Adaptive Fine-tuning for Vision and Language Pre-trained Models

Shentong Mo[†,*] **Jingfei Xia**[†]**, Ihor Markevych**
Carnegie Mellon University
Pittsburgh, PA 15213
`shentonm,jingfeix,imarkevy`@andrew.cmu.edu

Abstract

Visual and linguistic pre-training aims to learn vision and language representations together, which can be transferred to visual-linguistic downstream tasks. However, there exists semantic confusion between language and vision during the pre-training stage. Moreover, current pre-trained models tend to take lots of computation resources for fine-tuning when transferred to downstream tasks. In this work, we present a simple but effective approach for Adaptive Fine-tuning of Vision and Language pre-trained models, namely **AFVL**. Specifically, we introduce a pair-wise contrastive loss to learn alignments between the whole sentence and each image in the same batch during the pre-training process. At the fine-tuning stage, we introduce two lightweight adaptation networks to reduce model parameters and increase training speed for saving computation resources. We evaluate our AFVL on four main downstream tasks, including Visual Question Answering (VQA), Visual Commonsense Reasoning (VCR), Natural Language for Visual Reasoning (NLVR), and Region-to-Phrase Grounding (RPG). Compared to previous methods, our AFVL achieves comparable or better results while saving training time and GPU memory by a large margin for fine-tuning. Extensive experiments and ablation studies demonstrate the efficiency of contrastive pre-training and adaptive fine-tuning proposed in our AFVL.

1 Introduction

Visual and language representations pre-training [17, 19, 29] has been an active research area in the multi-modal community, as it allows for the usage of pre-trained models that achieve state-of-the-art comparable results for a variety of tasks without spending significant compute time for modeling language and visual distributions by leveraging features created by available pre-trained models. With a better pre-training model, it can be used in a variety of areas in visual and language fields such as Visual Question Answering (VQA) and Visual Commonsense Reasoning (VCR). In various architectures designed for different visual-linguistic tasks, a key point is to aggregate the multi-modal information in both visual and linguistic domains. However, there exists *semantic confusion* between vision and language at the pre-training stage, that is, misalignment between object/entities or image/text. Moreover, when transferred to downstream tasks, pre-trained models tend to take much training time and resources for fine-tuning.

In this work, we propose a simple but effective framework based on BERT for adaptive fine-tuning of vision and language pre-trained models, called AFVL. Specifically, to eliminate the misalignment between language and vision during pre-training, we apply a pair-wise contrastive loss to learn alignments between the whole sentence and each image, where we maximize the cosine similarity of visual and linguistic embeddings from correct pairs while minimizing the cosine similarity of

---

*Corresponding author,[†]These authors contributed equally to this work.

embeddings of false pairs. To further eliminate the need for much training time at the fine-tuning stage, we introduce two lightweight adaptation networks to learn adaptive representations in our AFVL. One adapter is to use a shortcut block to obtain task-specific features and merge generalized features from the pre-trained model with the output block for final predictions; the other adapter is to apply a bottleneck structure after attention and feed-forward modules in each layer in BERT. Our AFVL not only reduces the training parameters greatly but also maintain the performance at a comparable level.

We conduct extensive experiments on four main downstream tasks: VQA, VCR, Natural Language for Visual Reasoning, and Region-to-Phrase Grounding. Compared to previous state-of-the-art models, our AFVL achieves comparable or even better performance when transferred to visual-linguistic downstream tasks. Contrastive pre-training assists our AFVL in better understanding the relationship between image and text which improves results on downstream tasks. Ablation studies on adaptive fine-tuning also demonstrate the effectiveness and efficiency of the proposed adaptive fine-tuning in saving computation resources.

Our main contributions in this work can be summarized as follows:

- We propose a simple but effective approach for learning alignments between visual and linguistic representations during pre-training, namely AFVL.
- We present two lightweight adaptation networks in our AFVL to further ease the need for large computation resources at the fine-tuning stage.
- Our AFVL achieves comparable or better results when transferred to four main visual-linguistic downstream tasks.
- Extensive ablation studies demonstrate the efficiency of adaptive fine-tuning in reducing training parameters while achieving comparable performance.

2   Related Work

Visual representations pre-training. In recent years, visual representations pre-training has been applied to many downstream tasks, such as image classification, object detection, and segmentation. Typically, contrastive self-supervised learning [3, 4, 5, 13] is one of the popular methods to learn meaningful visual representations. Most previous methods learn visual representations from text paired with images in unsupervised, self-supervised, weakly supervised, and supervised. Since language and vision can share a similar semantic meaning, CLIP [22] is a commonly-used neural network trained on a variety of (image, text) pairs for learning transferable visual representations from natural language supervision. With the instruction of natural language and task-agnostic optimization, CLIP can predict the most relevant text snippet given an image, which is similar to the zero-shot capabilities of GPT-2 [24] and GPT-3 [2]. Huo *et al.* [11] apply a cross-modal contrastive learning framework called BriVL for image-text pre-training. Unlike CLIP that adopts a simple contrastive learning method, they make use of MoCo [8] for the cross-modal scenario by building a large queue-based dictionary to introduce negative samples in a batch for the pre-training. In this work, misalignments between visual and linguistic embeddings are mitigated by a pair-wise contrastive loss at the pre-training stage for the multi-modal scenario.

Linguistic representations pre-training. In the language modeling literature, there are two main frameworks for linguistic representations pre-training, including BERT [7] and GPT [23]. BERT [7] is a transformers-based [32] pre-trained model, which advanced state-of-the-art results for a lot of natural language processing tasks with self-attention module. GPT [23] is another language modeling architecture that is based on transformers [32]. Input for GPT models is represented on a byte level with an exception of spaces which allows handling variable vocabularies and deal with unknown during training time tokens. In this work, we introduce a simple but effective framework based on BERT and evaluate our network on four main visual-linguistic downstream tasks.

Fusion of visual and linguistic representations pre-training. There is a bunch of work [21, 28, 33, 6, 30, 31, 17, 19, 29, 9, 16] focusing on the fusion of visual and linguistic representations pre-training. Typically, UNITER [6] aims at learning to join image-text embeddings by using transformer on multi-model inputs with Masked Language Modeling (MLM), Masked Region Modeling, Image-Text Matching, and Word-Region Alignment as pre-training tasks. LXMERT [31] is a transformer based model with encoders: an object relationship encoder, a language encoder, and a cross-modality

encoder. VisualBERT [17] is a simple and flexible framework used for vision-and-language tasks, where they use the self-attention module in BERT structure to combine the image embedding in vision and text embedding in language. ViLBERT [19] extends the traditional BERT by using two parallel BERT-type models that operate over text segments and image regions. VL-BERT [29] is another BERT-based model that takes regions of interest from images and sub-word information, where they pre-trains the model by predicting masked words with image clues and predicting masked regions with text clues. To address the noisy label and domain bias problems, CVLP [27] introduces a contrastive loss in the visual branch to discriminate between positive examples and negative ones. More recently, DocFormer [1] is proposed to enforce the multi-modal interaction between visual, text and spatial features for Visual Document Understanding. However, in this paper, we adopt the pair-wise contrastive loss in both visual and linguistic branches to eliminate the misalignment between the whole sentence and each image during pre-training.

Adaptation networks in transformers. Adaptation is an important way when fine-tuning the BERT model, which allows us to achieve a promising result by updating much fewer parameters with less time and computation resources. This motivates researchers to use efficient adaptation methods in transformer-based models. MAD-X [20] adapts a multilingual model to arbitrary tasks and languages, where authors propose an invertible adapter and show good performance on different languages and tasks. Houlsby *et al.* [10] propose an intermediate layer inside transformer layers and train the intermediate layer with all other parameters frozen. In this paper, we propose two lightweight adapters to achieve comparable performance on domain-specific tasks with significantly reduced computational resources.

3 Method

In this part, we propose a simple but effective approach for learning contrastive and adaptive representations of vision and language, namely AFVL, which consists of contrastive pre-training and adaptive fine-tuning. Specifically, contrastive pre-training is applied to mitigate the semantic confusion between visual and linguistic representations at the pre-training stage. When transferred to downstream tasks during the fine-tuning process, adaptive fine-tuning eliminates the need for much training time and large GPU memories.

3.1 Contrastive pre-training

For contrastive pre-training, we adopt a self-attention mechanism within the BERT-based transformer to explicitly align elements of the input text and regions in the input image in a contrastive manner. Our AFVL consists of three main components: linguistic pre-training, visual pre-training, and contrastive fusion of visual-linguistic pre-training.

Linguistic pre-training. For language embeddings in the pre-training, we input three types of embeddings: 1) a token embedding $\mathbf{e}^t$ for each subword in a sentence; 2) a segment embedding $\mathbf{e}^s$ indicating which part of the text the token is from; 3) a position embedding $\mathbf{e}^p$ for the position of the token in the sentence. Then we sum up all three embeddings in a contextual representation $\mathbf{e}_n, n \in \{1, 2, ..., N\}$, where N denotes the number of subwords in the sentence. After being fed into a BERT-based transformer, those contextual embeddings become $\mathbf{e}'_i$. We adopt two similar objectives as BERT, including masked language modeling (MLM) and next sentence prediction (NSP). For the former objective, we randomly mask some parts of the input tokens with a special token (*i.e.*, [MASK]), and the model is trained to predict the masked token. As for NSP, we train the model using the embedding [CLS] to classify whether a pair of given sentences are consecutive in a context.

Visual pre-training. For vision features in the pre-training, we extract image ROIs from an objection detection framework (*i.e.*, Faster R-CNN) as the input $\mathbf{f}_k, k \in \{1, 2, ..., K\}$, where K is the number of image ROIs. The input $\mathbf{f}_k$ is also composed of three types of visual embeddings: 1) an image feature embedding $\mathbf{f}^i$ for each image ROI; 2) a segment embedding $\mathbf{f}^s$ indicating which token embedding the image embedding is opposed to; 3) a position embedding $\mathbf{f}^p$ for alignments between tokens and each image ROI. Following Visual-BERT in task-agnostic pre-training, we apply two captions for a text segment in the COCO dataset, where there are multiple captions corresponding to one image. Particularly, we use one of the captions as ground truth to describe the image, while we also apply a 50% probability to choose a caption from those two captions. Our model is trained to distinguish whether the caption is the ground truth or randomly drawn from two captions.

Visual-linguistic pre-training. To mitigate the semantic confusion between language and vision, we design a pair-wise contrastive learning mechanism on visual and linguistic representations from the multi-layer transformer. Specifically, we calculate the cosine similarity between each pair of linguistic embeddings $\mathbf{E}'_b$ and visual embeddings $\mathbf{F}'_b$ in a batch of size B, where $b \in \{1, 2, ..., B\}$. Then, those similarities are jointly learned for alignments between the whole sentence and each image in the same batch, where we maximize the cosine similarity of the visual and linguistic embeddings of the B correct pairs in the batch while minimizing the cosine similarity of the embeddings of the $B^2 - B$ false pairings. We apply a pair-wise contrastive loss over these similarities scores for optimization. Specifically, we define the Pair-wise Contrastive Loss (PwCL) between linguistic embeddings $\mathbf{E}'_i$ and visual embeddings $\mathbf{F}'_j$ as:

$$\mathcal{L}_{\text{PwCL}} = -\log \frac{\sum_{i=1}^{B}(\mathbf{E}'_i \cdot \mathbf{F}'_i)}{\sum_{i=1}^{B}\sum_{j=1}^{B} \mathbb{1}_{i \neq j}(\mathbf{E}'_i \cdot \mathbf{F}'_j)} \tag{1}$$

where $\mathbb{1}_{i \neq j}$ is an indicator function to check if linguistic embeddings $\mathbf{E}'_i$ and visual embeddings $\mathbf{F}'_j$ are aligned or not. In this way, we maximize the cosine similarity of visual and linguistic embeddings from correct pairs while minimizing the cosine similarity of embeddings of false pairs. Intuitively, alignments between the whole sentence and each image are learned in our AFVL to mitigate the semantic confusion existing in the pre-training process.

Note that this kind of contrastive pre-training methodology is different from a recent powerful framework, CLIP [22]. Concretely, we build a image-text alignment pre-training framework for the multi-modal scenario. First, this idea in CLIP is used to train a model that performs better in image classification with natural language supervision while in this paper we focus on solving multi-modal problems in visual and linguistic area. Second, they train an image encoder and a text encoder to generate vision and language embeddings. However, we combine two encoders into one BERT model and show that pair-wise contrastive learning enables model to learn better representations of image and text.
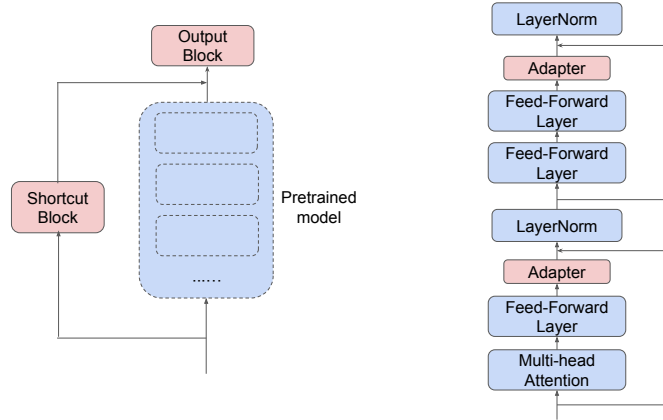

Figure 1: Adapters for adaptive fine-tuning in AFVL. Adapter I (**Left**): This adaptation method freezes the pre-trained model and trains a shortcut block and output block during fine-tuning. Adapter II (**Right**): This adaptation method follows [10] which adds an adapter right before each LayerNorm layer in the BERT. During the fine-tuning, it freezes feed-forward layers and attention modules while updating parameters in the adapter and LayerNorm layer. Blue blocks denote the frozen part and parameters in red blocks are updated during the fine-tuning phrase.

3.2 Adaptive fine-tuning

In this section, we design one adaptation methods and compared that with other efficient methods to fine tune on down-streaming tasks. One idea is that the frozen part in the model provides basic information for the generalized problem while the updated part generates a task-specific feature. Inspired by this idea, we proposed a method that used a shortcut block aside from the pre-trained model and merges the output from the pre-trained model and shortcut block with an additional output block. The pre-trained model obtains generalized features between image and text, and the shortcut

block acts as the selection neuron to capture the feature in each specific task. Then we apply an output block to combine generalized feature and task-specific feature together to get the final result. There are fewer layers in shortcut and output block than those in pre-trained model so the training time is reduced. We denote this type of adapter as **Adapter I**.

Motivated by Neil *et al.* [10], we propose another adaptation method denoted as **Adapter II** by adding a bottleneck structure within each BERT layer. Specifically, an adapter is added at the end of the attention module and another one at the end of the feed-forward layer. The adapter output and input of both the attention module and feed-forward layer are added together to pass through the LayerNorm module. During the fine-tuning process, we freeze the attention module and feed-forward layer. So, the adapter acts as a projection module which reflects the generalized feature to task-specific features. The adapter with fewer parameters is designed as a bottleneck structure, which contains one linear layer, a GELU function, and another linear layer.

We show the network details about two types of adapters for adaptive fine-tuning in Figure 1. We also conduct a comprehensive ablation study on the adapters in Section 4.3.

4 Experiments

4.1 Pre-training & Implementation Details

Following previous work [17, 19, 29], we apply the same setting for a fair comparison with those baselines. Specifically, we pre-train our AFVL on MS COCO [18] and Visual Genome [25]. For visual tokens, we apply the pre-trained Faster R-CNN [26] to extract the image ROIs (at most 100 ROIs with detection scores higher than 0.5 for each image). We apply Adam [15] for optimization and use a total batch size of 512 for 10 epochs. We use the warm-up step number of 15% of the total training steps. The pre-training and fine-tuning costs are 88 and 10 hours on 4 Tesla V100-32G GPUs respectively.

4.2 Downstream Tasks

We evaluate our AFVL pre-trained models on four downstream tasks: (1) Visual Question Answering (VQA), (2) Visual Commonsense Reasoning (VCR), (3) Natural Language for Visual Reasoning (NLVR^2), and (4) Region-to-Phrase Grounding (Flickr30K). Unless otherwise specified, we adopt Adapter I as the adaptive fine-tuning method.

Table 1: Comparison results on the NLVR^2 dataset.

Model	Dev	Test-P	Test-U	Test-U (Cons)
LXMERT [31]	-	74.45	76.20	42.10
VisualBERT [17]	67.40	67.00	67.30	26.90
UNITER [6]	76.93	75.58	-	-
CVLP [27]	-	76.20	-	-
AFVL (ours)	**79.16**	**78.31**	**79.87**	**46.23**

Natural Language for Visual Reasoning (NLVR). Following previous work [17], we evaluate our AFVL pre-trained models on the NLVR^2 dataset for joint reasoning about natural language and images. In this task, we focus on predicting whether a natural language caption match with a pair of images. We report the comparison results with state-of-the-art methods in Table 1. As can be seen, our AFVL achieves the new state-of-the-art performance in terms of all experimental settings compared to previous methods. This demonstrates the effectiveness of the pair-wise contrastive loss incorporated in both visual and linguistic branches to mitigate the existing semantic confusion between vision and language at the pre-training stage. The results also shows the advantage of our AFVL in joint reasoning about language and vision.

Region-to-Phrase Grounding (RPG). In order to test the performance of our AFVL on RPG, we fine-tune our AFVL pre-trained models on the Flickr30K Entities dataset consisting of 30k images and 250k annotations. Following BAN [14] and VisualBERT [17], we utilize image features from a pre-trained Faster R-CNN. For fine-tuning on Flickr30K Entities dataset, a self-attention block is applied to generate the average attention weights to predict the alignment between bounding boxes and phrases. Therefore, the box with the most attention from the phrase is the model prediction for a phrase to be grounded.

In this task, we need to choose the right bounding regions in an image that spans from a sentence belong to. Table 2 reports the comparison results with existing methods. We can observe that our AFVL outperforms previous multi-modal methods by a large margin under the same pre-trained setting, which demonstrates the effectiveness of our AFVL in visual-linguistic grounding tasks.

Table 2: Comparison results on the Flickr30K Entities dataset.

Model	R@1		R@5		R@10	
	Dev	Test	Dev	Test	Dev	Test
BAN [14]	-	69.70	-	84.20	-	86.40
VisualBERT [17]	70.40	71.33	84.49	84.98	86.31	86.51
MDETR [12]	78.90	-	88.80	-	90.80	-
AFVL (ours)	**80.35**	**81.76**	**90.63**	**91.12**	**93.16**	**94.21**

4.3 Ablation Study

In this section, we perform extensive ablation studies on the effect of adapters on the final performance of our AFVL, and the efficiency of the proposed adapters (Adapter I and II). Unless specified, we conduct all ablation studies on the VQA 2.0 dataset and report the mean and standard deviation of all results with 5 random seeds.

Effect of adapters. As shown in Table 3, we compare the performance of various models in terms of training parameters, fine-tuning costs, and accuracy on test-dev, test-std sets. We observe that all AFVL based models achieve better results than the baselines. This further demonstrates the effectiveness of contrastive visual and linguistic pre-training in eliminating the semantic confusion during the pre-training process. The proposed AFVL jointly learns alignments between the whole sentence and each image in the same batch to improve the pre-trained model's generalizability. In the meanwhile, our AFVL with both adapters with fewer training parameters achieves comparable performance to baselines with large fine-tuning parameters, which validates the efficiency of our proposed adapters in the AFVL on fine-tuning fewer parameters to save computation resources.

We also compare our AFVL with two types of adapters with current multi-modal methods [17, 29] in terms of the average fine-tuning cost to evaluate how much time our AFVL could save during the fine-tuning phase. Typically, both AFVL variants (Adapter I /II) achieves better results than previous work on both test-dev and test-std settings, which reducing remarkable parameters and costs for fine-tuning. While Adapter I performs slightly better than Adapter II, Adapter II reduces the fine-tuning parameters and costs by a large margin, *i.e.*, 59.40% and 76.17%. This implies the efficiency of our AFVL in the fine-tuning stage to learn effective representations for visual-linguistic downstream tasks.

Table 3: Ablation study on variants of adapters proposed in our AFVL.

Model	Fine-tune Params (M)↓	test-dev ↑	test-std ↑	Fine-tune Cost (h) ↓
VisualBERT [17]	113.90	70.08	71.00	24.5
VL-BERT$_{base}$ [29]	115.04	71.16	-	25.6
VL-BERT$_{large}$ [29]	342.55	71.79	72.22	70.2
AFVL (Adapter I)	98.40	**72.83±0.05**	**73.05±0.06**	10.0
AFVL (Adapter II)	**46.70**	72.67±0.08	72.86±0.11	**6.1**

5 Conclusion

In this work, we propose a simple but effective framework for learning contrastive and adaptive representations of vision and language, called AFVL, which involves contrastive pre-training and adaptive fine-tuning. The pair-wise contrastive loss is applied to mitigate the semantic confusion between language and vision during pre-training. To further eliminate the need for much computation time at the fine-tuning stage, we successfully introduce two lightweight adapters in our AFVL. Our AFVL achieves competitive performance on four main downstream tasks, including Visual Question Answering, Visual Commonsense Reasoning, Natural Language for Visual Reasoning, and Region-to-Phrase Grounding. We also conduct extensive experiments and ablation studies to demonstrate the efficiency of contrastive pre-training and adaptive fine-tuning proposed in our AFVL.

References

[1] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding. *arXiv preprint arXiv:2106.11539*, 2021.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020.

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *Advances in Neural Information Processing Systems*, pages 22243–22255, 2020.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

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

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pref-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[9] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. VLN BERT: a recurrent vision-and-language bert for navigation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[10] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799, 2019.

[11] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, Wanqing Cui, Danyang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chuhao Jin, Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanyan Lan, Wayne Xin Zhao, Ruihua Song, and Ji-Rong Wen. WenLan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021.

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, pages 18661–18673, 2020.

[14] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, page 1571–1581, 2018.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is More: clipbert for video-and-language learning via sparse sampling. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[17] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[20] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer, 2020.

[21] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. ImageBERT: cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2019.

[24] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[25] Oliver Groth Justin Johnson Kenji Hata Joshua Kravitz Stephanie Chen Yannis Kalantidis Li-Jia Li David A. Shamma Michael S. Bernstein Fei-Fei Li Ranjay Krishna, Yuke Zhu. Visual Genome: connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[27] Lei Shi, Kai Shuang, Shijie Geng, Peng Su, Zhengkai Jiang, Peng Gao, Zuohui Fu, Gerard de Melo, and Sen Su. Contrastive visual-linguistic pretraining. *arXiv preprint arXiv:2007.13135*, 2020.

[28] Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. KVL-BERT: knowledge enhanced visual-and-linguistic BERT for visual commonsense reasoning. *arXiv preprint arXiv:2012.07000*, 2020.

[29] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[30] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*, 2019.

[31] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, page 5998–6008, 2017.

[33] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.