Transformer-Exclusive Cross-Modal Representation for Vision and Language

Andrew Shin
Sony Group Corporation
andrew.shin@sony.com

Takuya Narihira
Sony Group Corporation
takuya.narihira@sony.com

Abstract

Ever since the advent of deep learning, cross-modal representation learning has been dominated by the approaches involving convolutional neural networks for visual representation and recurrent neural networks for language representation. Transformer architecture, however, has rapidly taken over the recurrent neural networks in natural language processing tasks, and it has also been shown that vision tasks can be handled with transformer architecture, with compatible performance to convolutional neural networks. Such results naturally lead to speculation upon the possibility of tackling cross-modal representation for vision and language exclusively with transformer. This paper examines transformer-exclusive cross-modal representation to explore such possibility, demonstrating its potentials as well as discussing its current limitations and its prospects.

1 Introduction

While early cross-modal models handled visuolinguistic tasks with template-based methods (Barbu et al., 2012; Elliott and Keller, 2013), or as a retrieval model (Farhadi et al., 2010; Ordonez et al., 2011), the advent of deep learning introduced end-to-end learning models for cross-modal tasks, in which convolutional neural networks (CNNs) (Krizhevsky et al., 2012) are employed for vision representation, whereas recurrent neural networks (RNNs), such as LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014), are employed for language representation. While a plethora of variations exist, most models proposed in the past few years have invariably relied on the CNN-RNN approach.

Such standardized scheme, however, started to change with the introduction of transformer architecture based on multi-head attention mechanism (Vaswani et al., 2017), which rapidly started to achieve state-of-the-art performance in natural language processing (NLP) (Peters et al., 2018; Dai et al., 2019; Yang et al., 2019) and speech recognition domains (Dong et al., 2018; Wang et al., 2020b), frequently outperforming RNNs. Furthermore, large-scale models based on transformer architecture, such as BERT (Devlin et al., 2019) or GPT-3 (Brown et al., 2020), started to appear, demonstrating that pre-training a sufficiently large model with a very large amount of data results in strong performance with versatility for various downstream tasks.

The success of transformer-based models in NLP and speech recognition naturally led to its adaptation in cross-modal tasks. (Lu et al., 2019) proposed ViLBERT, a pioneering BERT-inspired work that proposed to tokenize the images for compatibility with transformer architecture, and also to extend the pre-training objectives of BERT to reflect the nature of cross-modality. Many other cross-modal models followed, but mostly with similar approaches for image tokenization and pre-training objectives. This line of transformer-based cross-modal works described above, however, still heavily relied on CNN-based models, such as Faster R-CNN (Ren et al., 2015), to extract features from images, and the application of transformer was mostly limited to language representation.

Inspired by the observations made by recent works (Dosovitskiy et al., 2020), which demonstrate that vision tasks can be handled solely by transformer architecture with compatible performances to CNN-based models, this paper examines cross-modal representation for visuolinguistic tasks relying exclusively on transformer architecture, without using CNNs or RNNs. Without any structural modifications or advanced common embedding scheme, and without additional cross-modal pre-training that can be expensive both

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2719–2725
August 1–6, 2021. ©2021 Association for Computational Linguistics

computationally and data-wise, our model demonstrates comparable performances to conventional approaches based on CNNs and RNNs in exemplary cross-modal tasks.

2 Related Works

ViLBERT (Lu et al., 2019) was one of the first models to extend transformer architecture to cross-modal visuolinguistic tasks. They propose co-attentional transformer, in which separate transformer modules for each modality run in parallel, with the key and value inputs from one modality entering the transformer block for the other modality, thereby learning cross-modal dependence. In order to tokenize the image, they extract image regions using Faster R-CNN (Ren et al., 2015) along with 5-dimensional location vector. They also extend two unique pre-training objectives of BERT, namely masked language modeling and next-sentence prediction, to cross-modal setting, as masked multi-modal learning and image-sentence alignment classification. In masked multi-modal learning, visual tokens, along with language tokens, are randomly masked, and the model is trained to predict their probability distribution over object classes. In image-sentence alignment classification, a sequence of visual tokens and a sentence are juxtaposed, and the model performs a binary classification task, predicting whether the sentence describes the contents of the image. Many other models, such as VisualBERT (Li et al., 2019), LXMERT (Tan and Bansal, 2019) and Unicoder-VL (Li et al., 2020), also follow nearly identical pre-training objectives as VilBERT. On the other hand, UNITER (Chen et al., 2020) demonstrates improved performance by introducing additional pre-training objective of word region alignment, while MiniVLM (Wang et al., 2020a) achieves comparable performance with up to 70% fewer parameters by utilizing EfficinetNet (Tan and Le, 2019) with their own Compact BERT model.

While all models described above rely on CNN-based models to extract features from images, limiting the scope of applicability of transformer, recent works have demonstrated results that may imply a potential change in such workflow. (Dosovitskiy et al., 2020) proposed Vision Transformer (ViT), which demonstrates that pure transformer architecture without convolution can achieve comparable performance in image classification tasks, while requiring substantially less computational costs. Furthermore, (Touvron et al., 2021) showed via data-efficient image transformers (DeiT) that competitive performance can be achieved with training only on ImageNet (Deng et al., 2009) with no external data.

3 Model

We employ separate transformer models for vision and language, although internal mechanisms are essentially identical. Following (Dosovitskiy et al., 2020), we split an image into N patches x_p of $P \times P$ pixels, each of which is linearly projected into D-dimensional patch embedding, where $P = 16$, and $D = 768$. A learnable embedding x_{class} is prepended to patch embeddings, and positional embeddings are also added. The input sequence z_0 subsequently undergoes alternating layers of layer normalization (Ba et al., 2016) and multi-head self-attention, followed by a 2-layer MLP with GELUs (Hendrycks and Gimpel, 2020) as non-linearity:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos}, \quad (1)$$

$$E \in \mathbb{R}^{(P^2 C) \times D}, E_{pos} \in \mathbb{R}^{(N+1)+D}$$

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, l = 1...L \quad (2)$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l, l = 1...L \quad (3)$$

$$y_{img} = \text{LN}(z_L^0) \quad (4)$$

where

$$\text{MSA}(X) = W_{att}[\text{Att}_1(X), ..., \text{Att}_m(X)]^\top \quad (5)$$

$$\text{Att}_i(X) = \text{softmax}\frac{((W_{Q_i}X)^\top W_{K_i}X)}{\sqrt{D/m}}(W_{V_i}X)^\top \quad (6)$$

for input layer $X \in \mathbb{R}^{D \times N}$, and learnable parameters $W_{Q_i}, W_{K_i}, W_{V_i} \in \mathbb{R}^{\frac{D}{m} \times D}$, $W_{att} \in \mathbb{R}^{D \times D}$ for m attention heads.

For language representation, we employ an off-the-shelf BERT model. An input sequence $s_0 = [w_0, ..., w_S]$ is given with special tokens $[CLS]$ and $[SEP]$ inserted at the beginning and the end of the sequence respectively. In case of two sentences within the input sequences, $[SEP]$ token is also inserted in between the two. Each token is represented as the sum of word embedding, position embedding, and segment embedding, and undergoes bidirectional multi-head self-attention over multiple layers. The representation for the input sequence is obtained as $h_0, ..., h_S$ from the upper-

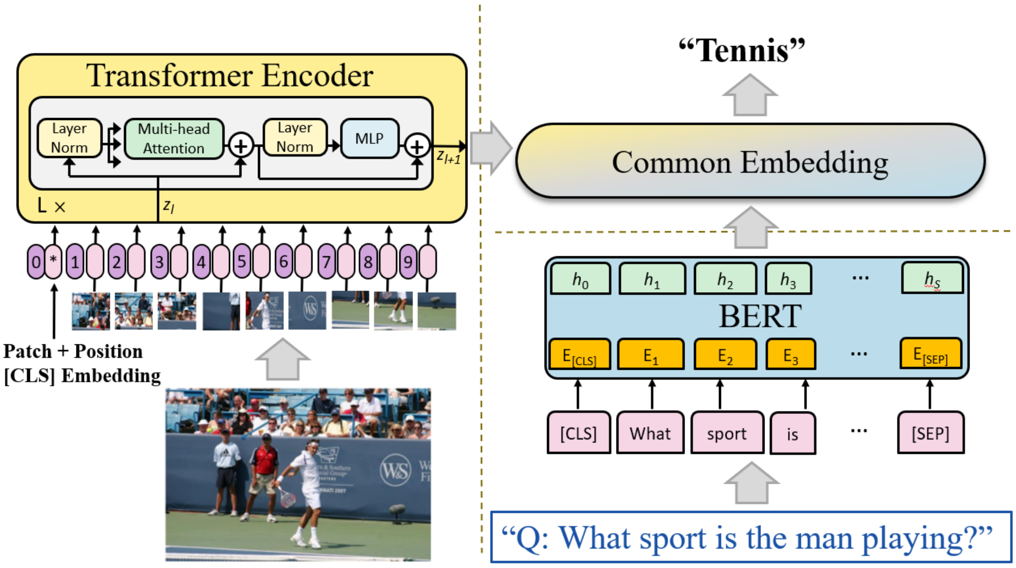


Figure 1: Overview of our model. Images are split into patches, sentences are split into tokens, and both are encoded with transformers, after which they are combined in common embedding space for final classification.

most attention layer:

$$s_0 = [[CLS], w_1, ..., w_{S-1}, [SEP]] \quad (7)$$

$$s'_l = \text{MSA}(\text{LN}(s_{l-1})) + s_{l-1}, l = 1...L \quad (8)$$

$$s_l = \text{MLP}(\text{LN}(s'_l)) + s'_l, l = 1...L \quad (9)$$

$$y_{lang} = \text{LN}(s^0_L) \quad (10)$$

We now project the image and language representations obtained into common embedding space by concatenation:

$$y = \text{Concat}(y_{img}, y_{lang}) \quad (11)$$

Note that we deliberately choose the most elementary common embedding scheme, as our focus is to examine the performance of the features themselves, rather than the embedding scheme. It is thus highly likely that, when coupled with more sophisticated embedding schemes, a significant performance boost will occur. Fig. 1 describes the overall architecture of our approach.

4 Experiments

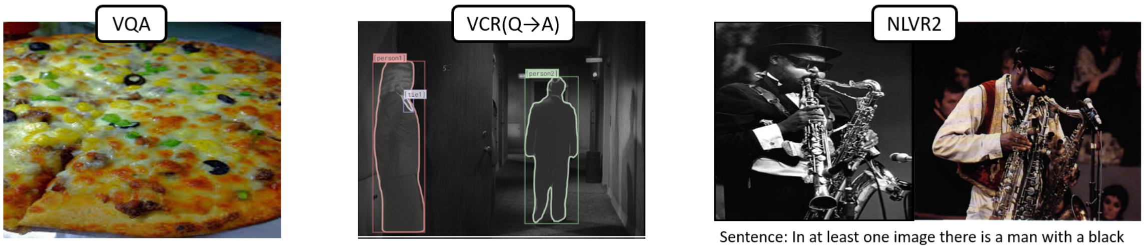
4.1 Setting

For images, we use ViT-B model pre-trained on ImageNet-21k. The model contains 12-layers with 12 attention heads and hidden size of 768, consisting of 86M parameters. For language, we use off-the-shelf BERT$_{\text{BASE}}$ mode, trained with BERT's pre-training objectives of masked language modeling and next sentence prediction on BookCorpus (Zhu et al., 2015) and English Wikipedia. Like ViT-B, the model contains 12-layers with 12 attention heads and hidden size of 768, and consists of 110M parameters. During both training and testing, image and language features are extracted from the uppermost layer of respective model, and we concatenate them to make a 1536-dimensional vector. Concatenated features are trained with cross-entropy loss and Adam (Kingma and Ba, 2014) optimizer.

We evaluate our model on the following commonly tackled cross-modal visuolinguistic tasks; visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017), visual commonsense reasoning (VCR) (Zellers et al., 2019), and reasoning about natural language grounded in photographs (NLVR2) (Suhr et al., 2019). For VCR, we followed (Lu et al., 2019) by making 4 possible pairs of question and answer. For NLVR2, we follow the *pair* approach of (Chen et al., 2020), by embedding each image and the query, as it is reported to outperform *triplet* approach of embedding two images with the query. We trained with 4 V100 GPUs with batch size 96 for VQA and NLVR2, and 48 for VCR, which were adjusted with respect to the memory constraint of the computational environment. Learning rate was initially set to 1e-4 under linearly decaying schedule with warm up. We trained the model for 25 epochs for each task.

4.2 Results

Table 1 compares our model's performance with other transformer-based cross-modal models. While our model's performance falls below that of state-of-the-art models, it is noteworthy that other models explicitly perform additional cross-modal pre-training on top of already pre-trained vision and



Q: What sort of food is this?
Answer: Pizza

Q: Why is person1 standing behind the open door?
Answer: person1 is trying to hide from person2

Sentence: In at least one image there is a man with a black hat and brown vest playing three saxophones at once.
Answer: True

Figure 2: Qualitative results for each task.

Method	Pre-train #samples	VQA test-dev	VCR QA	VCR QA-R	VCR Q-AR	NLVR2 dev
ViLBERT	3.3M	70.55	73.3	74.6	54.8	–
UNITER	9.5M	73.82	77.3	80.8	62.8	78.4
MiniVLM	7M	69.39	–	–	–	73.7
VisualBERT	0.5M	70.80	71.6	73.2	52.4	67.4
DeVLBERT	3.3M	71.1	–	–	–	–
CAPT	9.2M	72.78	–	–	–	75.1
ERNIE-ViL	4M	73.78	79.2	83.5	66.3	–
Ours	0	67.84	68.4	70.2	49.2	65.2

Table 1: Comparison of our model to other state-of-the-art cross-modal models. 2nd column refers to the number of image-caption pairs seen during cross-modal pre-training. Despite the disadvantage of not having seen a substantial amount of pre-training data, our model closely follows the state-of-the-art models. (CAPT (Luo et al., 2020), ERNIE-ViL (Yu et al., 2020)

language modules. On the other hand, our model simply concatenates two pre-trained models without additional cross-modal pre-training, and is immediately trained with the target task. For example, ViLBERT and DeVLBERT (Zhang et al., 2020) are pre-traiend with Conceptual Captions dataset (Sharma et al., 2018), and our model is at the disadvantage of not having seen 3.3M pairs of image and captions, yet comes fairly close to those pre-trained models. Fig. 2 shows qualitative examples of the model's performance on each task. In addition, while many papers on pre-trained cross-modal representations do not report specific number of parameters, our approximations of other models' sizes based on the implementation details reported in respective papers suggest that our model is reasonably smaller, especially since it completely eliminates the need for external region detector.

4.3 Further Experiments

In order to examine how much each component contributes to performance, we conduct further experiments, replacing each component with conventional modules. We first replace transformer encoder for images with CNN module, specifically with ResNet-50 (He et al., 2016) trained on ImageNet, using global average-pooled features. We also examine replacing BERT module with LSTM (Hochreiter and Schmidhuber, 1997) using early fusion with image features. For fair comparison, we used concatenation as common embedding scheme for all combinations.

Method	VQA test-dev	VCR QA	VCR QA-R	VCR Q-AR	NLVR2 dev
CNN+BERT	65.36	65.4	68.1	48.9	61.9
ViT+LSTM	62.27	62.9	66.4	45.8	60.1
w/o finetuning	56.54	58.2	60.3	42.5	52.4
Ours	67.84	68.4	70.2	49.2	65.2

Table 2: Comparison of our model to different combinations. Under the same condition of no explicit cross-modal fine-tuning and the same embedding scheme, our model outperforms other combinations.

Table 2 shows the results. While ResNet/BERT comes fairly close, it falls below our model, and performance drop is clearer with ViT/LSTM, possibly reflecting superior adaptability of BERT compared to LSTM. Our conjecture is that architectural integrity, *i.e.*, using the same architecture for both vision and language, throughout the model, plays an important role in learning cross-modal representations. Note that, however, it would require a more thorough and analytical study to conclusively claim that ViT is superior to ResNet, or that attention is superior to convolution, and our primary purpose in this experiment is simply to demonstrate that transformer-exclusive models can accomplish comparable performance to the models employing CNN. We also examined linearly training a classifier for target task while fixing the extracted features, without fine-tuning. As expected, there is a significant performance drop, reaffirming the premise that the competence of transformer and BERT is attainable via fine-tuning to its downstream tasks.

Note that, although we performed experiments on a small set of cross-modal tasks, given the superior performance of ViT over ResNet on image classification as reported by (Dosovitskiy et al., 2020), and also on other computer vision tasks as reported by models like pyramid vision transformer (Wang et al., 2021), we believe any task that involves vision and language is a potential beneficiary of transformer-exclusive approach, since it enables the architectural integrity for both modalities.

5 Conclusion

This paper proposed to handle cross-modal tasks for vision and language, solely based on transformer architecture, examining it in various cross-modal tasks. Our paper admittedly does not claim state-of-the-art performances, but to the best of our knowledge, our work is one of the first attempts, along with models like ViLT (Kim et al., 2021) and UniT (Hu and Singh, 2021), to examine cross-modal representation for vision and language solely based on transformer architecture, excluding CNNs and RNNs. Without any structural modifications or sophisticated common embedding scheme, and without additional cross-modal pre-training with millions of samples, our model demonstrates comparable performances to state-of-the-art cross-modal models. Since we deliberately chose the smallest baseline models for each component, and a very simple concatenation scheme, we can intuitively expect an enhanced performance by selecting larger pre-trained models at the cost of more parameters, or by selecting more sophisticated common embedding scheme. The same holds true for the amount of pre-training data used, as we can reasonably expect the performance to boost by using the same amount of pre-training data employed by previous models. With transformer's relative computational efficiency as reported by (Dosovitskiy et al., 2020), the architectural integrity proposed in our model is likely to lead to new research direction, and we hope to encourage more advanced models with novel ideas to follow in near future.

Acknowledgments

We thank Naofumi Akimoto, Akio Hayakawa, Masato Ishii, and Krishna Wadhwani for their help on the manuscript.

References

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)*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell Waggoner, Song Wang, Jinlian Wei, Yifan Yin, and Zhiqi Zhang. 2012. Video in sentences out.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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 *ECCV*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

L. Dong, S. Xu, and B. Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, USA. Association for Computational Linguistics.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision, ECCV 2010 - 11th European Conference on Computer Vision, Proceedings*, number PART 4 in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 15–29. Springer-Verlag Berlin Heidelberg. Copyright: Copyright 2019 Elsevier B.V., All rights reserved.; 11th European Conference on Computer Vision, ECCV 2010 ; Conference date: 10-09-2010 Through 11-09-2010.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Dan Hendrycks and Kevin Gimpel. 2020. Gaussian error linear units (gelus).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pretraining. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11336–11344.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, pages 13–23. Curran Associates, Inc.

Fuli Luo, Pengcheng Yang, Shicheng Li, Xuancheng Ren, and Xu Sun. 2020. Capt: Contrastive pretraining for learning denoised sequence representations.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24, pages 1143–1151. Curran Associates, Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 91–99. Curran Associates, Inc.

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

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020a. Minivlm: A smaller and faster vision-language model.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.

Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, and et al. 2020b. Transformer-based acoustic modeling for hybrid speech recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

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

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. Devlbert. *Proceedings of the 28th ACM International Conference on Multimedia*.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pages 19–27. IEEE Computer Society.