Improving Pre-trained Vision-and-Language Embeddings for Phrase Grounding

Zi-Yi Dou, Nanyun Peng
Department of Computer Science
University of California, Los Angeles
{zdou,violetpeng}@cs.ucla.edu

Abstract

Phrase grounding aims to map textual phrases to their associated image regions, which can be a prerequisite for multimodal reasoning and can benefit tasks requiring identifying objects based on language. With pre-trained vision-and-language models achieving impressive performance across tasks, it remains unclear if we can directly utilize their learned embeddings for phrase grounding without fine-tuning. To this end, we propose a method to extract matched phrase-region pairs from pre-trained vision-and-language embeddings and propose four fine-tuning objectives to improve the model phrase grounding ability using image-caption data without any supervised grounding signals. Experiments on two representative datasets demonstrate the effectiveness of our objectives, outperforming baseline models in both weakly-supervised and supervised phrase grounding settings. In addition, we evaluate the aligned embeddings on several other downstream tasks and show that we can achieve better phrase grounding without sacrificing representation generality.[1]

1 Introduction

Recent studies on vision-and-language pre-training (Tan and Bansal, 2019; Li et al., 2019; Lu et al., 2019; Su et al., 2019; Chen et al., 2020; Li et al., 2020b, 2021; Shen et al., 2021) demonstrate impressive performance across vision-and-language tasks, including image-text retrieval (Lin et al., 2014; Plummer et al., 2015), visual entailment (Xie et al., 2019) and visual question answering (Antol et al., 2015).

However, few existing papers have paid attention to the phrase grounding ability of their pre-trained embeddings, namely the ability to map natural language queries to their corresponding image regions, which can 1) benefit tasks requiring identifying objects based on language (Deng et al., 2018); 2) be a prerequisite for advanced multimodal reasoning (Plummer et al., 2015). Among the prior work, Li et al. (2020a) demonstrate certain grounding abilities of VisualBERT, yet their analysis is limited to attention heads and it is unclear how VisualBERT compares with state-of-the-art grounding models. Cao et al. (2020) provide insights on cross-modal interaction, but their analysis is primarily limited to the coreference relations between phrases and visual tokens.

In this paper, we study the phrase grounding ability of vision-and-language embeddings pre-trained on image-caption datasets. First, we propose a method to extract phrase-region pairs from the pre-trained embeddings without any fine-tuning. We find that while our method uncovers certain grounding abilities of the pre-trained embeddings, there is still much room for improvement. Therefore, we propose to fine-tune models with objectives designed for better aligning word and region representations on image-caption datasets. The fine-tuning objectives are designed to *maximize the symmetricity* between vision and language during fine-tuning for better phrase grounding while maintaining the representation transferability so that the learned representations are still useful for other downstream tasks. Specifically, we fine-tune models with 1) a *masked language modeling* objective conditioned on images; 2) an *adapted masked region modeling* objective with texts utilizing a dynamically constructed vision vocabulary; 3) a *modified object label prediction* objective that explicitly bridges the gap between vision and language; 4) a *proposed bidirectional attention optimization* objective encouraging the consistency between vision-to-language and language-to-vision alignments.

We fine-tune pre-trained models on COCO (Chen et al., 2015) and test them on two representative phrase grounding datasets,

[1]Code is available at https://github.com/pluslab/phrase_grounding.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6362–6371
November 7–11, 2021.

RefCOCO+ (Kazemzadeh et al., 2014) and Flickr30k Entities (Plummer et al., 2015). We find that our fine-tuning objectives can improve the model grounding ability significantly, improving baseline in both weakly-supervised and supervised phrase grounding settings. We also evaluate the aligned representations on several downstream tasks and show that our model can achieve better phrase grounding without sacrificing performance on other types of tasks.

2 Extracting Phrase-Region Pairs from Pre-Trained Embeddings

Formally, the phrase grounding task can be defined as: given an image $\mathbf{v}$ consisting of multiple regions $\langle v_1, \cdots, v_n \rangle$ and its corresponding caption $\mathbf{l}$ segmented into tokens $\langle l_1, \cdots, l_m \rangle$, for each noun phrase $\mathbf{p}_i = \langle l_{ix}, \cdots, l_{iy} \rangle$, a model needs to find its associated region v_j.

We first propose a way to directly extract the matched phrase-region pairs from pre-trained embeddings. Then, we evaluate this method on phrase grounding tasks with several popular pre-trained models, including LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020), ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019) and VL-BERT (Su et al., 2019).

2.1 Extraction Method

We propose to directly extract phrase-region pairs from pre-trained models based on representation similarities. Specifically, given an image $\mathbf{v}$ and its caption $\mathbf{l}$, we feed them to a pre-trained vision-and-language model and obtain their representations $h(\mathbf{v})$ and $h(\mathbf{l})$. Note that here representations of the k-th model layer are taken, where k is a hyper-parameter and is selected on the validation set.

Then, given a noun phrase $\mathbf{p}_i$, we average its token representations and get the phrase representation $h(\mathbf{p}_i) = \textsc{Mean}(\langle h(l_{ix}), \cdots, h(l_{iy}) \rangle)$. Afterwards, we score each candidate region v_j by computing the dot product between $h(\mathbf{p}_i)$ and $h(v_j)$. Regions with the highest scores are selected and we can measure the accuracy of the selected pairs.

2.2 Experiments

We evaluate the extraction method on RefCOCO+ using pre-trained models in a controlled setting (Bugliarello et al., 2020).

Model	RefCOCO+		
	val	testA	testB
LXMERT	13.62 (33.33)	10.41 (36.59)	16.53 (30.91)
UNITER	26.26 (43.27)	32.62 (50.90)	18.49 (35.80)
ViLBERT	14.47 (42.14)	10.79 (48.88)	18.76 (36.27)
VisualBERT	**33.26 (43.88)**	**33.59 (52.04)**	**33.34 (36.56)**
VL-BERT	23.52 (42.97)	32.54 (49.86)	13.93 (36.19)
Supervised	70.98	77.05	60.73

Table 1: Phrase grounding accuracy (%) of pre-trained models investigated with our proposed method. We also include the performance of probing classifiers (numbers in parenthesis) and a supervised VisualBERT model ('Supervised') for reference.

2.2.1 Setup

We follow the setting in Bugliarello et al. (2020) in this section. Specifically, all the vision-and-language models are pre-trained on a pruned Conceptual Captions dataset (Sharma et al., 2018), consisting of 2.77M images with weakly-associated captions automatically collected from billions of web pages. The image features are extracted using a Faster R-CNN (Ren et al., 2016) with a ResNet-101 backbone (Anderson et al., 2018) trained on the Visual Genome dataset (Krishna et al., 2017) and the vision-and-language models are trained with 36 extracted regions of interest.

2.2.2 Results

Table 1 shows the phrase grounding accuracy of the pre-trained models using our method. To provide upper-bound performance for our extraction method, we train a linear probing classifier with the frozen model embeddings as inputs (numbers in parenthesis). We find that our extraction method can better uncover the phrase grounding ability of single-stream models (UNITER, VisualBERT, VL-BERT), which process the vision and language inputs jointly. On the other hand, the grounding information in two-stream models (LXMERT, ViLBERT) can be hard to extract, probably because the parameters of two-stream models are not shared in the top layers and thus they are less likely to learn aligned representations.

Also, comparing with a supervised VisualBERT model, the pre-trained models can underperform their supervised counterparts by a large margin, indicating there is much room for improvement and additional efforts are required to align the pre-trained vision-and-language embeddings.

3 Aligning Pre-trained Vision-and-Language Embeddings

To improve the model phrase grounding ability, we then propose four fine-tuning objectives for vision-and-language models. We assume an image-caption dataset $\{\langle \mathbf{v}^k, \mathbf{l}^k \rangle\}$ is provided but no fine-grained phrase-region annotations are available. A pre-trained vision model (Anderson et al., 2018) is used to segment images into regions and produce region representations and object labels.

3.1 Fine-tuning Objectives

We investigate four objectives that fine-tune pre-trained vision-and-language models for phrase grounding:

Masked Language Modeling (MLM). MLM with images has proven to be useful for representation learning (Li et al., 2019) and here we investigate if it is also helpful for phrase grounding. Specifically, we randomly mask 15% of the tokens $\mathbf{l}$ and the model is trained to reconstruct $\mathbf{l}$ given the masked texts $\mathbf{l}^{mask}$ and regions $\mathbf{v}$:

$$\mathcal{L}_{\text{MLM}} = \log p(\mathbf{l}|[\mathbf{v}; \mathbf{l}^{mask}]). \quad (1)$$

Masked Region Modeling (MRM). Inspired by MLM, we propose its counterpart in the vision side to encourage the symmetricity between vision and language. While previous work (Tan and Bansal, 2019) regress the region features, we find it is unhelpful in our setting (in Appendix). Instead, by imitating MLM which uses a text vocabulary, we create a dynamic vision vocabulary on the fly, and the model tries to reconstruct the input regions given the dynamically constructed vocabulary.

Concretely, at each training step, we sample a batch of image-caption pairs $\{\langle \mathbf{v}^k, \mathbf{l}^k \rangle\}_{k=1}^B$ and randomly mask 15% of the regions, where B is the batch size. We treat all the regions in $\{\mathbf{v}^k\}_{k=1}^B$ as candidate regions, and for each masked region, the model needs to select the original region within the set of candidate regions given masked inputs. Denoting the pre-trained vision model representations and our model representations of $\{\mathbf{v}^k\}_{k=1}^B$ as $\{c(\mathbf{v}^k)\}_{k=1}^B$ and $\{h(\mathbf{v}^k)\}_{k=1}^B$ respectively, we can represent the output probability at position i for the k-th instance as:

$$p(\mathbf{v}_i^k|[\mathbf{v}^{k,mask}; \mathbf{l}^k]) = \frac{e^{\cos(h(\mathbf{v}_i^k), c(\mathbf{v}_i^k))}}{\sum_{j,k'} e^{\cos(h(\mathbf{v}_i^k), c(\mathbf{v}_j^{k'}))}},$$

where $\cos(\cdot, \cdot)$ refers to the cosine similarity.

The model is trained to maximize this probability similar to noise contrastive estimation (Gutmann and Hyvärinen, 2010; Jozefowicz et al., 2016):

$$\mathcal{L}_{\text{MRM}} = \log p(\mathbf{v}|[\mathbf{v}^{mask}; \mathbf{l}]). \quad (2)$$

Object Label Prediction (OLP). The object labels predicted by the pre-trained vision model provide us with good anchor points to bridge the gap between vision and language, and previous work has tried to incorporate the information by predicting the object labels for each region (Tan and Bansal, 2019; Chen et al., 2020). In this paper, to better share the information between the two modalities, we propose to 1) use simple heuristics to convert object labels into text tokens and train our model to predict the object labels $\mathbf{o_v}$ with a multi-class MLM objective; 2) share the classification layer of MLM and OLP.

For example, if the object label of $\mathbf{v}_i$ is "stop sign", we first tokenize it into "stop" and "sign", the model is then trained to maximize the joint probability of both the two tokens at $\mathbf{v}_i$:

$$\mathcal{L}_{\text{OLP}} = \log p(\mathbf{o_v}|[\mathbf{v}; \mathbf{l}]). \quad (3)$$

Bidirectional Attention Optimization (BAO). Inspired by the work on encouraging the consistency between forward and backward attentions (Cohn et al., 2016; Hu et al., 2020; Dou and Neubig, 2021), we propose an objective to encourage the symmetricity of vision-to-language and language-to-vision attentions. Specifically, after obtaining the representations $h(\mathbf{v})$ and $h(\mathbf{l})$, we compute the forward and backward attention matrices as:

$$\text{ATT}_{VL} = \text{SOFTMAX}(h(\mathbf{v})^{\text{T}} h(\mathbf{l})/\sqrt{d}),$$

$$\text{ATT}_{LV} = \text{SOFTMAX}(h(\mathbf{l})^{\text{T}} h(\mathbf{v})/\sqrt{d}),$$

where d denotes the feature dimension.

We then minimize the distance between them by maximizing the trace of $\text{ATT}_{VL}^{\text{T}} \text{ATT}_{LV}$:

$$\mathcal{L}_{\text{BAO}} = -\log(1 + \frac{\texttt{trace}(\text{ATT}_{VL}^{\text{T}} \text{ATT}_{LV})}{\min(|\mathbf{v}|, |\mathbf{l}|)}). \quad (4)$$

Combined Objective. Our final objective is a combination of the four objectives:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{MRM}} + \mathcal{L}_{\text{OLP}} + \alpha \mathcal{L}_{\text{BAO}}, \quad (5)$$

where α is selected from $\{0.1, 0.25, 0.5, 1.0\}$ and is set to 0.1 based on the validation performance.

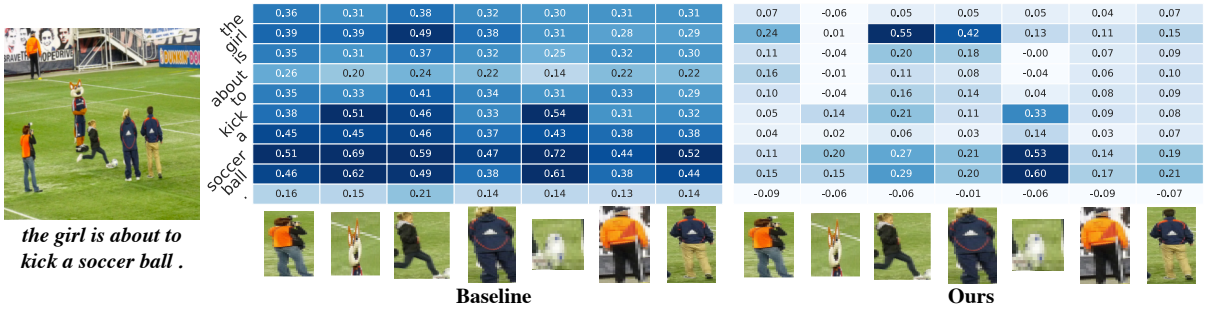

Figure 1: Visualizations of cosine similarities between text and image representations of the VisualBERT baseline and our model. Our model can learn more aligned representations than the baseline.

Model	Flickr30k	RefCOCO+	
		testA	testB
Weakly-Supervised			
MAF (Wang et al., 2020)	61.43	17.10	13.50
VisualBERT	34.53	42.19	35.44
Ours	**62.10**	**47.89**	**38.20**
Supervised			
VisualBERT	71.33	78.31	61.98
Ours	**72.49**	**78.64**	**62.86**

Table 2: Accuracy (%) in weakly-supervised and supervised grounding settings. Best scores are in **bold**.

Model	Flickr30k		RefCOCO+		
	val	test	val	testA	testB
Ours	59.59	62.10	42.79	47.89	38.20
-MLM	50.53	52.71	39.14	42.54	35.66
-MRM	51.21	53.49	40.58	44.53	36.87
-OLP	48.38	50.17	40.06	42.32	38.84
-BAO	57.20	59.19	41.49	44.48	37.12

Table 3: Ablation studies on each of our objectives. We measure the accuracy (%) numbers in weakly-supervised grounding settings.

Model	Flickr (Recall@1)	SNLI-VE (Accuracy)	VQAv2 (VQA-score)	
			test-dev	test-std
VisualBERT	58.94	76.41	69.68	69.92
Ours	59.84	76.83	69.89	70.16

Table 4: We can achieve better phrase grounding abilities while maintaining the representation transferability on other types of tasks, including image-text retrieval, visual entailment and visual question answering.

3.2 Experiments

We then train our model with the proposed objective and compare with several baselines.

3.2.1 Setup

Model/Datasets. We choose VisualBERT as our base architecture because it performs the best in Section 2 and pre-train it on COCO (Chen et al., 2015). We then further fine-tune models on COCO and evaluate them on RefCOCO+ and Flickr30k in both weakly-supervised and supervised settings. Details of the models and datasets are in Appendix.

Settings. In weakly-supervised settings where only the image-text pairs in COCO are given, we directly extract phrase-region pairs from models using our method in Section 2.1. In supervised settings where phrase-region annotations in RefCOCO+ and Flickr30k are available, we add a linear layer on top of each region representation and fine-tune models with the cross-entropy loss.

3.2.2 Results

We first present the main results of the models and some ablation studies of the training objectives.

Main Results. In the weakly-supervised settings, Table 2 demonstrates that our objectives can improve the model grounding ability significantly, outperforming all the baselines. Moreover, we find that while MAF (Wang et al., 2020) achieves strong performance on Flickr30k, it fails on RefCOCO+. We hypothesize that this is because MAF is based on static word embeddings and in the RefCOCO+ setting multiple objects of the same type will typically present in one image, making MAF unable to disambiguate the phrases. With the aligned representations, we can also achieve better grounding ability than VisualBERT in supervised settings.

Ablation Studies. We ablate each of our training objective and test their contributions in Table 3. We can see that all of the objectives are beneficial for phrase grounding, with OLP being the most effec-

tive one. BAO can bring marginal improvements, yet its contributions are still non-negligible. We also test most existing pre-training objectives in Appendix and show that our proposed objective works the best.

3.2.3 Analysis

We then perform analysis to provide insights on the fine-tuned model representations.

Transferring to Other Tasks. It is interesting to see if the aligned representations are still useful for other types of tasks. In Table 4, we test our model on image-text retrieval (Plummer et al., 2015), visual entailment (Xie et al., 2019) and visual question answering (Goyal et al., 2017) (details in Appendix). We find that our model can achieve comparable or superior performance compared with VisualBERT, especially on tasks relying more on the model grounding ability like image retrieval, which shows that our training paradigm can maintain the representation generality.

Qualitative Examples. We visualize the learned representations in Figure 1. We find it hard to observe clear patterns from the baseline representations. For example, while the token representation of "ball" have high similarity with its associated region embedding, it is also close to the representation of the mascot. By contrast, our model can clearly learn more aligned representations. It is interesting to note that our model can learn there is a partial correspondence between the word "kick" and the soccer ball region, indicating that our objectives can also align verb and region representations.

4 Related Work

We overview two lines of related work in this part.

Vision-and-Language Representation Learning. Learning multimodal representations has been an active research area (Ngiam et al., 2011; Silberer and Lapata, 2014; Hill and Korhonen, 2014; Hubert Tsai et al., 2017) and the progresses on model pre-training in computer vision (Doersch et al., 2015; Pathak et al., 2016) and natural language processing (Peters et al., 2018; Devlin et al., 2019) have motivated research on vision-and-language representation learning with pre-training (Tan and Bansal, 2019; Li et al., 2019; Lu et al., 2019; Su et al., 2019; Chen et al., 2020; Sun et al., 2019; Li et al., 2020b). The pretraining-finetuning paradigm has proven to be effective across tasks, such as image-text retrieval (Lin et al., 2014; Plummer et al., 2015), visual entailment (Xie et al., 2019) and visual question answering (Antol et al., 2015). A few studies (Li et al., 2020a; Cao et al., 2020) analyze the pre-trained multimodal models, while to the best of our knowledge, no existing work have focused on the phrase grounding ability of learned representations.

Phrase Grounding. Many vision-and-language tasks, such as visual question answering and vision-language navigation, rely on or can benefit from phrase grounding. Both supervised (Rohrbach et al., 2016a; Yu et al., 2018; Liu et al., 2020) and weakly-supervised (Rohrbach et al., 2016b; Yeh et al., 2018; Chen et al., 2018; Wang and Specia, 2019; Hessel et al., 2019; Wang et al., 2020) phrase grounding approaches have been proposed. While pre-trained vision-language models have been applied in vision grounding tasks in supervised settings (Chen et al., 2020; Li et al., 2020b), it is unclear whether the models can perform phrase grounding by directly using the representations learned during pre-training.

5 Conclusion

In this paper, we first propose a method to extract matched phrase-region pairs from pre-trained vision-and-language embeddings and evaluate its performance across models. Then, we propose several fine-tuning objectives for phrase grounding and demonstrate their effectiveness in both weakly-superivsed and supervised phrase grounding tasks. We also evaluate our aligned representations on other downstream tasks and show that we can achieve better phrase grounding without sacrificing the representation transferability to other downstream tasks. Future directions include better utilizing the aligned representations and incorporating our objectives into pre-training.

Acknowledgement

We would like to thank the anonymous reviewers for valuable suggestions and Liunian Harold Li, Da Yin, Te-Lin Wu for helpful discussions. This work is supported by the Machine Common Sense (MCS) program under Cooperative Agreement N66001-19-2-4032 with the US Defense Advanced Research Projects Agency (DARPA). The views and the conclusions of this paper are those of the authors and do not reflect the official policy or position of DARPA.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2020. Multimodal pretraining unmasked: Unifying the vision and language BERTs. *arXiv preprint*.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*.

Kan Chen, Jiyang Gao, and Ram Nevatia. 2018. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2018. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2020. Explicit alignment objectives for multilingual bidirectional encoders. *arXiv preprint*.

Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. 2017. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*.

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. *arXiv preprint*.

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

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2021. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. 2020. Learning cross-modal context graph for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? *arXiv preprint*.

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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016a. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016b. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint*.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Josiah Wang and Lucia Specia. 2019. Phrase localization without paired training examples. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. 2020. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint*.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. 2020. A fast proximal point method for computing exact wasserstein distance. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Raymond A Yeh, Minh N Do, and Alexander G Schwing. 2018. Unsupervised textual grounding: Linking words to image concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization (IJCAI)*.

A Implementation Details

In Section 2, we follow (Bugliarello et al., 2020) and pre-train the vision-and-language models in a controlled setting. Specifically, all the models are pre-trained on a pruned Conceptual Captions dataset (Sharma et al., 2018), consisting of 2.77M images with weakly-associated captions automatically collected from billions of web pages. The image features are extracted using a Faster R-CNN (Ren et al., 2016) with a ResNet-101 backbone (Anderson et al., 2018) trained on the Visual Genome dataset (Krishna et al., 2017) and the vision-and-language models are trained with 36 extracted regions of interest. For the probing experiments, we use the default hyper-parameters in (Bugliarello et al., 2020) for training the probing classifiers.

In Section 3, we pre-train VisualBERT on the COCO dataset (Chen et al., 2015), consisting of 413K captions for 82K images (each image is paired with five different captions). VisualBERT is pre-trained with its original objectives for 11K steps and with two RTX 2080 GPUs, taking about 40 hours per experiment. Then, our models are further fine-tuned on two RTX 2080 GPUs for 11K steps, taking about 2 days per experiment. The batch size is set to 480 and the learning rate is set to 5e-5. The models are trained with 64 extracted regions of interest. α in Equation 5 is selected from $\{0.1, 0.25, 0.5, 1.0\}$ based on the validation performance on Flickr30k. The image features and labels are extracted from a ResNeXT-152 Faster-RCNN model trained on Visual Genome with attribute loss. For efficiency, we mask both vision and language inputs and perform MLM, MRM, OLP jointly on the masked inputs instead of training models with these objectives sequentially. We also tried to fine-tune VisualBERT with its original objectives for phrase grounding, but the grounding performance did not get improved.

For the phrase grounding datasets we use, the RefCOCO+ dataset is collected in an interactive game interface and we follow its standard split. During test, RefCOCO+ provides person vs. object splits for evaluation, where images containing multiple people are in "testA" and images containing multiple other objects are in "testB". The Flickr30k Entities dataset contains 224K phrases and 31K images in total, where each image is associated with five captions, and we follow its standard splits. Following previous work, we consider a prediction to be correct if the IoU (Intersection of Union) score between our predicted bounding box and the ground-truth box is larger than 0.5. We fine-tune the models for 20K steps, with the batch size set to 32 and the learning rate set to 2e-5. The models are trained with 100 extracted regions of interest.

For the image-text retrieval task, we evaluate models on Flickr30k (Plummer et al., 2015) with Recall@1 as the evaluation metric. The models are fine-tuned for 20 epochs with the batch size set to 256 and the learning rate set to 1e-4. The models are trained with 36 extracted regions of interest. For the visual entailment task, we experiment on the SNLI-VE dataset (Xie et al., 2019) and test the accuracy numbers. We fine-tune the models for 60K steps with the batch size set to 480 and the learning rate set to 5e-5. The models are trained with 100 extracted regions of interest. For the visual question answering task, we choose the VQAv2 dataset (Goyal et al., 2017) and evaluate models with the VQA-score[2] on both test-dev and test-standard datasets. The models are fine-tuned for 60K steps with the batch size set to 480 and the learning rate set to 5e-5. The models are trained with 100 extracted regions of interest.

[2]https://visualqa.org/evaluation.html

B Negative Results

In this part, we show some negative results of four fine-tuning objectives that we have tried in our settings.

B.1 Fine-tuning Objectives

In addition to the objectives presented in the main content, we also experiment with the following four objectives:

Masked Region Regression (MRR). Previous work (Tan and Bansal, 2019; Chen et al., 2020) have attempted to regress the region features by minimizing the L2 distance between the predicted and the original image features. An additional feed-forward layer is used to transform the hidden representations into the image feature space.

Masked Region Classification (MRC). Similar to our OLP objective, researchers (Tan and Bansal, 2019; Lu et al., 2019; Su et al., 2019; Chen et al., 2020) have also tried to utilize object labels by predicting the object semantic class without sharing the classification layer between vision and language modalities. The object labels are obtained from a pre-trained vision model. The main difference between MRC and our OLP is that we perform image classification in the text space and share the prediction layer between the two modalities.

Image-Text Matching (ITM). In ITM, a special token ([CLS]) is inserted at the beginning of the input sentence and it tries to learn a fused representation of both vision and language. We feed the model with either matched or mismatched image-caption pairs with equal probability. A classifier is added on the top of this token and its output is a binary label, indicating if the sampled image-caption pair is a match.

Optimal Transport (OT). Chen et al. (2020) use optimal transport to encourage word-region alignments, which is potentially beneficial for phrase grounding. Therefore, we follow their settings and implement the optimal transport objective. Specifically, for each pair of word l_i and region v_j, we first compute their cosine distance $c_{ij} = 1 - \frac{l_i^T v_j}{||l_i||_2||v_j||_2}$. Then, the optimal transport objective is defined as:

$$\mathcal{L}_{OT} = \min_{T} \sum_i \sum_j T_{ij} c_{ij},$$

Model	Flickr30k		RefCOCO+		
	val	test	val	testA	testB
Ours	59.59	62.10	42.79	47.89	38.20
+MRR	57.77	60.52	41.80	46.32	37.38
+MRC	58.13	60.35	41.96	46.32	37.81
+ITM	52.72	55.42	41.10	45.41	37.22
+OT	43.44	45.33	39.98	43.40	35.62
-MRM	51.21	53.49	40.58	44.53	36.87
-MRM+MRR	50.98	52.50	40.71	45.96	35.93
-OLP	48.38	50.17	40.06	42.32	38.84
-OLP+MRC	48.75	50.98	40.39	43.77	37.44
-BAO	57.20	59.19	41.49	44.48	37.12
-BAO+OT	43.00	44.78	39.63	43.17	35.71

Table 5: Negative results of four training objectives. Our proposed objectives (MRM, OLP, BAO) are better than previous methods (MRR, MRC, OT).

where T_{ij} is the transport plan between language and vision and is obtained using the IPOT algorithm (Xie et al., 2020).

B.2 Results

Table 5 shows the results of the four fine-tuning objectives on Flickr30k and RefCOCO+ in weakly-supervised settings. We can see that adding these objectives cannot improve the model performance in our phrase grounding settings. We think this is possibly because 1) the MRR and MRC objectives differ a lot from the MLM objective in the language part, and thus they can deviate the resulting vision-and-language representations; 2) ITM mainly cares about aligning sentence and image representations, while our phrase grounding tasks require fine-grained phrase-region alignments; 3) there can be multiple complicated many-to-many alignments for an image-caption pair, making it hard to find a reasonable transport plan between language and vision modalities, and thus the optimal transport techniques may not be not suitable for phrase grounding. Also, as shown in the last 6 rows of the table, our proposed MRM, OLP, BAO objectives are better than the MRR, MRC, OT objectives that previous work use.

C Phrase Grounding Abilities Across Layers

In this part, we plot the grounding performance of each model layer in Figure 2. Contrary to findings in multilingual encoders (Pires et al., 2019), we do not see coherent patterns from the performance of different models. While most models demonstrate

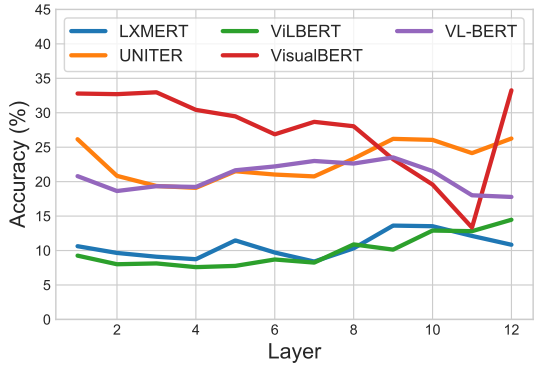

Figure 2: The phrase grounding ability across layers evaluated with representation similarity measures.

better grounding abilities in the top and bottom layers than the middle layers, VL-BERT exhibits an opposite behavior.