What Does BERT with Vision Look At?

Liunian Harold Li[†], **Mark Yatskar**[*], **Da Yin**[°], **Cho-Jui Hsieh**[†] **& Kai-Wei Chang**[†]
[†]University of California, Los Angeles
[*]Allen Institute for Artificial Intelligence
[°]Peking University
liunian.harold.li@cs.ucla.edu, marky@allenai.org,
wade_yin9712@pku.edu.cn, {chohsieh, kwchang}@cs.ucla.edu

Abstract

Pre-trained visually grounded language models such as ViLBERT, LXMERT, and UNITER have achieved significant performance improvement on vision-and-language tasks but what they learn during pre-training remains unclear. In this work, we demonstrate that certain attention heads of a visually grounded language model actively ground elements of language to image regions. Specifically, some heads can map entities to image regions, performing the task known as *entity grounding*. Some heads can even detect the syntactic relations between non-entity words and image regions, tracking, for example, associations between verbs and regions corresponding to their arguments. We denote this ability as *syntactic grounding*. We verify grounding both quantitatively and qualitatively, using Flickr30K Entities as a testbed.

1 Introduction

Recently, BERT (Devlin et al., 2019) variants with vision such as ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019), and UNITER (Chen et al., 2019) have achieved new records on several vision-and-language reasoning tasks, e.g. VQA (Antol et al., 2015), NLVR^2 (Suhr et al., 2019), and VCR (Zellers et al., 2019). These pre-trained visually grounded language models use Transformers (Vaswani et al., 2017) to jointly model words and image regions. They are pre-trained on paired image-text data, where given parts of the input the model is trained to predict the missing pieces. Despite their strong performance, it remains unclear if these models have learned the desired cross-modal representations.

Conversely, a large body of work (Liu et al., 2019; Tenney et al., 2019; Clark et al., 2019) has focused on understanding the internal behaviours of pre-trained language models (Peters et al., 2018b; Radford et al., 2018; Devlin et al., 2019) and find that they capture linguistic features such as POS, syntactic structures, and coreferences. This inspires us to ask: what do visually grounded language models learn during pre-training?

Following Clark et al. (2019), we find that certain attention heads of a visually grounded language model acquire an intuitive yet fundamental ability that is often believed to be a prerequisite for advanced visual reasoning (Plummer et al., 2015): grounding of language to image regions.

We first observe that some heads can perform **entity grounding**, where entities that have direct semantic correspondences in the image are mapped to the correct regions. For example, in Figure 1, the word "man" attends to the person on the left of the image. Further, non-entity words often attend to image regions that correspond to their syntactic neighbors and we call this **syntactic grounding**. For example, "wearing" is attending to its subject, the man in the image. We argue that syntactic grounding actually complements entity grounding and that it is a natural byproduct of cross-modal reasoning. For example, to ground "man" to the person on the left rather than other pedestrians, the model needs to identify the syntactic relationships among "man", "wearing", "white", and "shirt" and ground "shirt" and "man" subsequently. During such process, it is helpful and natural that "wearing" attends to the man in the image.

We verify such phenomena by treating each attention head as a ready-to-use classifier (Clark et al., 2019) that given an input word, always outputs the most-attended-to image region. Using Flickr30K Entities (Plummer et al., 2015) as a test bed, we demonstrate that certain heads could perform entity and syntactic grounding with an accuracy significantly higher than a rule-based baseline. Further, higher layers tend to have higher grounding accuracy, suggesting that the model is

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5265–5275
July 5 - 10, 2020.

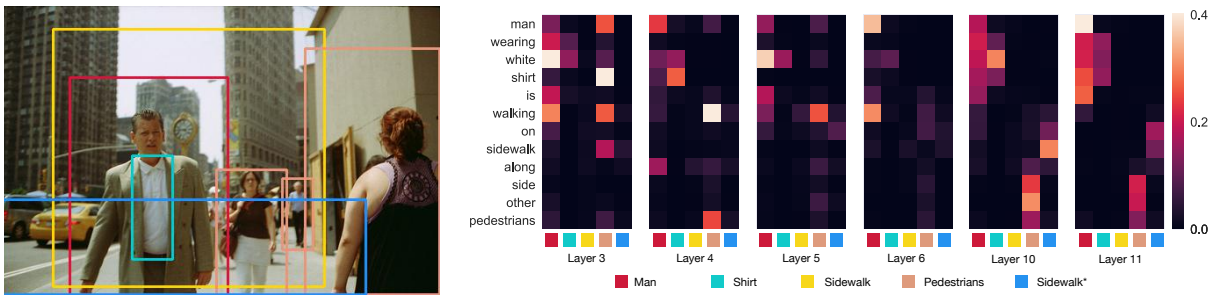

Figure 1: Attention weights of some selected heads in a pre-trained visually grounded language model. In high layers (e.g., the 10-th and 11-th layer), the model can implicitly grounding visual concepts (e.g., "other pedestrians" and "man wearing white shirt"). The model also captures certain syntactic dependency relations (e.g., "walking" is aligned to the *man* region in the 6-th layer). The model also refines its understanding over the layers, incorrectly aligning "man" and "shirt" in the 3-rd layer but correcting them in higher layers.

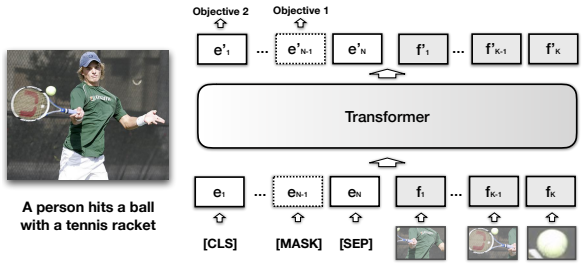

Figure 2: The architecture of VisualBERT. Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision. n. It is pre-trained with a masked language modeling (Objective 1), and sentence-image prediction task (Objective 2), on caption data and then fine-tuned for different tasks.

Task	Baseline	VisualBERT
VQA	68.71	70.80
VCR	44.0	52.4
NLVR^2	53.5	67.3
Flickr30K	69.69	71.33

Table 1: Performance of VisualBERT on four benchmarks. On VQA, we compare to Pythia v0.3 (Singh et al., 2019) and report on test-dev; on VCR, we compare to R2C (Zellers et al., 2019) and report test accuracy on Q→AR; on NLVR^2, we compare to MaxEnt (Suhr et al., 2019) and report on Test-P; on Flickr30K, we compare to BAN (Kim et al., 2018) and report the test recall@1.

refining its understanding of vision and language layer by layer. Additionally, we provide a qualitative analysis exemplifying these phenomena. A long version of this paper is at https://arxiv.org/abs/1908.03557. Our code is available at https://github.com/uclanlp/visualbert.

2 Model

Several pre-trained visually grounded models have been proposed recently, and they are conceptually similar yet vary in design details, making evaluating them complicated and difficult. Thus for simplicity, we propose a simple and performant baseline, VisualBERT (see Figure 2), and base our analysis on this model. We argue that our analysis on VisualBERT can be generalized to other similar models as all these models share the following two core ideas: (1) image features extracted from object detectors such as Faster-RCNN (Ren et al., 2015) are fed in a Transformer-based model along with text; (2) the model is pre-trained on image-text data with a masked visually grounded langauge model objective. Below we introduce VisualBERT briefly and leave details to the Appendix A.

Input to VisualBERT includes a text segment and an image. The image is represeted as a set of visual embeddings, where each embedding vector corresponds to a bounding region in the image, derived from an object detector (Ren et al., 2015). Text and visual embeddings are then passed through multiple Transformer layers to build joint representations. VisualBERT is pre-trained on the COCO dataset (Chen et al., 2015), concisting of around 100K images with 5 captions each. We use two objectives for pre-training. (1) Masked language modeling with the image. Some elements of text input are masked and the model learns to predict the masked words based on the remaining text and visual context. (2) Sentence-image prediction. For COCO, where there are multiple captions corresponding to one image, we provide a text segment consisting of two captions. One of the caption is describing the image, while the other has a 50% chance to be another corresponding caption and a

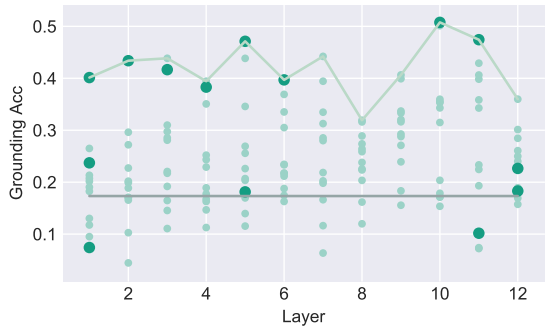

Figure 3: Entity grounding accuracy of the attention heads organized by layer. The rule-based baseline is drawn as the grey line. We find that certain heads achieve high accuracy while the accuracy peaks at higher layers.

Type	Baseline	Acc	Head
det	19.59	54.01	10-1
pobj	17.34	32.82	11-11
amod	18.67	45.96	10-9
nsubj	23.19	44.64	5-1
prep	20.61	49.27	9-11
dobj	9.82	30.24	11-11
punct	23.32	48.80	3-6
partmod	21.41	38.15	4-9
nn	16.33	34.06	10-9
num	23.15	67.44	9-11

Table 2: The best performing heads on grounding 10 most common dependency relationships. We only consider heads that are allocating on average more than 20% of the attention from source words to all image regions. A particular attention head is denoted as <layer>-<head number>.

50% chance to be a randomly drawn caption. The model is trained to distinguish these two situations.

Extensive experiments on four vision-and-language datasets (Goyal et al., 2017; Zellers et al., 2018; Suhr et al., 2019; Plummer et al., 2015) verify that pre-trained VisualBERT exceeds all comparable baselines significantly. A summary of the results is present in Table 1. See the Appendix B for details. Some of the afore-mentioned pre-trained visually grounded language models use additional pre-training data or parameters and achieve better performance. As this paper focuses on the analysis, we do not focus on comparing the performance of VisualBERT and other similar models. For the rest of the paper, we analyze a VisualBERT that is configured the same as BERT$_{Base}$ with 12 layers and 144 self-attention heads in total. The model is pre-trained on COCO. To mitigate the domain difference between the diagnostic dataset Flickr30K and COCO, we perform additional pre-training on the training set of Flickr30K with the fore-mentioned masked language modeling objective with the image.

3 Experiment

3.1 Quantitative analysis

Entity Grounding We first focus on entity grounding and use the validation set of Flickr30K Entities for evaluation. The dataset contains image-caption pairs and annotates the entities in the captions and the corresponding image regions. For each annotated entity and for each attention head of VisualBERT, we take the bounding region which receives the most attention weight as the prediction. An entity could attend to not only the image regions but also other words in the text. For this evaluation, we regard the image region that receives the most attention weight compared to other image regions as the prediction, without considering other words in the text. The predicted region is considered correct as long as it overlaps with the gold bounding region with a IoU ≥ 0.5 (Kim et al., 2018). We also consider a rule-based baseline that always chooses the region with the highest detection confidence. We report the accuracy for all 144 attention heads in VisualBERT and the baseline in Figure 3. Despite that some heads are accurate at entity grounding, they are not actively attending to the image regions. For example, a head might be allocating 10% of its attention weights to all image regions, but it assigns the most of the 10% weights to the correct region. We regard heads paying on average more than 20% of its attention weights from the entities to the regions as "actively paying attention to the image" and draw then as dark and large dots, while the others are drawn as light and small dots.

We make the following two observations. First, *certain heads perform entity grounding with a remarkably high accuracy*. This is consistent with the observations in Clark et al. (2019) and Voita et al. (2019) that the attention heads specialize in different things. The best of all heads even achieves a high accuracy of 50.77 compared to the baseline 17.33. Further, *the grounding accuracy peaks in higher layers*. This resembles what Tenney et al. (2019) find, in that BERT also refines its understanding of the input over the layers.

Syntactic Grounding As motivated before, alignments between words other than nouns and

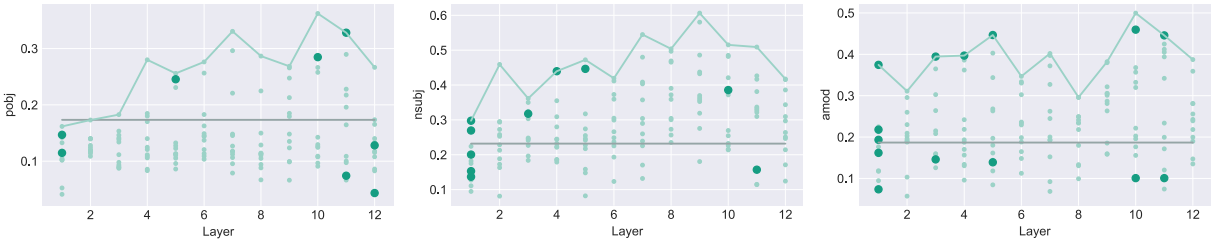

Figure 4: Accuracy of attention heads of VisualBERT for syntactic grounding on specific dependency relationships ("pobj", "nsubj", "amod"). The grey lines denote a baseline that always chooses the region with the highest detection confidence. We observe that VisualBERT is capable of detecting these dependency relationships without direct supervision.

image regions could also be helpful for visual reasoning. More specifically, if two words are connected with a dependency relation, $w_1 \overset{r}{\longleftrightarrow} w_2$, and w_1 is an entity aligned to an image region, we would like to know how often the attention heads attend from w_2 to the regions corresponding to w_1. For evaluation, we parse all sentences in the validation set of Flickr30K using AllenNLP (Dozat and Manning, 2017; Gardner et al., 2018) and use the parser output as the gold parsing annotation.

We find that *for each dependency relationship, there exists at least one head that significantly outperforms guessing the most confident bounding region.* We report the 10 most common relations in Table 2 and plot the syntactic grounding accuracy of three particularly interesting dependency relationships in Figure 4. Similar to what we observe for entity grounding, *the model becomes more accurate on syntactic grounding in higher layers.*

3.2 Qualitative Analysis

Finally, we showcase several interesting examples of how VisualBERT performs grounding in Figure 1 and Figure 5. To generate these examples, for each ground-truth box, we show a predicted bounding region closest to it and manually group the bounding regions into different categories. We also include regions that the model is actively attending to, even if they are not present in the gold annotations (marked with an asterisk). We then aggregate the attention weights from words to those regions in the same category. We show the best heads of 6 layers that achieve the highest entity grounding accuracy but we find that they also exhibit a certain level of syntactic grounding.

We observe the same behaviours as in the quantitative analysis, in that VisualBERT not only performs grounding but also refines its predictions through successive Transformer layers. For example, in the bottom image in Figure 5, initially the word "husband" and the word "woman" both assign significant attention weight to regions corresponding to the woman. By the end of the computation, VisualBERT has disentangled the woman and man, correctly aligning both. Furthermore, there are many examples of syntactic alignments. In the same image, the word "teased" aligns to both the man and woman while "by" aligns to the man.

4 Related Work

There is a long research history of bridging vision and language (Chen et al., 2015; Antol et al., 2015; Zellers et al., 2019) with the lasted advances being visually grounded language models (Lu et al., 2019; Alberti et al., 2019; Li et al., 2019; Su et al., 2019; Tan and Bansal, 2019; Chen et al., 2019). However, little analysis has been done on understanding what vision-and-language models learn. Previous works on VQA and image captioning (Yang et al., 2016; Anderson et al., 2018; Kim et al., 2018) have only shown qualitative examples on the grounding ability of the models, while another line of work focuses on designing dedicated models for the entity grounding task (Xiao et al., 2017; Datta et al., 2019). We, however, present a quantitative study on whether visually grounded language models acquire the grounding ability during pre-training without explicit supervision.

Our work is inspired by papers on analyzing pre-trained language models. One line of work uses probing tasks to study the internal representations (Peters et al., 2018a; Liu et al., 2019; Tenney et al., 2019) while another studies the attention mechanism (Clark et al., 2019; Voita et al., 2019; Kovaleva et al., 2019). We follow the latter but we believe the grounding behaviour could also be probed in the internal representations of VisualBERT.

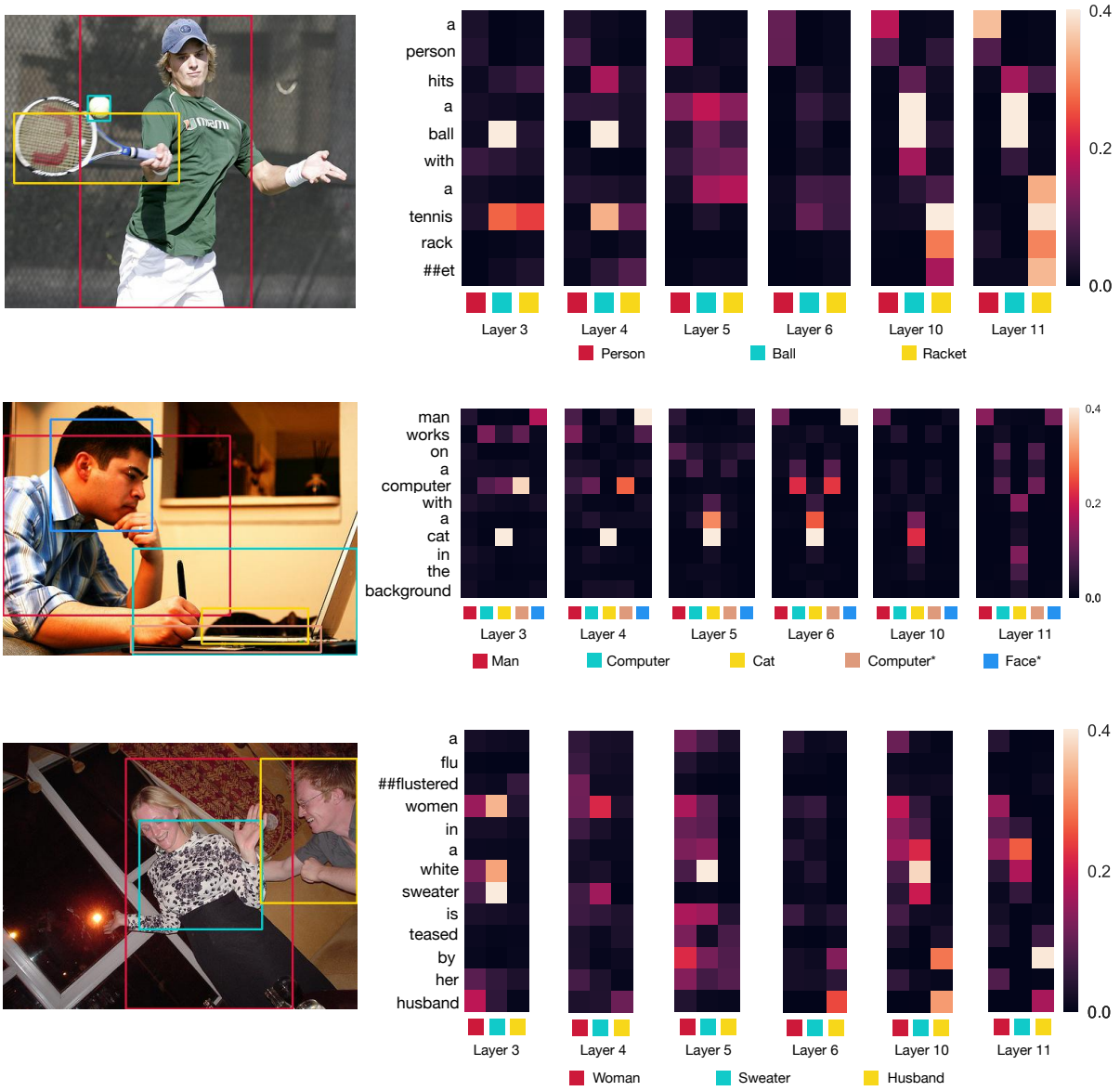

Figure 5: Attention weights of 6 selected heads in VisualBERT where alignments match Flickr30k annotations.

5 Conclusion and Future Work

We have presented an analysis on the attention maps of VisualBERT, a proposed visually grounded language model. We note that the grounding behaviour we have found is linguistically inspired, as entity grounding can be regarded as cross-modal entity coref resolution while syntactic grounding can be regarded as cross-modal parsing. Moreover, VisualBERT exhibits a hint of cross-modal pronoun resolution, as in the bottom image of Figure 5, the word "her" is resolved to the woman. For future work, it would be interesting to see if more linguistically-inspired phenomena can be systematically found in cross-modal models.

Acknowledgement

We would like to thank Xianda Zhou for help with experiments as well as Patrick H. Chen, members of UCLA NLP, and anonymous reviewers for helpful comments. We also thank Rowan Zellers for evaluation on VCR and Alane Suhr for evaluation on NLVR^2. Cho-Jui Hsieh acknowledges the support of NSF IIS-1719097 and Facebook Research Award. This work was supported in part by DARPA MCS program under Cooperative Agreement N66001-19-2-4032. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. *ArXiv*, abs/1908.05054.

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 *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? an analysis of BERT's attention. *BlackboxNLP*.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. *ICCV*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. *ICLR*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software*.

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. 2018. Detectron. https://github.com/facebookresearch/detectron.

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 *CVPR*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0. 1: the winning entry to the VQA challenge 2018. *arXiv preprint arXiv:1807.09956*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *NeurIPS*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. *ArXiv*, abs/1908.06066.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *NAACL-HLT*.

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*.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018a. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP*, pages 1499–1509.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. In *NAACL-HLT*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *CVPR*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. *ACL*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

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. *ACL*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. 2017. Weakly-supervised visual grounding of phrases with linguistic structures. *CVPR*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *CVPR*.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019a. Multimodal transformer with multi-view visual representation for image captioning. *arXiv preprint arXiv:1905.07841*.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019b. Deep modular co-attention networks for visual question answering. In *CVPR*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*.

Appendix

We first introduce the model architecture and training process of VisualBERT (Section A). We then show experiments on four vision-and-language benchmarks (Section B). Ablation study is performed to verify our design choices (Section C).

A VisualBERT

First we give background on BERT, then summarize the adaptations we made to allow processing images and text jointly, and finally explain our training procedure.

A.1 Background

BERT (Devlin et al., 2019) is a Transformer (Vaswani et al., 2017) with subwords (Wu et al., 2016) as input and trained using language modeling objectives. All of the subwords in an input sentence are mapped to a set of embeddings, E. Each embedding $e \in E$ is computed as the sum of 1) a token embedding e_t, specific to the subword, 2) a segment embedding e_s, indicating which part of text the token comes from (e.g., the hypothesis from an entailment pair) and 3) a position embedding e_p, indicating the position of the token in the sentence. The input embeddings E are then passed through a multi-layer Transformer that builds up a contextualized representation of the subwords.

BERT is commonly trained with two steps: pre-training and fine-tuning. Pre-training is done using a combination of two language modeling objectives: (1) masked language modeling, where some parts of the input tokens are randomly replaced with a special token (i.e., [MASK]), and the model needs to predict the identity of those tokens and (2) next sentence prediction, where the model is given a sentence pair and trained to classify whether they are two consecutive sentences from a document. Finally, to apply BERT to a particular task, a task-specific input, output layer, and objective are introduced, and the model is fine-tuned on the task data from pre-trained parameters.

A.2 Model

The core of our idea is to reuse the self-attention mechanism within the Transformer to implicitly align elements of the input text and regions in the input image. In addition to all the components of BERT, we introduce a set of visual embeddings, F, to model an image. Each $f \in F$ corresponds to a bounding region in the image, derived from an object detector. It is computed by summing three embeddings: (1) f_o, a visual feature representation of the bounding region of f, computed by a convolutional neural network, (2) f_s, a segment embedding indicating it is an image embedding as opposed to a text embedding, and (3) f_p, a position

embedding, which is used when alignments between words and bounding regions are provided as part of the input, and set to the sum of the position embeddings corresponding to the aligned words (see Section B.2). The visual embeddings are then passed to a multi-layer Transformer along with the original set of text embeddings, allowing the model to implicitly discover alignments between both sets of inputs, and build up a joint representation.[1]

A.3 Training VisualBERT

We would like to adopt a similar training procedure as BERT but VisualBERT must learn to accommodate both language and visual input. Therefore we reach to a resource of paired data: COCO (Chen et al., 2015) that contains images each paired with 5 independent captions. Our training procedure contains three phases:

Task-Agnostic Pre-Training As introduced before, we pre-train VisualBERT on COCO using two *visually-grounded* language model objectives. (1) Masked language modeling with the image. Some elements of text input are masked and must be predicted but vectors corresponding to image regions are not masked. (2) Sentence-image prediction. We supply two captions in one training example and one of the caption has a 50% chance to not match the image. The model is trained to determine if the provided captions is describing the image.

Task-Specific Pre-Training Before fine-tuning VisualBERT to a downstream task, we find it beneficial to train the model using the data of the task with the masked language modeling with the image objective. This step allows the model to adapt to the new target domain.

Fine-Tuning This step mirrors BERT fine-tuning, where a task-specific input, output, and objective are introduced, and the model is trained to maximize performance on the task.

B Experiment

We evaluate VisualBERT on four different types of vision-and-language applications: (1) Visual Question Answering (VQA 2.0) (Goyal et al., 2017), (2) Visual Commonsense Reasoning (VCR) (Zellers et al., 2019), (3) Natural Language for Visual Reasoning (NLVR^2) (Suhr et al., 2019), and (4) Region-to-Phrase Grounding (Flickr30K) (Plummer et al., 2015), each described in more details in the following sections. For all tasks, we use the Karpathy train split (Karpathy and Fei-Fei, 2015) of COCO for task-agnostic pre-training, which has around 100k images with 5 captions each. The Transformer encoder in all models has the same configuration as BERT$_{Base}$: 12 layers, a hidden size of 768, and 12 self-attention heads. The parameters are initialized from BERT$_{Base}$ released by Devlin et al. (2019).

For the image representations, each dataset we study has a different standard object detector to generate region proposals and region features. To compare with them, we follow their settings, and as a result, different image features are used for different tasks (see details in the subsections).[2] For consistency, during task-agnostic pre-training on COCO, we use the same image features as in the end tasks. For each dataset, we evaluate three variants of our model:

VisualBERT: The full model with parameter initialization from BERT that undergoes pre-training on COCO, pre-training on the task data, and fine-tuning for the task.

VisualBERT w/o Early Fusion: VisualBERT but where image representations are not combined with the text in the initial Transformer layer but instead at the very end with a new Transformer layer. This allows us to test whether interaction between language and vision throughout the whole Transformer stack is important for performance.

VisualBERT w/o COCO Pre-training: VisualBERT but where we skip task-agnostic pre-training on COCO captions. This allows us to validate the importance of this step.

Following Devlin et al. (2019), we optimize all models using SGD with Adam (Kingma and Ba, 2015). We set the warm-up step number to be 10% of the total training step count unless specified otherwise. Batch sizes are chosen to meet hardware constraints and text sequences whose lengths are longer than 128 are capped. Experiments are conducted on Tesla V100s and GTX 1080Tis, and all experiments can be replicated on 4 Tesla V100s each with 16GBs of GPU memory. Pre-training on COCO generally takes less than a day on 4 cards while task-specific pre-training and fine-tuning usually take less. Other task-specific training details are in the corresponding subsections.

[1] If text and visual input embeddings are of different dimension, we project the visual embeddings into a space of the same dimension as the text embeddings.

[2] Ideally, we can use the best available detector and visual representation for all tasks, but we would like to compare

Model	Test-Dev	Test-Std
Pythia v0.1 (Jiang et al., 2018)	68.49	-
Pythia v0.3 (Singh et al., 2019)	68.71	-
VisualBERT w/o Early Fusion	68.18	-
VisualBERT w/o COCO Pre-training	70.18	-
VisualBERT	70.80	71.00
Pythia v0.1 + VG + Other Data Augmentation (Jiang et al., 2018)	70.01	70.24
MCAN + VG (Yu et al., 2019b)	70.63	70.90
MCAN + VG + Multiple Detectors (Yu et al., 2019b)	72.55	-
MCAN + VG + Multiple Detectors + BERT (Yu et al., 2019b)	72.80	-
MCAN + VG + Multiple Detectors + BERT + Ensemble (Yu et al., 2019b)	75.00	75.23

Table 3: Model performance on VQA. VisualBERT outperforms Pythia(s), which are tested under a comparable setting.

Model	Q → A		QA → R		Q → AR	
	Dev	Test	Dev	Test	Dev	Test
R2C (Zellers et al., 2019)	63.8	65.1	67.2	67.3	43.1	44.0
VL-BERT (Su et al., 2019)	73.7	74.0	74.5	74.8	55.0	55.5
VisualBERT w/o Early Fusion	70.1	-	71.9	-	50.6	-
VisualBERT w/o COCO Pre-training	67.9	-	69.5	-	47.9	-
VisualBERT	70.8	71.6	73.2	73.2	52.2	52.4

Table 4: Model performance on VCR. VisualBERT w/o COCO Pre-training outperforms R2C, which enjoys the same resource while VisualBERT further improves the results.

B.1 VQA

Given an image and a question, the task is to correctly answer the question. We use the VQA 2.0 (Goyal et al., 2017), consisting of over 1 million questions about images from COCO. We train the model to predict the 3,129 most frequent answers and use image features from a ResNeXt-based Faster RCNN pre-trained on Visual Genome (Jiang et al., 2018). We report the results in Table 3, including baselines using the same visual features and number of bounding region proposals as our methods (first section), our models (second section), and other incomparable methods (third section) that use external question-answer pairs from Visual Genome (+VG) , multiple detectors (Yu et al., 2019a) (+Multiple Detectors) and ensembles of their models. In comparable settings, our method is significantly simpler and outperforms existing work.

B.2 VCR

VCR consists of 290k questions derived from 110k movie scenes, where the questions focus on visual commonsense. The task is decomposed into two multi-choice sub-tasks wherein we train individual models: question answering (Q → A) and answer justification (QA → R). Image features are obtained from a ResNet50 (He et al., 2016) and "gold" detection bounding boxes and segmentations provided in the dataset are used[3]. The dataset also provides alignments between words and bounding regions that are referenced to in the text, which we utilize by using the same position embeddings for matched words and regions. Results on VCR are presented in Table 4. We compare our methods against the model released with the dataset which builds on BERT (R2C) and list the top performing single model on the leaderboard when we submit VisualBERT to the leaderloard (VL-BERT). Our ablated VisualBERT w/o COCO Pre-training enjoys the same resource as R2C, and despite being significantly simpler, outperforms it by a large margin. The full model further improves the results. Despite substantial domain difference between COCO and VCR, with VCR covering scenes from movies, pre-training on COCO still helps significantly.

methods on a similar footing.

[3]In the fine-tuning stage, for VisualBERT (with/without Early Fusion), ResNet50 is fine-tuned along with the model as we find it beneficial. For reference, VisualBERT with a fixed ResNet50 gets 51.4 on the dev set for Q → AR. The ResNet50 of VisualBERT w/o COCO Pre-training is not fine-tuned with the model such that we could compare it with R2C fairly.

Model	Dev	Test-P	Test-U	Test-U (Cons)
MaxEnt (Suhr et al., 2019)	54.1	54.8	53.5	12.0
LXMERT (Tan and Bansal, 2019)	75.0	74.5	76.2	42.1
VisualBERT w/o Early Fusion	64.6	-	-	-
VisualBERT w/o COCO Pre-training	63.5	-	-	-
VisualBERT	67.4	67.0	67.3	26.9

Table 5: Comparison with the state-of-the-art models on NLVR^2. The two ablation models significantly outperform MaxEnt while the full model widens the gap.

Table 6: Comparison with the state-of-the-art model on the Flickr30K. VisualBERT holds a clear advantage over BAN.

Model	R@1		R@5		R@10		Upper Bound	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
BAN (Kim et al., 2018)	-	69.69	-	84.22	-	86.35	86.97	87.45
VisualBERT w/o Early Fusion	70.33	-	84.53	-	86.39	-		
VisualBERT w/o COCO Pre-training	68.07	-	83.98	-	86.24	-	86.97	87.45
VisualBERT	70.40	71.33	84.49	84.98	86.31	86.51		

B.3 NLVR^2

NLVR^2 is a dataset for joint reasoning about natural language and images, with a focus on semantic diversity, compositionality, and visual reasoning challenges. The task is to determine whether a natural language caption is true about a pair of images. The dataset consists of over 100k examples of English sentences paired with web images. We modify the segment embedding mechanism in VisualBERT and assign features from different images with different segment embeddings. We use an off-the-shelf detector from Detectron (Girshick et al., 2018) to provide image features and use 144 proposals per image.[4] Results are in Table 5. VisualBERT w/o Early Fusion and VisualBERT w/o COCO Pre-training surpass the best model in Suhr et al. (2019) (MaxEnt) by a large margin while VisualBERT widens the gap. LXMERT is pre-trained on a much larger dataset and thus shows superior performance.

B.4 Flickr30K Entities

Flickr30K Entities dataset tests the ability of systems to ground phrases in captions to bounding regions in the image. The task is, given spans from a sentence, selecting the bounding regions they correspond to. The dataset consists of 30k images and nearly 250k annotations. We adapt the setting of BAN (Kim et al., 2018), where image features from a Faster R-CNN pre-trained on Visual Genome are used. For task specific fine-tuning, we introduce an additional self-attention block and use the average attention weights from each head to predict the alignment between boxes and phrases. For a phrase to be grounded, we take whichever box receives the most attention from the last sub-word of the phrase as the model prediction. Results are listed in Table 6. VisualBERT outperforms the current state-of-the-art model BAN. In this setting, we do not observe a significant difference between the ablation model without early fusion and our full model, arguing that perhaps a shallower architecture is sufficient for grounding when supervision is available.

C Ablation Study

In this section we conduct ablation study on what parts of our approach are important to VisualBERT's strong performance. We compare two ablation models in the Experiment section and four additional variants on NLVR^2. For ease of computations, these models are trained with only 36 features per image (including the full model). Our analysis (Table 7) aims to investigate the contributions of the following four components in VisualBERT:

C1: Task-agnostic Pre-training We investigate the contribution of task-agnostic pre-training by

[4] We conducted a preliminary experiment on the effect of the number of object proposals kept per image. We tested models with 9, 18, 36, 72, and 144 proposals, which achieve an accuracy of 64.8, 65.5, 66.7, 67.1, and 67.4 respectively on the development set.

	Model	Dev
	VisualBERT	66.7
C1	VisualBERT w/o Grounded Pre-training	63.9
	VisualBERT w/o COCO Pre-training	62.9
C2	VisualBERT w/o Early Fusion	61.4
C3	VisualBERT w/o BERT Initialization	64.7
C4	VisualBERT w/o Objective 2	64.9

Table 7: Performance of the ablation models on NLVR^2. Results confirm the importance of task-agnostic pre-training (C1) and early fusion of vision and language (C2).

entirely skipping such pre-training (VisualBERT w/o COCO Pre-training) and also by pre-training with only text but no images from COCO (VisualBERT w/o Grounded Pre-training). Both variants underperform, showing that pre-training on paired vision and language data is important.

C2: Early Fusion We include VisualBERT w/o Early Fusion to justify allowing early interaction between image and text features, confirming again that multiple interaction layers between vision and language are important.

C3: BERT Initialization All models discussed before are initialized from a pre-trained BERT. To understand its contribution, we introduce a variant that is randomly initialized and then trained as the full model. While it seems weights from language-only pre-trained BERT are important, performance does not degrade as much as we expect, arguing that the model is likely learning many of the same useful aspects about grounded language during COCO pre-training.

C4: The sentence-image prediction objective We introduce a model without the sentence-image prediction objective during pre-training (VisualBERT w/o Objective 2). Results suggest that this objective has positive but less significant effect, compared to other components.

Overall, the results confirm that the most important design choices are task-agnostic pre-training (C1) and early fusion of vision and language (C2). In pre-training, both the inclusion of additional COCO data and using both images and captions are paramount.