


Scaling Up Vision-Language Pre-training for Image Captioning

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang,
Zicheng Liu, Yumao Lu, Lijuan Wang
Microsoft

{xiaowei.hu, zhe.gan, jianfw, zhengyang, zliu, yumaolu, lijuanw}@microsoft.com

Abstract

In recent years, we have witnessed significant performance boost in the image captioning task based on vision-language pre-training (VLP). Scale is believed to be an important factor for this advance. However, most existing work only focuses on pre-training transformers with moderate sizes (e.g., 12 or 24 layers) on roughly 4 million images. In this paper, we present LEMON , a Large-scale iMage captiONer, and provide the first empirical study on the scaling behavior of VLP for image captioning. We use the state-of-the-art VinVL model as our reference model, which consists of an image feature extractor and a transformer model, and scale the transformer both up and down, with model sizes ranging from 13 to 675 million parameters. In terms of data, we conduct experiments with up to 200 million image-text pairs which are automatically collected from web based on the alt attribute of the image (dubbed as ALT200M¹). Extensive analysis helps to characterize the performance trend as the model size and the pre-training data size increase. We also compare different training recipes, especially for training on large-scale noisy data. As a result, LEMON achieves new state of the arts on several major image captioning benchmarks, including COCO Caption, nocaps, and Conceptual Captions. We also show LEMON can generate captions with long-tail visual concepts when used in a zero-shot manner.

1. Introduction

Recent advances in image captioning [1, 5, 35] can be largely attributed to vision-language pre-training (VLP) [26, 30, 37, 40], the current prevailing training paradigm for vision-language (VL) research. VLP [6] is usually conducted on a combined image-text dataset comprising of several or tens of millions images in total, e.g., Visual Genome [20], SBU [32] and Conceptual Captions [4, 35]. While previous studies [29, 48, 49] have ana-

¹The dataset is released at <https://github.com/xiaoweihu/ALT200M>.

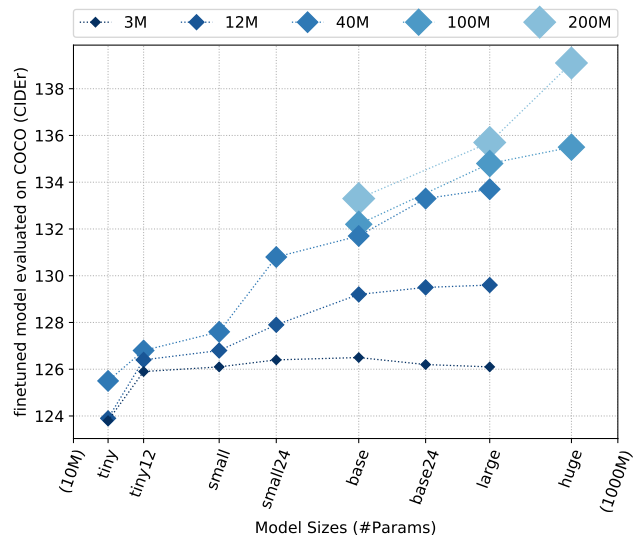


Figure 1. **Image captioning performance on COCO when up-scaling model for each dataset size.** The x-axis plots the number of parameters for each model size (e.g., tiny, small, huge) in a logarithmic scale. The definition of model sizes is detailed in Table 2. Increasing the model size is not significantly beneficial at small pre-training dataset scales. However, when we use sufficiently large datasets, we see strong performance boost from a larger model.

lyzed various choices of pre-training objectives and model architectures, it remains unclear to what extent the pre-training dataset would impact the performance, and how it correlates with different model settings. Along the journey of pushing the limit of VLP, it becomes increasingly important to answer this question.

Scale is believed to be an important ingredient in attaining excellent performance [17, 33, 43]. Recent work has investigated the Pareto frontier of training transformer models, often referred to as the neural scaling law, in the domains of natural language processing [2, 18, 41] and computer vision [12, 47], via unsupervised or weakly-supervised learning methods. These studies have observed consistent benefits of increasing the model size to billions of param-

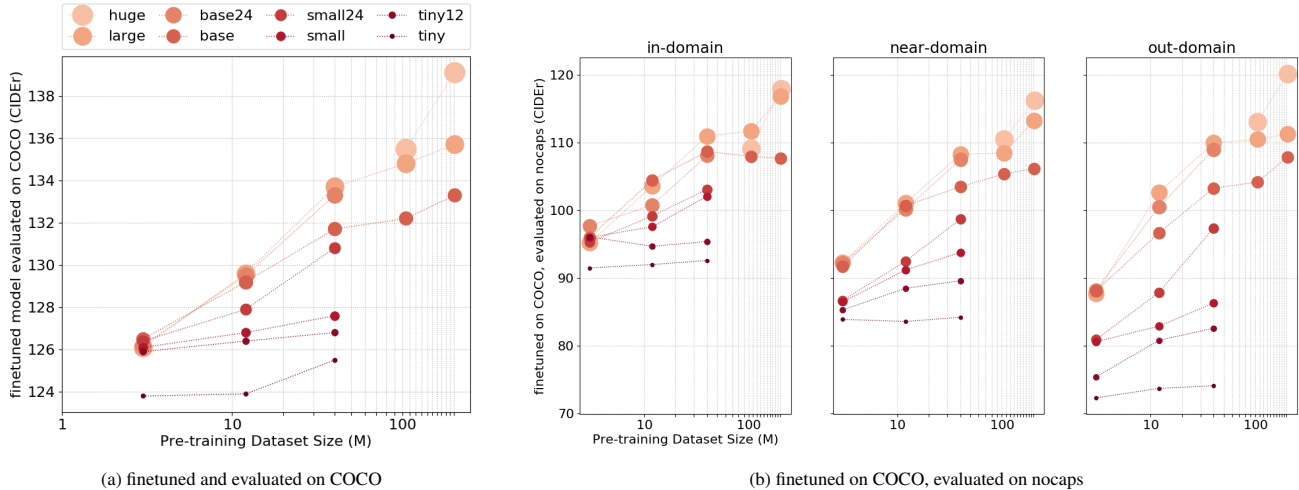


Figure 2. **Image captioning performance in data upscaling for each model size.** The x-axis shows the number of image-text pairs used in pre-training. The y-axis shows the evaluation score (CIDer) on COCO “Karpathy” test split and `nocaps` validation set, respectively. The models are first pre-trained, then finetuned on COCO caption training split. Note that x-axis is plotted in a logarithmic scale.

ters, given billion magnitude of pre-training data available.

More recently, contrastive image-text pre-training [17, 33] has also been scaled up to 400 million and 1.8 billion data sizes for image representation learning and image-text retrieval. Both CLIP [33] and ALIGN [17] employ two individual networks to encode the image and the text separately for alignment, which well fits the image-text retrieval task, but little is known about the scaling properties when it comes to image captioning.

To study the characteristics of this scaling trend on the captioning task, we first construct a large-scale image-text dataset (dubbed as ALT200M), consisting of up to 200 million image-text pairs from web based on the alt attribute of the images. Then, we conduct extensive experiments to scale VLP for image captioning from both the *data* and *model* perspectives, and name our model as LEMON 🍋, short for a Large-scale iMAGE captiONer. To simulate the process of data scaling, we create multiple subsets of ALT200M, ranging from 3 to 200 million. In terms of model, we use the state-of-the-art image captioning model VinVL [48] as our reference model, composed of an image feature extractor and a transformer model. We adapt the pre-training task to be consistent with the captioning task, and then scale the width and depth of the transformer model with the number of parameters ranging from 13 (*i.e.*, tiny) to 675 (*i.e.*, huge) millions. Combining different models and pre-training data sizes, we summarize our results in Figure 1 and 2, which characterize the linear-logarithmic scaling trend. Larger models tend to benefit more when we have more than 10 million data for pre-training. However, with only 3 million data, the performance starts to saturate early as the model size increases. Moreover, we also investigate other design choices of VLP, *e.g.*, model architectures and

training objectives.

Our contributions are summarized as follows.

- We present the VLP scaling rule for image captioning. Not only does this prove the effectiveness of learning from large-scale noisy data, but it also sheds lights on how performance can be efficiently improved by increasing the model and pre-training data sizes together to avoid a saturation plateau.
- We achieve new state-of-the-art results for image captioning across several major benchmarks, including COCO Caption, `nocaps`, and Conceptual Captions.

2. Related Work

Vision-Language Pre-training. Since the birth of ViL-BERT [30] and LXMERT [40], we have witnessed a boom of methods for vision-language pre-training [6, 7, 13, 22, 26, 37, 45, 46]. Prominent examples include UNITER [6], VL-BERT [37], OSCAR [29], UNIMO [28], and VinVL [48]. Along the journey of VLP, researchers have investigated different training strategies [11, 31], robustness [25], compression [9, 10, 42], probing analysis [3, 27], and the extension to video-text modeling [21, 24, 38, 39, 50]. More recently, instead of using object detectors for image feature extraction, end-to-end VLP based on convolution networks and transformers are becoming popular [14, 15, 19, 23, 44].

However, as another important factor in achieving superior performance, the scaling behavior of VLP is less studied. While most works pre-train transformer of base/large sizes on no more than 4M images, we train models from tiny to huge, on up to 200M images. CLIP [33] and ALIGN [17] scaled up contrastive pre-training to 400M and 1.8B images, and SimVLM [43] further use 1.8B images for prefix language modeling pre-training. However, CLIP and ALIGN

Dataset	#images (M)	#cap./image	Unigram		Caption lengths	
			#unique	#unique in 0.1% tail	mean \pm std	P5%/50%/95%
COCO Caption [5]	0.1	5	19,264	1,184	10.44 \pm 2.24	8/10/14
CC3M [35]	3.1	1	49,638	22,677	10.25 \pm 4.64	5/9/19
CC12M [4]	12.2	1	1,319,284	193,368	17.17 \pm 12.76	6/13/43
ALT200M (Ours)	203.4	1	2,067,401	1,167,304	13.01 \pm 8.85	2/11/27

Table 1. **Statistics of existing and our collected datasets.** The number of images in CC3M and CC12M are calculated for valid RGB images at the time we downloaded them. The unigrams are counted and sorted by occurrences from large to small to form a distribution curve for each dataset. Our dataset features much more long-tail concepts, as indicated by the number of unigrams included in the 0.1% distribution tail. The datasets used in CLIP [33] and ALIGN [17] are not included, since we do not know the corresponding statistics.

focus on image-text retrieval, while SimVLM did not study its scaling behavior w.r.t. pre-training data sizes. Compared with them, we focus on image captioning, provide a more comprehensive study on the scaling behavior via altering data and model sizes, and show that by using 200M images, we can outperform SimVLM on image captioning.

Scaling Law. With the success of large-scale pre-trained models in both the language and vision domains, there has been a surging research interest in discovering the empirical scaling law of these models. [18] presented that the language model performance scales as power-law across many orders of magnitude with dataset size, model size, and computation used in training. [12] further studied the scaling of autoregressive generative modeling. Aside from the model size, [41] showed that the model shape also matters for efficient transfer from upstream pre-training to downstream finetuning. In the vision domain, [47] scaled a series of vision transformer models evaluated on image classification tasks. While the scaling protocols have been investigated for many NLP and vision tasks, we are the first to study the scaling behavior of VLP for image captioning, and push multimodal transformer pre-training to a much larger scale.

In Appendix, we also provide a detailed related work review on non-pretraining-based image captioning methods.

3. Method

In this section, we present the pre-training dataset in Section 3.1, the model structure in Section 3.2, and training objective in Section 3.3.

3.1. Pre-training Dataset

We construct a data collection pipeline to crawl the images from the Internet and the associated *alt* attribute, which usually provides the description of the image content. In order to scale up easily, we follow the natural distribution of images without re-balancing, and apply only minimal rule-based filtering. We keep images with the longer side more than 200 pixels and aspect ratio smaller than 3. As some alt-texts are too long, we split them up by punctuation marks, such as period and exclamation mark, and select the longest part. To filter out some rare or misspelled words, we build



Figure 3. **Word cloud of the top 200 words** in our pre-training dataset ALT200M, excluding the stop words, e.g., a, the, of, etc.

a vocabulary of unigrams with English Wikipedia titles and body text. We remove unigrams that are present less than 5 times, resulting in approximately 250 million unique unigrams. We remove the alt-text if any of its unigrams cannot be found in the vocabulary. Afterwards, we count the frequency of all the remaining sentences, and filter out some boilerplate sentences that are too generic, e.g., stock image, 3D illustration, vector photo. For the sake of privacy, we use a Named Entity Recognition model spaCy² to identify person and location names, and replace them with special tokens \langle PERSON \rangle , \langle LOC \rangle , respectively. At last, we perform duplication check on all the collected images to ensure that they do not overlap with existing test sets, such as COCO, nocaps, and Conceptual Captions.

The final dataset, named as ALT200M, contains more than 200 million images, each corresponding to one alt-text. The word cloud of 200 most frequent words is visualized in Figure 3. As shown in Table 1, compared to CC12M, ALT200M has nearly 16 \times more images. The vocabulary is almost doubled. We observe that 56% of unigrams sum up to only 0.1% of total occurrences, characterizing an extremely long tail of rarely occurring unigrams. The average length of the captions is 13.01, more than that of the COCO caption dataset (10.44). We also observe that our dataset contains much more shorter captions with only 2 or 3 unigrams. This indicates a shift in the distribution of captions from pre-training to finetuning.

²<https://github.com/explosion/spaCy>

Model	Layers	Width	MLP	Heads	Param (M)	FLOPs
tiny	6	256	1024	4	13.4	1.1
tiny12	12	256	1024	4	18.1	1.5
small	12	384	1536	6	34.3	2.9
small24	24	384	1536	6	55.6	4.8
base	12	768	3072	12	111.7	9.5
base24	24	768	3072	12	196.7	16.8
large	24	1024	4096	16	338.3	28.9
huge	32	1280	5120	16	675.4	57.7

Table 2. **Details of model architecture.** FLOPs are calculated via taking 50 image region features and 35 text tokens as input in one forward pass. The dimension of image region feature is 2054, which is mapped to the transformer width via a linear layer.

Besides CC12M, there also exist some other large-scale image-text datasets, such as WIT [36], WenLan [16], LAION-400M [34], and the datasets used in CLIP [33] and ALIGN [17]. More detailed discussions on them are provided in Appendix.

3.2. VLP Model for Captioning

We use the pre-trained Faster R-CNN detector from [48] to extract image region features, which are concatenated with scaled bounding boxes as position encoding. Following [29, 48], we also add the detected object tags as input. The text input, including the caption and objects tags, are tokenized by WordPiece, with a vocabulary of 30522 tokens. A multi-layer transformer model is used for multimodal fusion, which consists of a stack of encoder layers, each of which has a multi-head self-attention (MSA) layer followed by a feed-forward layer. To enable text generation with the encoder layers, we use the sequence-to-sequence attention mask [49] in each self-attention layer for the captioning module. Specifically, the input consists of image embeddings $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^N$, object tag embeddings $\mathbf{T} = \{\mathbf{t}_j\}_{j=1}^M$, and token embeddings for the caption $\mathbf{W} = \{\mathbf{w}_k\}_{k=1}^L$, where N, M, L are the number of image regions, tags, and caption tokens, respectively. The corresponding outputs are:

$$\mathbf{R}_{v_i} := \text{MSA}(\mathbf{v}_i, \mathbf{V} \cup \mathbf{T}), \quad (1)$$

$$\mathbf{R}_{t_j} := \text{MSA}(\mathbf{t}_j, \mathbf{V} \cup \mathbf{T}), \quad (2)$$

$$\mathbf{R}_{w_k} := \text{MSA}(\mathbf{w}_k, \mathbf{V} \cup \mathbf{T} \cup \{\mathbf{w}_l\}_{l=1}^k), \quad (3)$$

where $\text{MSA}(\mathbf{x}, \mathbf{Y})$ is the MSA layer with \mathbf{x} mapped to query, and \mathbf{Y} mapped to key/value. \cup means concatenation of matrices, and the index of \mathbf{R}_{v_i} denotes the position corresponding to \mathbf{v}_i . The output representation is fed into the next layer, or used for prediction at the end. In this way, during inference, the model can decode the token from left

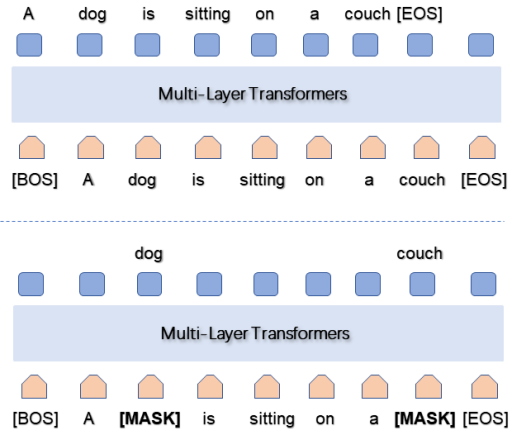


Figure 4. **Comparison of training objectives.** (Top) Language Modeling (LM), to predict the next token at each position. (Bottom) Masked Language Modeling (MLM), to predict the masked and/or possibly polluted tokens at the masked positions. Both use causal masking for model training.

to right in an auto-regressive manner. To study the scaling trend, we experiment with 8 model configurations, ranging from “tiny” of 13M parameters to “huge” of 674M parameters, detailed in Table 2.

3.3. Training Objective

While bidirectional Masked Language Modeling (MLM) has been widely used in both language and vision-language pre-training, its bidirectional nature makes it sub-optimal for text generation. In contrast to VLP works that are mostly evaluated on VL understanding tasks, we use sequence-to-sequence MLM for generation tasks. During training, we randomly mask out 15% of caption tokens following BERT [8] to form a “corrupted” caption $\tilde{\mathbf{W}} = \{\tilde{w}_k\}_{k=1}^L$, where \tilde{w}_k is either equal to w_k , or replaced with [MASK] token or another token sampled from vocabulary. The training loss is defined as:

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \mathbf{V}, \mathbf{T}) &= \sum_{k \in D} \text{CE}(\mathbf{w}_k, \mathbf{R}_{\tilde{w}_k}) \\ &= \sum_{k \in D} (-\log p(\mathbf{w}_k | \mathbf{V}, \mathbf{T}, \{\tilde{w}_l\}_{l=1}^k)), \end{aligned} \quad (4)$$

where $\text{CE}(\cdot, \cdot)$ is the cross-entropy loss with softmax, D is the subset of masked positions. The loss for the recovery of the possibly polluted tokens by intuition fits into the scenario of training with noisy captions. Note that we use the same loss in pre-training and finetuning. During inference, at step s , given the previous predicted tokens $\{\hat{w}_k\}_{k=1}^{s-1}$, we set \tilde{w}_s to [MASK], and $\tilde{w}_k = \hat{w}_k$ for $k < s$. Therefore, the generation process simulates recovering the [MASK] token at the end in each step. Since the representations of caption tokens do not depend on the subsequent tokens, the intermediate representations of predicted tokens can be saved to

#	Model	Pre-training data	in-domain		near-domain		out-of-domain		overall	
			CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Validation Set										
1	Encoder-Decoder [4]	CC3M [35]	81.8	11.6	73.7	11.1	65.3	10.1	73.2	11.0
2		CC12M [4]	88.3	12.3	86.0	11.8	91.3	11.2	87.4	11.8
3		CC3M+CC12M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1
4	VinVL _{base} * [48]	N/A	96.8	13.5	90.7	13.1	87.4	11.6	90.9	12.8
5	VinVL _{base} †	5.65M combined	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5
6	VinVL _{large} †	5.65M combined	106.3	14.5	98.0	14.0	88.8	12.6	97.3	13.8
7	SimVLM _{huge} [43]	1.8B	113.7	-	110.9	-	115.2	-	112.2	-
8	LEMON _{base}	N/A	91.4	13.3	81.4	12.5	62.6	10.6	79.0	12.3
9	LEMON _{base}	CC3M	96.0	13.8	91.7	13.2	88.1	11.8	91.6	13.0
10	LEMON _{base}	CC12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8
11	LEMON _{large}	CC12M	103.6	14.4	101.1	13.8	102.7	12.6	101.8	13.6
12	LEMON _{base}	ALT200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1
13	LEMON _{large}	ALT200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0
14	LEMON _{huge}	ALT200M	118.0	15.4	116.3	15.1	120.2	14.5	117.3	15.0
Test Set										
15	Human		80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6
16	SimVLM _{base}	1.8B	-	-	-	-	-	-	94.8	13.1
17	SimVLM _{large}	1.8B	-	-	-	-	-	-	108.5	14.2
18	SimVLM _{huge} ‡	1.8B	109.0	14.6	110.8	14.6	109.5	13.9	110.3	14.5
19	LEMON _{large}	ALT200M	111.2	15.6	112.3	15.2	105.0	13.6	110.9	15.0
20	LEMON _{huge}	ALT200M	112.8	15.2	115.5	15.1	110.1	13.7	114.3	14.9

Table 3. **Results on nocaps validation and test sets.** All our models are trained with cross-entropy loss only, without CIDEr optimization. The VinVL model with * is not pre-trained, but use SCST+CBS as reported in the paper. The VinVL results with † are reproduced by us via finetuning from the released checkpoints, which are pre-trained on the combined datasets including 5.65M images, 2.5M QAs, 4.68M captions and 1.67M pseudo-captions. The numbers with ‡ are copied from the nocaps leaderboard.

avoid duplicate computation, thereby making the generation efficient. We also experimented with other model structures and training objectives, such as predicting the next token with language modeling, as shown in Figure 4 and will be detailed later in Section 4.3.

4. Experiments

In this section, we first present our experimental setup in Section 4.1, and then detail our results in Section 4.2, followed by comprehensive analysis in Section 4.3.

4.1. Setup

Datasets. To measure the progress brought about by large-scale pre-training, we aim to evaluate the model’s capability of describing varieties of (long-tail) visual concepts, which is essential for captioning in the wild. For this purpose, we choose nocaps [1] as the evaluation benchmark, which is developed to evaluate object captioning at scale. The dataset consists of 15100 images from Open Images, and covers more than 600 object categories, of which nearly 400 of them are unseen from the training set in COCO [5]. Based on whether the image contains novel

objects unseen in the COCO training set, the nocaps images are divided into three domains: “in”, “near”, and “out”. None of the objects in the out-domain are seen in COCO. This discrepancy raises the importance of learning from external resources for recognizing novel objects, rather than relying on the clean and fully annotated captioning training data. As the external training resources may vary for different methods, in Table 3, we only compare our model with other models that also use extra image-caption pairs, and take the pre-training dataset size into account.

Implementation details. To study the scaling trend, we experiment with 8 model configurations and 5 pre-training data sizes. We train all the models from scratch if not otherwise specified. In the pre-training, we do not include COCO or Visual Genome data, to exclude the possible impact of data quality when plotting the scaling trend, as these datasets are manually annotated instead of web collected. To create pre-training dataset of different sizes, we randomly sample from ALT200M at different data scales. Note that the larger dataset is a superset of the smaller ones.

We use AdamW optimizer with linearly decaying learning rate. During pre-training, the batch size is 8192. The

Model	Pre-training data	Cross-entropy optimization				CIDEr optimization			
		B@4	M	C	S	B@4	M	C	S
Encoder-Decoder [4]	CC12M	-	-	110.9	-	-	-	-	-
VinVL _{base}	5.65M combined	38.2	30.3	129.3	23.6	40.9	30.9	140.4	25.1
VinVL _{large}		38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2
SimVLM _{base}	1.8B	39.0	32.9	134.8	24.0	-	-	-	-
SimVLM _{large}		40.3	33.4	142.6	24.7	-	-	-	-
SimVLM _{huge}		40.6	33.7	143.3	25.4	-	-	-	-
LEMON _{base}	ALT200M	40.3	30.2	133.3	23.3	41.6	31.0	142.7	25.1
LEMON _{large}		40.6	30.4	135.7	23.5	42.3	31.2	144.3	25.3
LEMON _{huge}		41.5	30.8	139.1	24.1	42.6	31.4	145.5	25.5

Table 4. **Results (single model) on COCO “Karpathy” test split.** B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE.

initial learning rate is set to 2×10^{-4} for the base and large model, and to 1×10^{-4} for the huge model. The models are trained for 60 epochs. The maximum length of image regions, tags and caption tokens are 50, 15, 20, respectively. During finetuning, the model is trained for 40 epochs with batch size 512. The initial learning rate is 1×10^{-5} , 1×10^{-6} , and 8×10^{-7} for the base, large, and huge models, respectively. During inference, the caption is generated with beam search and the beam size is 5. The generation ends when the $\langle \text{EOS} \rangle$ token is predicted, or the maximum length of 20 tokens is reached. More training details are provided in Appendix.

4.2. Captioning Results

Results on nocaps validation and test sets are shown in Table 3. By leveraging large-scale pre-training on the automatically collected alt-texts, LEMON has achieved remarkable improvement, especially for out-of-domain images. Compared to the baseline trained on COCO only (row 8), after pre-training on ALT200M (row 12), the CIDEr score is improved by 16.3 for the in-domain part, and 45.3 for the out-of-domain part. This evidences that large-scale pre-training improves the model’s ability to recognize a wide range of long-tailed visual objects. We also present results of models pre-trained on CC3M and CC12M. Compared to the best reported results on these datasets (row 1, 2), our evaluated CIDEr scores (row 9, 10) are increased by 18.4 and 13.0, respectively. This demonstrates the performance improvement in our captioning results brought about by the proposed training scheme when the pre-training dataset is the same. On the leaderboard³ test set, our large and huge models (row 19, 20) both surpassed the top-ranking model (row 18) that is pre-trained on 1.8B image-text pairs, creating the new state-of-the-art of 114.3 in CIDEr. We also achieve the state of the art on other image captioning benchmarks, including COCO Caption and Conceptual Captions, as summarized in Table 4 and 5.

³<https://eval.ai/web/challenges/challenge-page/355/leaderboard/1011>

Model	B@4	M	C	S
w/o pre-training [4]	-	-	100.9	-
pre-trained on CC12M [4]	-	-	105.4	-
LEMON _{base} w/o PT	10.1	12.1	104.4	19.0
LEMON _{base} on CC12M	10.1	11.9	108.1	19.8
LEMON _{base}	10.1	12.0	111.9	20.5
LEMON _{large}	10.8	12.3	117.4	21.0
LEMON _{huge}	13.0	13.9	136.8	23.2

Table 5. **Results on the Conceptual Captions (CC3M) dev set.** All the models are finetuned on CC3M with cross-entropy loss only. We compare with the best results reported on the dev set with and without pre-training. PT: pre-training.

Large-scale pre-training not only benefits VL representation learning, but also equips the model with the capability to zero-shot generalization. We use the pre-trained model to generate captions directly without further finetuning. The prefix “a picture of” is added as prompt to improve the quality of generated captions. Some examples are illustrated in Figure 5. The pre-trained model demonstrates strong ability in recognizing various long-tail visual concepts. Compared to the model trained only on small clean set, it shows the knowledge of many fine-grained categories (*e.g.*, “metal instrument” vs. “tuba”), which are learned from the large-scale noisy supervision of alt-texts from web. We also notice that our pre-trained model tends to generate very short descriptions when used in a zero-shot manner, but this is mitigated after finetuning on COCO. We posit that the reason for this is the relatively large proportion of short alt-texts in our pre-training datasets.

4.3. Ablation and Analysis

Scaling law: influence of data and model sizes. We conduct comprehensive experiments to understand how much gain can be obtained in the downstream tasks by scaling up pre-training. Figure 2 shows the relationship between the number of images used in pre-training and the CIDEr scores evaluated in the downstream captioning tasks. All

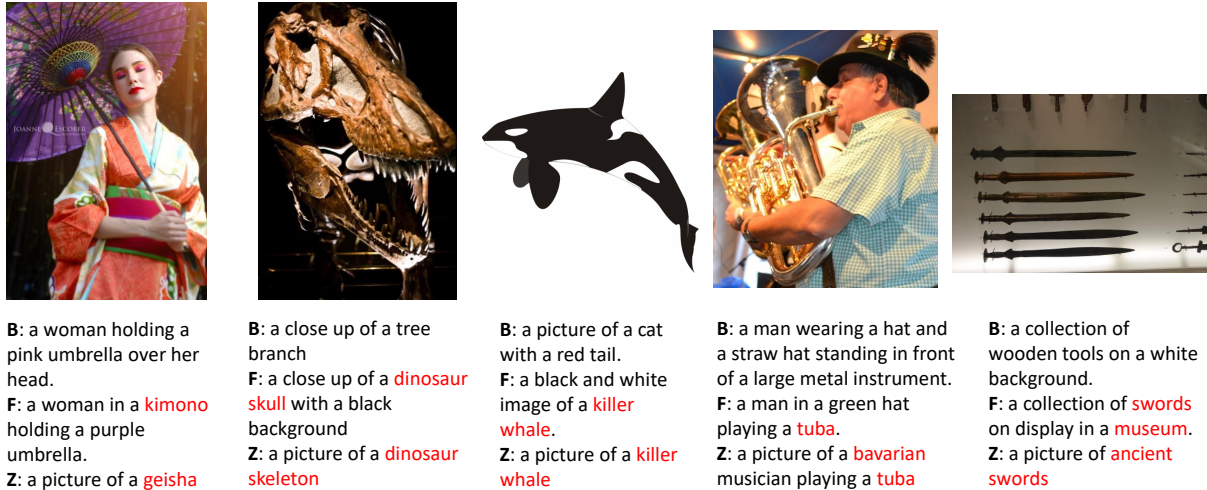


Figure 5. **Examples of generated captions on nocaps validation set.** **B:** the baseline model trained on COCO caption only without pre-training. **F:** the model finetuned on COCO after pre-training on ALT200M. **Z:** the pre-trained model without finetuning, where we add the prefix “a picture of” during inference as the prompt to improve the quality of zero-shot generation following [43].

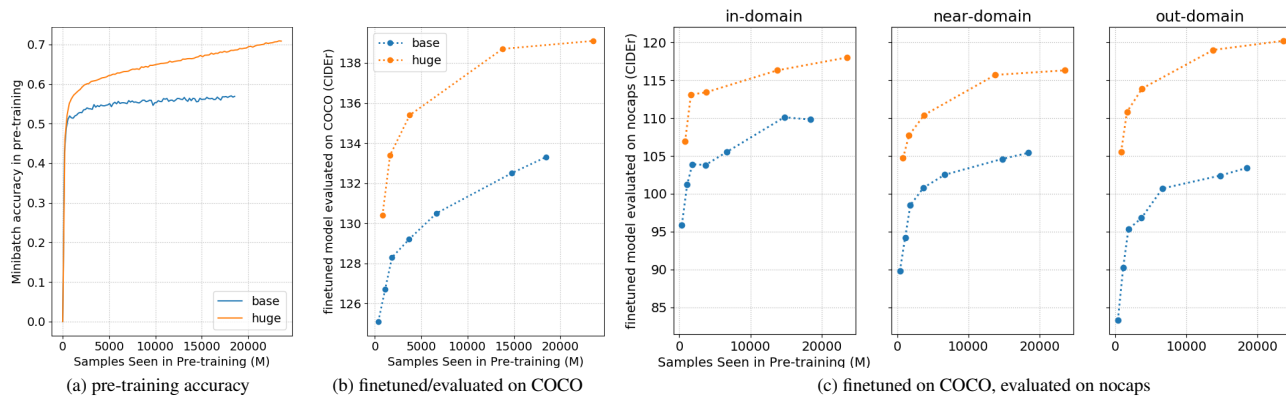


Figure 6. **Comparison of sample efficiency for different model sizes.** Figure (a) shows the learning curve in pre-training, measured by the accuracy of cross-entropy loss for masked token prediction. Figures (b) and (c) show the results of finetuned intermediate checkpoints, evaluated on COCO “Karpathy” test set and nocaps validation set, respectively. The larger model can consistently achieve better results in downstream tasks with far fewer pre-training epochs, especially for out-of-domain data.

the models are pre-trained from scratch, and then finetuned on COCO. While all the models can be improved after pre-training with more data, the improvement is obviously less significant for the smaller models than for the larger models. On COCO, the gap between “small” and “large” models is negligible at 3M scale, but it becomes large as the data size increases. Moreover, when evaluating on nocaps, the gap in out-of-domain set is consistently larger than that in in-domain. This implies the advantage of large models in transferring knowledge from pre-training to downstream tasks, especially when the finetuning data are too limited to cover all test scenarios.

Besides, we observe that the model capacity becomes the performance bottleneck as the amount of available data increases. Figure 1 plots the scaling trend w.r.t. the number of model parameters. When pre-training with 3M data, the

“base” size appears to be sufficient, and there is no significant benefit to using larger models. However, with more than 40M data, the larger models start to outperform the smaller ones by a significant margin. When the data magnitude reaches hundreds of millions, and if the observed trend from “base” to “huge” can be kept, there is promise in training an even larger model to push the limits of VLP for captioning tasks.

At last, to have a better understanding of the data quality, we perform pre-training with the same settings on CC12M and the 12M subset of ALT200M. With the only difference in pre-training data source, the models yield fairly similar results (0.1 to 0.3 differences in CIDEr) on COCO and nocaps. This indicates that our data quality is comparable to that of CC12M. The observed performance improvement should be attributed to the pre-training scale.

Arch.	Obj.	COCO		CC3M	
		CIDEr	SPICE	CIDEr	SPICE
Enc-Dec	LM	120.9	21.8	94.9	18.1
	s2s-MLM	120.4	22.1	99.9	18.9
Encoder	LM	119.2	21.5	96.1	18.0
	s2s-MLM	119.9	21.9	104.4	19.0

Table 6. **Ablation of models with different architectures, and trained with different objectives.** Results are reported on COCO Caption “Karpathy” test split and Conceptual Captions val split. All the models are trained from scratch. s2s-MLM indicates sequence-to-sequence MLM as described in Sec. 3.3.

Sample efficiency. We examine the improvement of learned representations along with the progress of pre-training. Progress is measured quantitatively by the number of image-text paired samples seen in pre-training, *i.e.*, the effective batch size multiplied by the training steps. In Figure 6, we report the results on COCO Caption after finetuning intermediate pre-trained checkpoints. We also evaluate the finetuned models on `nocaps`, indicating the ability of generalization under domain shift. We present two models in the figure, one with “base” size, the other with “huge” size. Both models are pre-trained on ALT200M.

We observe that both models continue to improve after seeing more samples in pre-training, but the larger model learns much “faster”. To achieve similar results in the downstream COCO captioning task, the base model must see more than 2 to 8 times more samples in pre-training. This factor is even greater when evaluating on the `nocaps` out-of-domain images. The result of the “base” model seeing 19 billion samples is still slightly worse than that of the “huge” model seeing 0.8 billion samples. This demonstrates the efficiency of large models in learning from large-scale data, as well as the robustness in generalization.

Further ablation. We compare with other common model structures and training objectives, such as the encoder-decoder transformer model and unidirectional language modeling (LM). Experiments are conducted with models of “base” size as specified in Table 2. For the encoder-decoder structure, we use 6 encoder layers (with self-attention) followed by 6 decoder layers (with cross-attention after self-attention), while other model configurations remain unchanged. The training objectives are illustrated in Figure 4. For each experiment setting, we sweep the hyperparameters, *e.g.*, pre-training epochs from 40 to 200, finetuning epochs from 10 to 60, and learning rates from 1×10^{-6} to 3×10^{-5} . The results of the best hyperparameters are reported.

We train the models under 4 different settings on COCO and CC3M, respectively. Results are summarized in Table 6. On COCO, the differences among the 4 settings are small (1.41% relative change in CIDEr), with the worst

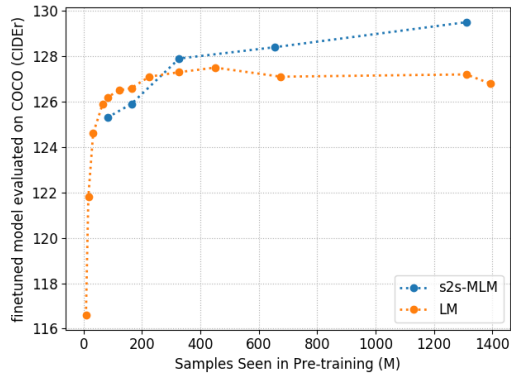


Figure 7. **Comparison of different training objectives** by pre-training on CC12M and finetuning on COCO. The models are finetuned from intermediate checkpoints using the same objective as used in pre-training.

being 119.2 from encoder+LM, and the best being 120.9 from encoder-decoder+LM. In contrast, on CC3M, the difference is much larger (9.10% relative change in CIDEr). The worst is 94.9 from encoder-decoder+LM, while the best is 104.4 from encoder+MLM. As CC3M is collected over the Internet and contains much more noise, we assume that the model that tends to overfit is prone to error when data quality is low, even though it performs well given well-annotated data.

Moreover, to compare training objectives, we first pre-train the models on CC12M, using s2s-MLM or LM, then finetune the intermediate checkpoints on COCO. As shown in Figure 7, we observe that although the model trained with LM converges faster at the beginning, it enters saturation early, and does not achieve scores as high as the model using s2s-MLM. We also find that training with LM is very sensitive to learning rates. Given the above results, we choose the s2s-MLM model and the encoder structure to scale up with the noisy pre-training data.

5. Conclusions

In this paper, we study the scaling behavior of VLP models for image captioning, and construct our own large-scale dataset ALT200M. Our experiments show that scaling up pre-training leads to remarkable improvement for the downstream captioning tasks. Our model LEMON has achieved new state-of-the-arts on multiple benchmarks, including COCO Caption, `nocaps`, and Conceptual Captions. LEMON also has impressive capability of recognizing a wide range of long-tail visual objects, even in the zero-shot manner. Moreover, our study on large transformer models indicates that with orders of magnitude larger training data available, the model capacity tends to be the bottleneck. It is a promising direction to train a substantially larger model to take more advantage from the large amounts of alt-text data widely circulated on the Internet.

References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 1, 5
- [2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1
- [3] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *ECCV*, 2020. 2
- [4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1, 3, 5, 6
- [5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 3, 5
- [6] 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 *ECCV*, 2020. 1, 2
- [7] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 2
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 4
- [9] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *ICCV*, 2021. 2
- [10] Zhe Gan, Yen-Chun Chen, Linjie Li, Tianlong Chen, Yu Cheng, Shuohang Wang, and Jingjing Liu. Playing lottery tickets with vision and language. *arXiv preprint arXiv:2104.11832*, 2021. 2
- [11] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 2
- [12] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. 1, 3
- [13] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. In *AAAI*, 2021. 2
- [14] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, 2021. 2
- [15] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2
- [16] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021. 4
- [17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2, 3, 4
- [18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1, 3
- [19] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2
- [20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1
- [21] 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 *CVPR*, 2021. 2
- [22] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 2
- [23] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2
- [24] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020. 2
- [25] Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020. 2
- [26] 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. 1, 2
- [27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *ACL*, 2020. 2
- [28] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL*, 2021. 2
- [29] Xiujuan Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 1, 2, 4

- [30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 2
- [31] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 2
- [32] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011. 1
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4
- [34] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4
- [35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypertexted, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1, 3, 5
- [36] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*, 2021. 4
- [37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 1, 2
- [38] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019. 2
- [39] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 2
- [40] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1, 2
- [41] Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021. 1, 3
- [42] Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujuan Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Minivlm: A smaller and faster vision-language model. *arXiv preprint arXiv:2012.06946*, 2020. 2
- [43] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1, 2, 5, 7
- [44] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing intermodality: Visual parsing with self-attention for vision-language pre-training. In *NeurIPS*, 2021. 2
- [45] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *CVPR*, 2021. 2
- [46] 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. In *AAAI*, 2021. 2
- [47] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021. 1, 3
- [48] Pengchuan Zhang, Xiujuan Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 1, 2, 4, 5
- [49] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020. 1, 4
- [50] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 2