VL-BEIT: Generative Vision-Language Pretraining

Hangbo Bao,* **Wenhui Wang**,* **Li Dong**, **Furu Wei**†
Microsoft Research
https://github.com/microsoft/unilm

Abstract

We introduce a vision-language foundation model called VL-BEIT, which is a bidirectional multimodal Transformer learned by generative pretraining. Our minimalist solution conducts masked prediction on both monomodal and multimodal data with a shared Transformer. Specifically, we perform masked vision-language modeling on image-text pairs, masked language modeling on texts, and masked image modeling on images. VL-BEIT is learned from scratch with one unified pretraining task, one shared backbone, and one-stage training. Our method is conceptually simple and empirically effective. Experimental results show that VL-BEIT obtains strong results on various vision-language benchmarks, such as visual question answering, visual reasoning, and image-text retrieval. Moreover, our method learns transferable visual features, achieving competitive performance on image classification, and semantic segmentation.

1 Introduction

Generative pretraining has achieved great success in natural language processing (Radford et al., 2018; Devlin et al., 2019; Dong et al., 2019; Liu et al., 2019; Conneau et al., 2020; Chi et al., 2021) and computer vision (Bao et al., 2022; He et al., 2021). Specifically, BERT (Devlin et al., 2019) introduces masked language modeling, which learns to recover masked tokens based on the bidirectional contextualized representations encoded by Transformer (Vaswani et al., 2017). BEIT (Bao et al., 2022) introduces masked image modeling to pretrain vision Transformer (Dosovitskiy et al., 2020), which randomly masks image patches and predicts the corresponding visual tokens.

In this work, we explore the mask-then-predict paradigm for multimodal (i.e., vision-language) pretraining. Our model, namely VL-BEIT, is simple and effective, which is trained from scratch with one unified masked prediction task, one shared Transformer, and one-stage training. We perform masked prediction on both monomodal (i.e., unpaired images and text) and multimodal data (image-text pairs). Specifically, the unified objective contains masked language modeling and masked image modeling to learn monomodal representations from large-scale monomodal data, and masked vision-language modeling to aggregate and align visual and linguistic information from multimodal data. After pretraining, our model can be finetuned on various vision-language and vision tasks. In addition, we employ mixture-of-modality-experts (MOME) Transformer (Wang et al., 2021a) as the shared backbone network. Each block of MOME Transformer consists of a shared self-attention module across different modalities to align the contents, and a pool of modality experts to capture modality-specific information. Benefiting from the multimodal pretraining objective and the shared Transformer backbone, VL-BEIT can be used as a image encoder for downstream vision tasks, or finetuned as a dual encoder or fusion encoder for vision-language tasks.

We conduct extensive experiments on vision-language benchmarks including visual question answering, visual reasoning, and image-text retrieval. Experimental results demonstrate that our model obtains competitive performance across vision-language benchmarks. We also evaluate our model on vision tasks including image classification and semantic segmentation, achieving strong results.

* Equal contribution. † Corresponding author.

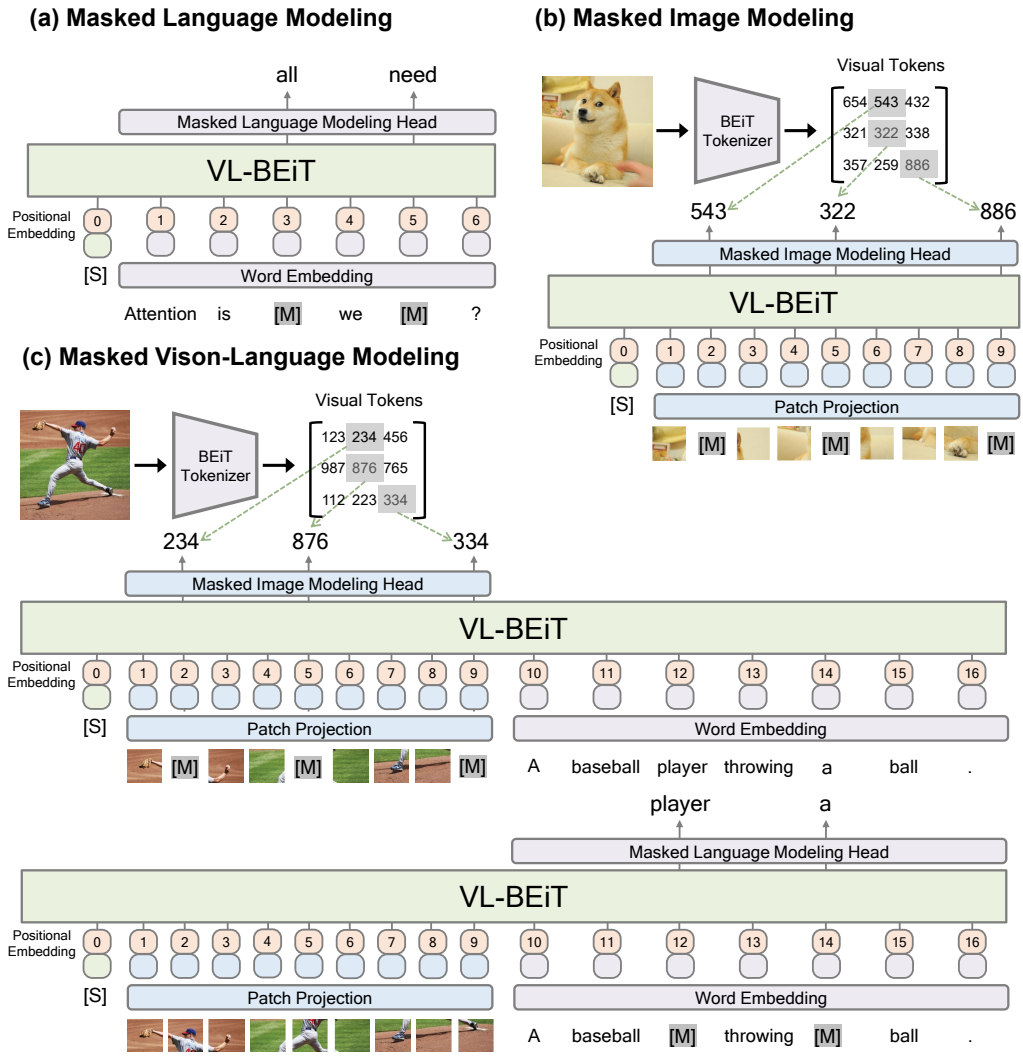

Figure 1: VL-BEIT is pretrained by masked prediction on both monomodal and multimodal data with a shared Transformer.

Ablation studies show that the pretraining tasks and MOME Transformer positively contribute to the final performance.

Our contributions are summarized as follows:

- We introduce a vision-language foundation model named VL-BEIT, which is pretrained by the mask-then-predict task on both multimodal and monomodal data.
- We propose a simple and effective framework that uses one unified generative pretraining task, one shared bidirectional Transformer, and one-stage training from scratch.
- Experimental results across various downstream tasks show that our method learns transferable vision-language and visual features.

2 Methods

As illustrated in Figure 1, VL-BEIT is pretrained by the mask-then-predict task with a shared multimodal Transformer. We perform masked image modeling on monomodal image data, masked language modeling on monomodal text data, and masked vision-language modeling on multimodal

image-text pairs. After pretraining, the modal can be finetuned as an image encoder, dual encoder, or fusion encoder for various vision and vision-language downstream tasks.

2.1 Input Representations

Image Representations Following (Dosovitskiy et al., 2020), we split the image $\boldsymbol{v} \in \mathbb{R}^{H \times W \times C}$ into a sequence of patches, so that the image can be encoded by standard Transformer. The number of patches is $N = HW/P^2$, where C is the number of channels, (H, W) is the image resolution, and (P, P) is the patch resolution. We then flatten these image patches and obtain patch embeddings $(\{\boldsymbol{v}_i^p\}_{i=1}^N)$ via a linear projection layer. A learnable special token `[I_CLS]` is prepended to the sequence of patch embeddings. Finally, we sum image patch embeddings and learnable position embeddings to obtain the final representations $\boldsymbol{H}^v = [\boldsymbol{v}_{\texttt{[I\_CLS]}}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N] + \boldsymbol{V}_{pos}$.

Text Representations We tokenize the input text and project the tokens to word embeddings $(\{\boldsymbol{w}_i\}_{i=1}^M)$, where M is the length of tokenized text sequence. Two special tokens, including a start-of-sequence token (`[T_CLS]`) and a special boundary token (`[T_SEP]`), are added to the sequence. Finally, text representations are obtained via summing the word embeddings and text position embeddings $\boldsymbol{H}^w = [\boldsymbol{w}_{\texttt{[T\_CLS]}}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_M, \boldsymbol{w}_{\texttt{[T\_SEP]}}] + \boldsymbol{T}_{pos}$.

Image-Text Pair Representations Given an image-text pair, we first obtain the image and text input representations as above, respectively. Then we concatenate these vectors to get the image-text pair representations $\boldsymbol{H}^{vl} = [\boldsymbol{H}^w; \boldsymbol{H}^v]$.

2.2 Backbone Network

We use a shared multimodal Transformer as the backbone network. Given the image and text representations of monomodal data, and the representations of image-text pairs, we employ a mixture-of-modality-experts (MOME) Transformer (Wang et al., 2021a) to encode different modalities. Specifically, MOME Transformer stacks multiple layers of blocks. In each block, MOME Transformer contains a multi-head self-attention layer and a feed-forward expert layer. The self-attention module is shared across different modalities. In contrast, each feed-forward expert layer has a pool of modality-specific experts, which performs as a substitute of the feed-forward network in standard Transformers. In other words, we use the modality of input token to conduct hard routing over the pool of feed-forward networks.

MOME Transformer is flexible to support various downstream tasks by activating different modality-specific experts. For example, we can use the backbone as monomodal Transformers (i.e., vision or language encoder), multimodal encoders (i.e, with deep fusion), and crossmodal Transformers (i.e., dual encoders).

2.3 Pretraining Tasks

VL-BEIT is jointly optimized by masked image modeling on images, masked language modeling on texts, and masked vision-language modeling on image-text pairs.

Masked Language Modeling VL-BEIT uses masked language modeling (MLM) to learn language representations from large-scale text-only data. Following BERT (Devlin et al., 2019), we randomly mask 15% tokens of monomodal text data. Each masked token is replaced by a `[MASK]` token 80% of the time, a random token 10% of the time and kept the original tokens 10% of the time. The pretraining objective is to recover the masked tokens from the corrupted input text.

Masked Image Modeling In addition to masked language modeling, we employ masked image modeling (MIM) to learn vision representations from large-scale image data. Following BEIT (Bao et al., 2022), we apply block-wise masking strategy to mask 40% of image patches. The pretraining objective of MIM is to reconstruct the discrete visual tokens of masked patches. We use image tokenizer of BEITV2 (Peng et al., 2022) to obtain the discrete tokens as the reconstructed targets.

Masked Vision-Language Modeling We introduce masked vision-language modeling (MVLM), which extends masked language modeling and masked image modeling to multimodal data. The

task aims at recovering masked image patches and text tokens based on visual and linguistic clues. Specifically, we randomly mask text tokens (with 50% mask ratio) as in MLM, and recover the masked text tokens based on the joint image-text representations. In addition, we mask image patches as in MIM and predict its corresponding visual tokens based on the image-text pair. The masking strategy is the same as in MIM. The MVLM task encourages the model to learn alignments between the pairs of image and text.

3 Experiments

We evaluate the pretrained model on vision-language and visual tasks. We also present ablation studies of pretraining tasks and the backbone architecture.

3.1 Pretraining Setup

Our pretraining data consists of monomodal and multimodal data. For monomodal data, we use ImageNet-22K as the image data, English Wikipedia and BookCorpus (Zhu et al., 2015) as the text data. The multimodal data combines four datasets of image-text pairs: Conceptual Captions (Sharma et al., 2018), SBU Captions (Ordonez et al., 2011), COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017). The multimodal data has about 4M images and 10M image-text pairs.

Following previous work (Dosovitskiy et al., 2020; Bao et al., 2022; Wang et al., 2021a), we adopt the same base-size network architecture which consists of 12-layer Transformer blocks with 768 hidden size and 12 attention heads. We follow the parameter initialization method used in BEiT (Bao et al., 2022). The image resolution used for pretraining is 224×224, and the image patch size is 16×16. We mix the data and pretrain the model from scratch with a total batch size of $6,144$ for 480k steps (i.e., 100 epochs of the image-text pairs). Each batch contains $2,048$ images, $2,048$ text and $2,048$ image-text pairs. For the ablation experiments, we train the model for 40 epochs. Following BEiT, we use random resized cropping, horizontal flipping, and color jittering (Wu et al., 2018) to perform image augmentation. We use a SentencePiece tokenizer (Kudo and Richardson, 2018) with 64k vocab size to tokenize the text data. Adam (Kingma and Ba, 2015) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ is utilized to optimize the model. The peak learning rate is 2e-3, with linear warmup over the first $10,000$ steps and cosine learning rate decay. The weight decay is 0.05. We disable dropout, and use stochastic depth (Huang et al., 2016) with a rate of 0.1.

3.2 Vision-Language Downstream Tasks

We conduct vision-language finetuning experiments on the widely used visual question answering (Goyal et al., 2017), natural language for visual reasoning (Suhr et al., 2019) and image-text retrieval (Plummer et al., 2015; Lin et al., 2014) tasks. We use 480×480 image resolution for VQA fine-tuning, and 384×384 for other tasks.

Visual Question Answering (VQA) VQA aims to answer questions based on the given image. Following previous work (Kim et al., 2021; Wang et al., 2021a), we use VQA 2.0 dataset (Goyal et al., 2017), and formulate the task as a classification problem to choose the answer from $3,129$ most frequent answers. We finetune our model as a fusion encoder to jointly encode the image and question. The final encoding vector of the `[T_CLS]` token is used as the representation of the image-question pair, and then fed into a classifier layer to predict the label.

Natural Language for Visual Reasoning (NLVR2) For visual reasoning task, a text description and a pair of images are given, the task is to predict whether the description is true about the visual input. We use NLVR2 (Suhr et al., 2019) dataset to evaluate the model. Following OSCAR (Li et al., 2020) and VinVL (Zhang et al., 2021), we create two image-text pairs based on the triplet input. Our model is used as a fusion encoder to jointly encode the image and text. The final vectors of `[T_CLS]` token of the two pairs are concatenated to predict the label.

Image-Text Retrieval Depending on the target modality, the task can be divided into two subtasks: image-to-text retrieval and text-to-image retrieval. We use the widely used COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) datasets to evaluate the model, and adopt the Karpathy split (Karpathy and Fei-Fei, 2015) following common practices. We employ image-text contrast and

Model	VQA		NLVR2	
	test-dev	test-std	dev	test-P
Base-size models pretrained on the same data				
UNITER	72.70	72.91	77.18	77.85
VILLA	73.59	73.67	78.39	79.30
UNIMO	73.79	74.02	-	-
ViLT	71.26	-	75.70	76.13
ALBEF	74.54	74.70	80.24	80.50
VLMo	76.64	76.89	**82.77**	**83.34**
VL-BEiT	**77.53**	**77.75**	81.93	82.66

Table 1: Finetuning results of base-size models on vision-language classification tasks. We report vqa-score on VQA test-dev and test-standard split, accuracy for NLVR2 development and public test set (test-P).

Model	COCO		Flickr30K	
	TR	IR	TR	IR
Fusion encoder				
UNITER	64.4	50.3	85.9	72.5
VILLA	-	-	86.6	74.7
ViLT	61.5	42.7	83.5	64.4
Dual encoder				
VLMo	74.8	57.2	92.3	79.3
Dual encoder + Fusion encoder reranking				
ALBEF	73.1	56.8	94.3	82.8
VL-BEiT	**79.5**	**61.5**	**95.8**	**83.9**

Table 2: Finetuning results of base-size models on image-text retrieval tasks. We report top-1 recall for image retrieval (IR) and text retrieval (TR).

image-text matching with hard negative mining objectives as in VLMo (Wang et al., 2021a) to jointly finetune the model. During inference, we first use our model as a dual encoder to obtain the top-k candidates, then the model is used as a fusion encoder to rank the candidates based on its image-text matching scores.

Table 1 reports the results on vision-language classification tasks, including VQA and NLVR2. We compare VL-BEiT with other base-size models pretrained on the same image-text pair data. VL-BEiT outperforms previous base-size models on VQA and achieves competitive performance on NLVR2. The unified mask-then-predict pretraining task effectively learns multimodal representations.

Our model also achieves promising performance on image-text retrieval tasks. As shown in Table 2, we compare with fusion-encoder models, dual-encoder models and the reranking models. Fusion-encoder models jointly encode all image-text combinations and obtain the similarity scores via the image-text matching objective. Dual-encoder models encode images and text separately, and compute the similarity scores via a simple interaction layer (i.e., dot product). The reranking models first obtain topk-k candidates from the dual encoder, and then rank the candidates via the image-text matching scores computed by the fusion encoder. VL-BEiT outperforms ALBEF, which is also a reranking model, without using image-text contrast/matching during pretraining.

3.3 Vision Downstream Tasks

Image Classification The task aims to classify the input image to the corresponding category. We use the ILSVRC-2012 ImageNet dataset (Russakovsky et al., 2015), which consists of 1.3M images with 1k classes. Following BEiT (Bao et al., 2022), we perform average pooling over the final vectors, and then feed the resulted vector into a linear classfier layer to predict the label.

Semantic Segmentation The task is to predict the label for each pixel of the input image. We evaluate our model on the ADE20K dataset (Zhou et al., 2019). The dataset contains 25K images with 150 semantic categories. We use the same task layer as in UperNet (Xiao et al., 2018).

As shown Table 3, we compare VL-BEiT with two base-size vision Transformers on image classification and semantic segmentation. For BEiT and VL-BEiT, we perform intermediate finetuning on ImageNet-22k to compare with ViT pretrained on ImageNet-22k. VL-BEiT outperforms previous state-of-the-art supervised and self-supervised models on ImageNet-1k. The model also performs competitively on ADE20k.

3.4 Ablation Studies

We conduct ablation studies to analyze the contributions of pretraining tasks and MoME Transformer used in VL-BEiT. We evaluate the models on visual reasoning (NLVR2) and image-text retrieval (Flickr30k).

Models	ImageNet (acc@1)	ADE20K (mIoU)
Vision Pretraining		
ViT (Dosovitskiy et al., 2020)	83.6	-
BEiT (Bao et al., 2022)	85.2	52.8
Vision-Language Pretraining		
VL-BEiT	**85.9**	**53.1**

Table 3: Results of base-size models on image classification (ImageNet-1K) and semantic segmentation (ADE20K). We report top-1 accuracy for ImageNet, and mean Intersection of Union (mIoU) averaged over all semantic categories for ADE20k.

	Pretraining Tasks			NLVR2		Flickr30k	
	MVLM	MIM	MLM	dev	test-P	TR R@1	IR R@1
[1]	✓	✗	✗	79.15	80.78	91.2	75.8
[2]	✓	✓	✗	80.44	81.36	**92.2**	77.4
[3]	✓	✓	✓	**81.10**	**82.19**	**92.2**	**77.9**

Table 4: Ablation studies of pretraining tasks.

Pretraining Task Table 4 presents the results using different pretraining tasks. Masked image modeling and masked language modeling on monomodal data positively contribute to our method. In addition, we find that only performing MLM and MIM training on monomodal data gives a relatively low accuracy on NLVR2. Masked vision-language modeling plays a critical role in our method.

Backbone Architecture We also compare MoME Transformer used in our model with standard Transformer. The results are shown in Table 5. Using MoME performs better than standard Transformer on both visual reasoning and image-text retrieval. Modality experts used in MoME effectively capture modality-specific information and improve the model.

4 Related Work

Vision-language pretraining (Tan and Bansal, 2019; Lu et al., 2019; Su et al., 2020; Zhang et al., 2021; Radford et al., 2021; Li et al., 2020; Kim et al., 2021; Li et al., 2021; Wang et al., 2021b;a; 2022b; Alayrac et al., 2022; Yu et al., 2022) aims to learn multimodal representations from large-scale image-text pairs. Model architecture and pretraining objectives are critical to the effectiveness of vision-language models.

Model Architectures There are two mainstream architectures widely used in previous models: *dual-encoder* and *fusion-encoder* models. Dual-encoder model (Radford et al., 2021; Jia et al., 2021) consists of an image encoder and a text encoder. It encodes images and text separately, and then employs cosine similarity to model the interaction of image and text vectors. Dual-encoder models achieve promising results for image-text retrieval tasks with linear time complexity. However, the simple fusion module is not enough to handle complex vision-language understanding tasks such as visual reasoning. Fusion-encoder models employ a complex fusion module with cross-modal attention, to jointly encode images and text. Previous models (Lu et al., 2019; Su et al., 2020; Li et al., 2020; Zhang et al., 2021) use an off-the-shelf object detector like Faster R-CNN (Ren et al., 2017) to obtain image region features. Text features are usually word embeddings or contextual vectors encoded by a text encoder. These image and text features are then jointly encoded by the fusion module, which usually adopts a multi-layer Transformer network. Recently, Pixel-BERT (Huang et al., 2020) and ALBEF (Li et al., 2021) use CNN/vision Transformer to encode images and remove object detector. ViLT (Kim et al., 2021) uses a shared Transformer network to jointly encode image patches and word embeddings. Fusion-encoder models achieve superior performance on vision-language understanding tasks such as vison reasoning. But it requires quadratic time complexity for retrieval tasks, which leads to a much slower inference speed than the dual-encoder models. VLMo (Wang et al., 2021a) unifies dual-encoder and fusion-encoder models and introduces mixture-of-modality-experts (MoME) Transformer to encode various modalities within a shared Transformer

Architecture	NLVR2		Flickr30k	
	dev	test-P	TR R@1	IR R@1
Standard Transformer	80.77	81.42	91.7	75.8
MoME Transformer	**81.10**	**82.19**	**92.2**	**77.9**

Table 5: Ablation study of MoME Transformer.

block. In this work, we adopt the MoME Transformer as the backbone network given its simplicity and flexibility. VL-BEiT can also be finetuned as a dual-encoder model or fusion-encoder model.

Pretraining Objectives Multiple cross-modal pretraining objectives have been proposed, including image-text contrastive learning (Radford et al., 2021; Jia et al., 2021), image-text matching (Tan and Bansal, 2019; Kim et al., 2021; Li et al., 2021; Wang et al., 2021a), masked language modeling (Tan and Bansal, 2019; Su et al., 2020; Kim et al., 2021) or prefix language modeling (Wang et al., 2021b), masked region classification (Tan and Bansal, 2019), word-patch/region alignment (Chen et al., 2020; Kim et al., 2021). SimVLM (Wang et al., 2021b) proposes to train the vision-language model using prefix language modeling on image-text pairs and text-only data. FLAVA (Singh et al., 2021) combines masked image modeling with masked language modeling, image-text contrast and matching based on a fusion-encoder model. Masked image modeling and masked language modeling are applied on the monomodal encoders. Masked multimodal modeling, image-text contrast and matching losses are used for the multimodal encoder. Compared with SimVLM, VL-BEiT introduces richer visual supervision via masked image modeling and masked vision-language modeling. Different form FLAVA, we use a shared MoME Transformer network for different modalities and adopt one-stage training from scratch.

5 Conclusion

In this work, we introduce VL-BEiT, a simple and effective approach to pretraining a bidirectional multimodal Transformer encoder for both vision-language and vision tasks. It solely employs generative pretraining tasks, including masked language modeling on texts, masked image modeling on images, and masked vision-language modeling on image-text pairs. We show that VL-BEiT effectively leverages monomodal data like images and texts as well as multimodal data like image-text pairs. Experimental results show that VL-BEiT gets strong performance on both vision-language and vision tasks.

In the future, we would like to improve VL-BEiT from the following perspectives:

- We will scale up the model size (Wang et al., 2022a; Chi et al., 2022) and data for VL-BEiT training. We would like to explore whether the success of scaling up generative pretraining in NLP can be reproduced for multimodal pretraining under the VL-BEiT framework.
- Following the research from multilingual language model pretraining (Chi et al., 2021), we will integrate contrastive objectives like CLIP (Radford et al., 2021) into VL-BEiT, either in pretraining stage by joint learning of generative and contrastive objectives or as an intermediate finetuning task.
- We are also interested in the zero-shot cross-modality transferability (Song et al., 2022) across different modalities like vision and language.

Acknowledgement

We would like to acknowledge Zhiliang Peng for the helpful discussions.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian

Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198, 2022. doi: 10.48550/arXiv.2204.14198. URL https://doi.org/10.48550/arXiv.2204.14198.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. doi: 10.1007/978-3-030-58577-8_7. URL https://doi.org/10.1007/978-3-030-58577-8_7.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.naacl-main.280.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. On the representation collapse of sparse mixture of experts. *ArXiv*, abs/2204.09179, 2022.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *preprint arXiv:2010.11929*, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. URL https://arxiv.org/abs/2111.06377.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 646–661. Springer, 2016.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849, 2020. URL https://arxiv.org/abs/2004.00849.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society, 2015.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21k.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021. URL https://arxiv.org/abs/2107.07651.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. doi: 10.1007/978-3-030-58577-8_8. URL https://doi.org/10.1007/978-3-030-58577-8_8.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

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 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *ArXiv*, abs/2208.06366, 2022.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303. URL https://doi.org/10.1109/ICCV.2015.303.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. URL https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/language-unsupervised/languageunderstandingpaper.pdf.

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. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39 (6):1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031. URL https://doi.org/10.1109/TPAMI.2016.2577031.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018. URL https://aclanthology.org/P18-1238/.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. *CoRR*, abs/2112.04482, 2021.

Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 6088–6100, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.421.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SygXPaEYvH.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1644. URL https://doi.org/10.18653/v1/p19-1644.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1514. URL https://doi.org/10.18653/v1/D19-1514.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. DeepNet: Scaling Transformers to 1,000 layers. *ArXiv*, abs/2203.00555, 2022a.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022b. URL https://arxiv.org/abs/2202.03052.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. *CoRR*, abs/2111.02358, 2021a. URL https://arxiv.org/abs/2111.02358.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021b. URL https://arxiv.org/abs/2108.10904.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. Computer Vision Foundation / IEEE Computer Society, 2018.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 432–448. Springer, 2018. doi: 10.1007/978-3-030-01228-1_26. URL https://doi.org/10.1007/978-3-030-01228-1_26.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *CoRR*, abs/2205.01917, 2022. doi: 10.48550/arXiv.2205.01917. URL https://doi.org/10.48550/arXiv.2205.01917.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.html.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019. doi: 10.1007/s11263-018-1140-0. URL https://doi.org/10.1007/s11263-018-1140-0.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.