, we believe that our methods failed to beat VL-BERT due to two reasons: catastrophic forgetting, and small-scale attention supervision training data. To address catastrophic forgetting, applying knowledge distillation (Hinton et al., 2015) that can incorporate both knowledge from the pre-trained VL-BERT model and the VGP fine-tuned model might be effective. For the small-scale attention supervision training data issue, a possible direction could be applying visual grounding on the conceptual captions dataset and training VL-BERT from scratch with attention supervision on the conceptual captions dataset.

3.2.3 Visualization

To gain more insights into what models learn with the different attention supervision methods, we visualize attention heads using Bertviz[2] (Vig, 2019).

By zooming in on individual attention heads, we noticed that when the model was fine-tuned using

[2] https://github.com/jessevig/bertviz

either the Semi-direct or Direct attention supervision methods, every grounded entity text token attributes more attention to image tokens to image tokens corresponding to the visual grounding of the entity. We also observed that even though the Direct method seemed to have an uniform impact on all attention heads in every layer, with the Semi-direct method attention heads displayed varying attention patterns across different layers. A possible explanation is that the attention rollout transformation makes the attention supervision problem slightly different across different layers whereas it is not the case for the Direct method which imposes the same constraint on the raw attention in all attention heads (and it is the raw attention we are visualizing). More details about the visualization and images are provided in the supplementary materials.

4 Related Work

Vision and language pre-trained models on large image caption datasets have been proposed such as VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019, 2020), VL-BERT (Su et al., 2020; Lu et al., 2020), LXMERT (LXM) and UNITER (Chen et al., 2020). Those vision and language pre-training models differ from the model architecture. We study visually grounded attention supervision in VL-BERT.

Clark et al. (2019); Kovaleva et al. (2019) analyzed on language pre-trained models and showed that different attention heads share similar patterns and behaviors. For neural machine translation, Liu et al. (2016) proposed to use word alignment for cross-attention supervision during decoding in a recurrent neural network based architecture. We work specifically on vision-and-language transformers and use phrase visual grounding for attention supervision in order to help the model learn how to align phrases with their associated regions in the images.

5 Conclusion

Motivated by similar works in language models, we have presented three different methods that attempt to guide the model in its learning of entity grounding. We observed that the indirect method which is the most similar to the structure used for downstream tasks had a little or negative effect on the performance of the model. We also found that supervising attention heads through attention rollout is the best performing method nevertheless all these methods fell short of the performances of the model before being fine-tuned on the VGP dataset.

Despite the performance, we have shown which attention supervision methods give better results and more interpretable attention patterns[3] (i.e., direct and semi-direct attention) than others that should not be used (i.e., indirect attention). Therefore, we believe that our work can pave the way for further analyses of how this mechanism could be made to improve the performance of vision-and-language models. For future work, we plan to study how direct supervision methods could be applied on some selected heads instead of supervising uniformly all attention heads in every layer.

Acknowledgments

This work was supported by ACT-I, JST.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer fine-tuning: A BERT case study. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 5393–5404.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision*, pages 104–120.

Chenhui Chu, Mayu Otani, and Yuta Nakashima. 2018. iParaphrasing: Extracting visually grounded paraphrases via an image. In *International Conference on Computational Linguistics*, pages 3479–3492.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

[3] A visualization comparison among the attention supervision methods can be found in the supplementary material.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Ross Girshick. 2015. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6325–6334.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*, pages 787–798.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 17 pages.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. In *CoRR arXiv:1908.03557*, 14 pages.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In *International Conference on Computational Linguistics*, pages 3093–3102.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. In *CoRR arXiv:1907.11692*, 13 pages.

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*, pages 13–23.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition)*, pages 10437–10446.

Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. Visually grounded paraphrase identification via gating and phrase localization. *Neurocomputing*, 404:165–172.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations.

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*, pages 5998–6008.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5753–5763.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. MAttNet: Modular attention network for referring expression comprehension. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315.