Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts

Soravit Changpinyo, Piyush Sharma, Nan Ding, Radu Soricut
Google Research
schangpi,piyushsharma,dingnan,rsoricut@google.com

Abstract

The availability of large-scale image captioning and visual question answering datasets has contributed significantly to recent successes in vision-and-language pre-training. However, these datasets are often collected with overrestrictive requirements inherited from their original target tasks (e.g., image caption generation), which limit the resulting dataset scale and diversity. We take a step further in pushing the limits of vision-and-language pre-training data by relaxing the data collection pipeline used in Conceptual Captions 3M (CC3M) [54] and introduce the Conceptual 12M (CC12M), a dataset with 12 million image-text pairs specifically meant to be used for vision-and-language pre-training. We perform an analysis of this dataset and benchmark its effectiveness against CC3M on multiple downstream tasks with an emphasis on long-tail visual recognition. Our results clearly illustrate the benefit of scaling up pre-training data for vision-and-language tasks, as indicated by the new state-of-the-art results on both the nocaps and Conceptual Captions benchmarks.[1]

Figure 1: **CC12M** Even when the alt-texts do not precisely describe their corresponding Web images, they still provide rich sources for learning long-tail visual concepts such as sumo, mangosteen, and jellyfish. We scale up vision-and-language pre-training data to 12 million by relaxing overly strict filters in Conceptual Captions [54].

1. Introduction

Transfer learning using pre-training and fine-tuning has become a prevalent paradigm in computer vision, natural language processing, and vision-and-language (V+L) research. It has been shown, for instance, that V+L pre-training leads to transferrable joint representations that benefit multiple downstream V+L tasks, including visual question answering, image and text retrieval, and referring expression comprehension [42, 36, 12, 58, 2, 56, 66, 35, 43].

What makes V+L pre-training successful? On one hand, this is due to advances in architectures and modeling that are mainly inspired by BERT and similar models in natural language understanding and generation [15, 40, 62, 33, 16, 51]. In particular, the idea of using flexible self-attention mechanisms via high-capacity multi-layer Transformers [59], in combination with self-supervised learning objectives such as masked language modeling [15], has proven to be effective and widely applicable. On the other hand, the availability of large-scale labeled and weakly-labeled data in the V+L domain [46, 11, 30, 54] is truly what enables such models to learn associations between the two modalities.

In *either* vision *or* language community, one notable trend is that scaling up training data is useful. In contrast, datasets in V+L research remain relatively limited in terms of scale and diversity. The capability of JFT-300M [57] and Instagram [45] over orders-of-magnitude smaller ImageNet [53] has been put to test on multiple downstream image classification and object detection tasks. In NLP, the size of pre-training data sources for training deep language models rose from the 20GB BooksCorpus [68]+English Wikipedia in BERT[15], to the 570GB dataset in GPT-3 [9] and the 745GB C4 dataset in T5 [51].

[1]Our dataset is available at https://github.com/google-research-datasets/conceptual-12m.

In contrast, V+L datasets are limited in two ways. First, the *effective* sizes of popular V+L datasets are low. The number of images in these datasets range from fewer than a few hundred thousands [64, 11, 31, 18] to several millions [54], with lower text quality as the scale increases. Second, many of the small-sized datasets share the same, limited visual domain; COCO-Captions [11], Visual Genome [31], and VQA2 [17] are (mostly) based on several hundreds thousand of COCO images [39]. The lack in scale and diversity of visual concepts (with respect to vision/language-only counterparts) makes it hard for V+L models to perform adequately in the wild.

One major reason for these gaps is the difficulty in collecting such datasets. Unlike in image classification, "text" in V+L datasets is longer and less likely to be agreed upon, making the annotation process more costly and time-consuming. One approach to remedy this is to make use of large amounts of the alt-texts accompanying images on the Web. For instance, Sharma et al. introduced Conceptual Captions (CC3M) [54], a dataset of 3.3M ⟨image, caption⟩ pairs that result from a filtering and post-processing pipeline of those alt-texts. Despite being automatically collected, CC3M is shown to be effective in both image captioning in the wild [54, 10] and V+L pre-training [42, 36, 12, 58, 2, 56, 66, 35, 43]. In other words, it provides a promising start for large-scale V+L annotations.

In this paper, we explore pushing the limits of V+L data using this approach. Our key insight is that specific downstream V+L tasks (e.g., VQA, image captioning) can be overly restrictive if the goal is to collect large-scale V+L annotations. For instance, CC3M was collected to favor high-precision texts that are fit for the downstream task of image captioning. Yet, we have witnessed this dataset being increasingly adopted for V+L pre-training [42, 12, 2, 56, 66, 35, 43], arguably beyond its original purpose.

We hypothesize that the V+L field could benefit from such an insight, and therefore we introduce Conceptual 12M (CC12M), a high(er)-recall V+L dataset for the purpose of V+L pre-training. By relaxing multiple image and text filters used in CC3M, we obtain a less precise but 4x larger V+L set of ⟨image, text⟩ pairs. We perform an analysis of this dataset and show that it covers a wider range of visual concepts.

We test our hypothesis by benchmarking the effectiveness of CC12M as a pre-training data source on several V+L tasks, in comparison to CC3M. We explore two main pre-training strategies (and more in the Supplementary material): one for vision-to-language generation and the other for vision-and-language matching. Our experiments indicate that scaling up pre-training V+L has a dramatic positive effect on image captioning, novel object captioning, and (zero-shot) image retrieval.

In summary, our main contributions are:

(a) A public larger-scale V+L pre-training dataset that covers a much wider range of concepts than existing ones.
(b) Evaluation on downstream vision-to-language generation and vision-and-language matching with an emphasis on long-tail recognition that consistently shows the superiority of this dataset over CC3M.
(c) State-of-the-art results on the nocaps (novel object captioning) and Conceptual Captions benchmarks.

2. Vision-and-Language Pre-Training Data

We first review the data collection pipeline for the Conceptual Captions 3M (CC3M) outlined in Sect. 3 of [54], which we followed closely. We then describe a series of relaxation and simplification to the pipeline that results in CC12M, a much larger set of image-text pairs. Finally, we perform an analysis of CC12M in comparison with CC3M and other existing V+L datasets.

2.1. Conceptual Captions 3M: Pipeline for extracting and cleaning Image Alt-Text from the Web

The Conceptual Captions dataset consists of about 3.3M Web images and their corresponding cleaned, hypernymized Alt-texts [54]. This approach leverages a promising source of (weak) supervision for learning correspondance between visual and linguistic concepts: once the pipeline is established, the data collection requires no additional human intervention. It consists of the following 4 steps: (i) image-based filtering based on size, aspect ratio, encoding format and offensive content, (ii) text-based filtering based on language, captialization, token frequency, pre-defined unwanted phrases, as well as part-of-speech (POS), sentiment/polarity, and adult content detection (using Google Cloud Natural Language APIs), (iii) image-text–based filtering based on the number of image tags (as predicted by Google Cloud Vision APIs) that overlap with the existing text, (iv) text transformations, most notably hypernymization of named entities, including proper names of persons, organizations and locations (e.g., both "Harrison Ford" and "Calista Flockhart" are replaced by "actor"), deletion of time-related spans, and digit replacement (using # as a digit abstraction).

The large scale nature and the high degree of textual and visual diversity make this dataset particularly suited to V+L pre-training [42, 12, 56, 66, 35, 43].

2.2. CC12M: Relaxing filters for higher recall

Conceptual Captions has been created to work out-of-the-box for training image captioning models, and thus it involves substantial image, text, and image-text filtering and processing to obtain clean, high-precision captions. As a result, this approach comes at the cost of low recall (many potentially useful ⟨image, Alt-text⟩ pairs are discarded). However, this trade-off may not be optimal if the dataset is to

Dataset	# examples	token/type	caption length
CC3M train	3,318,333	804.8	10.3 ± 4.5
CC12M	12,423,374	370.0	20.2 ± 16.3

Table 1: Basic statistics of CC12M vs. CC3M

be used primarily for V+L pre-training. Motivated by this, we follow a similar procedure as the one described in [54] but relax some of its filters, and construct the dataset called Conceptual 12M (CC12M), as detailed below.

Filtering. As described above, the construction of CC3M used three main filtering types [54]: image-based, text-based, and image-text–based. To arrive at CC12M, we keep the image-text filtering intact, and relax the unimodal filters only. First, for image-based filtering, we set the maximum ratio of larger to smaller dimension to 2.5 instead of 2. We still keep only JPEG images with size greater than 400 pixels, and still exclude images that trigger pornography detectors. Second, in text-based filtering, we allow text between 3 and 256 words in the alt-text. We still discard candidates with no noun or no determiner, but permit ones without prepositions. We discard the heuristics regarding high unique-word ratio covering various POS tags and word capitalization. We set the maximum fraction of word repetition allowed to 0.2. Given a larger pool of text due to the above relaxations, the threshold for counting a word type as rare is increased from 5 to 20.

Text transformation. The main motivation for CC3M to perform text transformation is that a majority of candidate captions contain ultrafine-grained entities such as proper names (people, venues, locations, etc.), making it extremely difficult to learn as part of the image captioning task. In contrast, we are not restricted by the end task of image caption generation. Our intuition is that relatively more difficult pre-training data would lead to better transferability. We thus do not perform hypernimization or digit substitution as in [54]. The only exception to the "keep alt-texts as raw as possible" rule is performing person-name substitutions, which we identify as necessary to protect the privacy of the individuals in these images. For this step, we use the Google Cloud Natural Language APIs to detect all named entities of type Person, and substitute them by a special token ⟨PERSON⟩. Around 25% of all the alt-texts in CC12M are transformed in this fashion.

2.3. Characteristics of CC12M

We provide an analysis of CC12M along multiple dimensions, focusing on comparing it to the most relevant CC3M. Additional analyses are in the supplementary material.

Basic statistics. As seen in Table 1, CC12M consists of 12.4M image-text pairs[2], about 4x larger than CC3M. It has a much lower token (word count) to type (vocab size) ratio,

Figure 2: Word clouds of top 100 tokens in CC3M (the top cloud) and in CC12M (the bottom cloud).

indicating a longer-tail distribution and a higher diversity degree of the concepts captured. Lastly, the average caption length of CC12M is much longer. This is overall achieved by our relaxation of the filters, especially the text one.

Quality. We compute a rough estimate of precision on 100 examples by asking two annotators to rate how well the given alt-text fits the image on a 1–5 scale: 1 (no fit), 2 (barely fit), 3 (somewhat), 4 (good fit, but disfluent language), 5 (perfect). We define precision as the fraction of captions with a score 4 or above. We see a drop in precision, 76.6% vs. 90.3% as reported for CC3M (Table 2 in [54]). This analysis points to the precision/recall tradeoff in transitioning from CC3M to CC12M. Fig. 1 illustrates such a tradeoff: the "jellyfish" example would have been filtered out from CC3M (due to a high percentage of nouns and a lack of proprositions), but it is included in CC12M.

Visual concept distribution. We use the caption text tokens to represent the visual concepts. The long tail of visual concepts that emerge in CC12M spans many categories, and can be attributed to (1) a dramatic increase in scale, and (2) the absence of fine-grained entity hypernymization. We list some of them here to illustrate this point, in the format of "⟨word⟩ ⟨frequency in CC3M⟩ $\rightarrow$ ⟨frequency in CC12M⟩": luffy 0 $\rightarrow$ 152, mangosteen 0 $\rightarrow$ 212, zanzibar 0 $\rightarrow$ 1138, sumo 1 $\rightarrow$ 661, pokemon 1 $\rightarrow$ 8615, chevrolet 1 $\rightarrow$ 12181, mehndi 3 $\rightarrow$ 9218, pooh 4 $\rightarrow$ 7286, cyberpunk 5 $\rightarrow$ 5247, keto 6 $\rightarrow$ 6046, hound 9 $\rightarrow$ 3392, quiche 50 $\rightarrow$ 1109, durian 61 $\rightarrow$ 552, jellyfish 456 $\rightarrow$ 2901.

We also visualize the head of the distribution in Fig. 2. We observe that "person" becomes much more frequent due to person substitution with the token "⟨PERSON⟩". Moreover, there are fewer "actor", "artist", "(football) player", as a result of removing hypernymization.

Finally, we inspect tokens that are unseen in CC3M.

[2]Extracted as of May 2020.

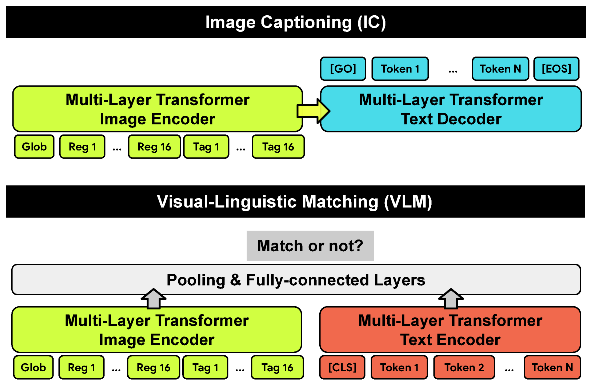

Figure 3: **Main Pre-Training Tasks:** image captioning (vision-to-language generation) and visual-linguistic matching (vision-and-language understanding).

Downstream task	Downstream datasets	
	Train	Eval
Novel object captioning	COCO Captions	nocaps
Novel object captioning	LocNar COCO	LocNar OID
Image captioning	CC3M	
Zero-shot IR	None	Flickr30K
IR	Flickr30K	
IR	LocNar Flickr30K	

Table 2: Generation (top) and matching (bottom) tasks and datasets considered in this paper. IR = Image Retrieval.

We observe that these tokens may occur very frequently in CC12M if they are fine-grained instances such as locations (“france,” “africa,” “dc,” “toronto”) or digits (“2019”, “10”, “2018”, “2020”). This is due to the removal of hypernymization and the dropping of time-related span deletion.

Biases. We study the context in which several sensitive terms related to gender, age, race, ethnicity appear such as “black”/“white”/“asian”/“african”/“indian”, “man”/”woman”, “young”/“old”, etc. We observe no large biases in the distribution of these terms, either in terms of co-occurrence between sensitive term pairs or with other tokens. Furthermore, we check the distribution of web domains and, similar to visual concepts, we find this to be diverse and long-tail: >100K with >40K contributing >10 samples. We take our preliminary study as a positive indication of no severe biases stemming from particular domains or communities. Finally, we provide a Broader Impact statement in the supplementary material.

3. Evaluating Vision-and-Language Pre-Training Data

The previous section describes CC3M and our CC12M. In this section, we evaluate both datasets on their ability to benefit V+L downstream tasks, measuring the impact from visual grounding produced under the two settings. For the sake of comparison, we do not include the images that appear in CC3M in CC12M in our experiments.

We focus on the two most fundamental V+L tasks: vision-to-language **generation** and vision-and-language **matching**. In both cases, our emphasis is on (i) the simplest setting in which the learning objectives during pre-training and downstream tasks match, and (ii) long-tail recognition and out-of-distribution generalization, as we believe this is where pre-training has the most impact. Fig. 3 and Table 2 summarize our experimental setup, in terms of the downstream tasks and the fine-tuning and evaluation datasets.

3.1. Vision-to-Language Generation

3.1.1 Pre-Training Tasks

We use **image captioning (`ic`)** as the pre-training task. The task is to predict the target caption given image features. To train the model parameters, we use the standard cross entropy loss given the groundtruth caption.

Note that there exist vision-to-language generation pre-training strategies that are different from ours. For instance, Zhou et al. [66] adapt BERT [15] to generate text. As masked language modeling is used for pre-training, there is no decoder and, at inference time, text is generated using the encoder network one token at a time, appending the mask token to the image and the text generated so far. Thus, this approach is inefficient as the number of passes over the input image is linear in the desired caption length. It is also unclear how to incorporate advanced decoding schemes such as beam search, top-k sampling, or nucleus sampling (see, e.g., [21]) with such an approach. Finally, our experiments (see Supplementary material) suggest that the `ic` pre-training task is superior to its masked variants and justify using the simple `ic` learning objective.

3.1.2 Downstream Tasks

Our downstream tasks are selected to measure progress toward solving image captioning in the wild. They also stand to benefit from visual grounding, especially since pre-training, by definition, is expected to cover a wider range of (long-tail) visual concepts than fine-tuning datasets.

nocaps [1] is a recent object-captioning-at-scale benchmark consisting of 4,500 validation and 10,600 test images with 10 hidden reference captions. Unlike in the standard image captioning setting, nocaps’s distributions of images during training (COCO Captions) and evaluation (Open Images) are different: the Open Images dataset [32, 29] covers one order of magnitude more objects (600 classes) than COCO [39] (80 classes). This discrepancy defines the challenge: solutions must be able to learn to describe novel concepts from sources external to the COCO training set, such as text corpora, knowledge bases, or object detection datasets. In the Supplementary material, we provide details on the nocapsleaderboard. In addition, besides CC3M and CC12M, we also explore using the Open Images Localized Narratives dataset (LocNar) [49], as an alternative "in-domain" (from a visual standpoint) pre-training data source.

Localized Narratives (LocNar) [49] is a collection of datasets with images that are paired with captions obtained by converting speech to text via ASR and manual post-processing it[3]. Inspired by the setting in nocaps, we use the COCO[39] portion (train split of around 130K images) for training/fine-tuning, and Open Images [32] portion of evaluation (val split of around 40K images). Note that the LocNar captions are much longer than standard captioning datasets (41.8 words/caption), setting it apart from nocaps.

Conceptual Captions 3M [54] is our main reference for V+L pre-training data source. At the same time, the image captioning task on this dataset itself is a valuable benchmark for vision-to-language generation in the wild. Thus, we adopt it as a downstream task for CC12M. This means that, in the case of CC3M, from-scratch and pre-training settings collapse.

Evaluation metrics. To measure the performance on image caption generation, we consider the standard metrics BLEU-1,4 [47], ROUGE-L [38], METEOR [8], CIDEr-D [60], and SPICE [3].

3.2. Vision-and-Language Matching

3.2.1 Pre-training Tasks

In **visual-linguistic matching (`vlm`)**, the task takes as input both image and text features and predicts whether the input image and text are matched. To train the model's parameters, we use a contrastive softmax loss, for which the original image-text pairs are used as positive examples, while all other image-text pairs in the mini-batch are used as negative examples [42, 58].

3.2.2 Downstream Tasks

The task of **caption-based image retrieval (IR)** is to identify a relevant image from a pool given a caption describing its content. The Flickr30K dataset [48] consists of 31,000 images from Flickr, each associated with five captions. Following existing work [34, 42], we use 1,000 images for validation, 1,000 images for testing, and use the rest of image-text pairs for model training.

We further consider **zero-shot caption-based image retrieval** [42] on the Flickr30K dataset. The term "zero-shot" refers to the setting in which we discard training data and apply pre-trained models "as-is", i.e., without fine-tuning on the target task.

Finally, we further evaluate our retrieval system on the Localized Narratives dataset [49] (see Sect. 3.1.2). We use the LocNar Flickr30K portion (train split of 30,546 images, and test split of 1000 images) for training and evaluation.

Evaluation metrics. To measure the performance on image retrieval, we consider the standard metrics Recall@1 (R1), Recall@5 (R5), and Recall@10 (R10).

3.3. Implementation Details

Representing images and texts. We use Graph-RISE [27, 28] to featurize the entire image. We train a Faster-RCNN [52] on Visual Genome [30], with a ResNet101 [19] backbone trained on JFT [20] and fine-tuned on ImageNet [53]. We select top-16 box proposals and featurize each of them with Graph-RISE, similar to [10]. Inspired by [37], we obtain up to 16 image tags from the Google Cloud Vision APIs, and treat them as text inputs to our model. These global, regional, and tag features end up being represented as a bag of 1+16+16 vectors, serving as bottom-up features [6] for our model.

Model and Learning. For `ic`-based pre-training and downstream tasks, we follow the state-of-the-art architecture that heavily rely on self-attention [59] or similar mechanisms [54, 65, 10, 23, 13]. We implement a Transformer-based encoder-decoder model, using [10] as a starting point. In addition, we encode each feature vector with a deeper embedding layer and apply layer normalization [7]. Following [42], we encode the corners and the area of bounding boxes and apply layer normalization when combining geometric and regional semantic features. These modifications lead to an improved CIDEr score of 100.9 on the CC3M dev benchmark (Table 7), vs. 93.7 as reported by [10]. We describe additional details in the supplementary material, including infrastructure description, runtime, model size, hyperparameter ranges and tuning methods, and the configuration of the best-performing model.

For the `vlm`-based pre-training and downstream tasks, we reuse the architecture above but discard the decoder. We use mean pooling to obtain a fixed-length vector for each modality, and compute the product of the transformed (last-layer Transformer encoder representation) image and the transformed text before applying softmax.

4. Experimental Results

4.1. Vision-to-Language Generation

Table 3 shows our results on **nocaps**. We report in Row 1 the performance of our baseline model without pre-training. Rows 2-3 show the performance of off-the-shelf captioning systems trained on CC3M and CC12M, respectively. This indicates the "raw" power (zero-shot setting) of the pre-trained network in generating captions out of the box. We note that, without fine-tuning on COCO Captions, the model underperforms our baseline numbers on all

[3]This dataset also contains mouse traces synchronized with the text, but we do not use the traces here.

Pretraining data	Train or finetune on coco cap?	nocaps val											
		in-domain		near-domain		out-of-domain		overall					
		CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	BLEU1	BLEU4	METEOR	ROUGE	CIDEr	SPICE
None	✓	72.8	11.1	57.1	10.2	34.1	8.3	69.8	14.5	21.9	47.9	54.7	10.0
CC3M		29.2	7.4	27.5	6.9	37.3	7.4	36.0	2.8	12.6	29.1	29.7	7.1
CC12M		20.7	6.9	24.1	6.9	41.6	8.0	31.8	2.9	12.1	26.8	27.1	7.2
CC3M	✓	81.8	11.6	73.7	11.1	65.3	10.1	74.6	19.1	24.1	51.5	73.2	11.0
CC12M	✓	88.3	12.3	86.0	11.8	91.3	11.2	78.5	23.4	25.9	54.5	87.4	11.8
CC3M+CC12M	✓	**92.6**	**12.5**	**88.3**	**12.1**	**94.5**	**11.9**	**79.2**	**24.4**	**26.1**	**55.1**	**90.2**	**12.1**

Table 3: Automatic metric scores on the nocaps val set: performance of from-scratch (Row 1), pre-trained (Rows 2-3), and fine-tuned (Rows 4-5) models. CC12M outperforms CC3M by a large margin after fine-tuning (Row 4 vs. 5). Together, they achieve a new best, surpassing 90 CIDEr points on nocaps val. Bold indicates best-to-date, underline indicates second-best.

Figure 4: **Qualitative results on nocaps**. Each example comes with a caption predicted by the model that is trained on COCO Captions without pre-training (very top, right under the image), as well as captions predicted by models pre-trained on CC3M (middle) and CC12M (bottom), where the left/right column indicates if the model is fine-tuned on COCO Captions.

metrics, which is indicative of the need for the model to learn the COCO captioning style, to which the existing automatic metrics are quite sensitive. In addition, we observe a slightly better performance by CC3M except for BLUE4 and SPICE. This illustrates the benefit of data processing and bias toward high-precision captions present in CC3M.

With a fine-tuned model, the benefit of transfer learning using pre-training on this task is clear (Row 1 vs. Rows 4,5,6), with CC12M outperforming CC3M by +14.2 CIDEr points and another +2.8 with CC3M+CC12M. Fig. 4 illustrates this effect; scaling up pre-training data benefits learning multimodal correspondences from a much larger pool of concepts, potentially making the model less susceptible to hallucinations (e.g., guessing "microphone" as it has not seen "bagpipes" in the training set), and also more informative (e.g. choosing "sumo wrestlers" over "men"/"people").

Table 4 compares our best model (`ic` pre-trained on CC3M+CC12M) to existing state-of-the-art results on nocaps, and show that ours achieves state-of-the-art performance on CIDEr, outperforming a concurrent work [22] that uses a different pre-training approach *directly on the Open Images dataset, which nocaps is based on.* Importantly, we observe that the gain in the overall score can be largely attributed to the out-of-domain performance (3rd column). This result indicates that, although the annotation protocol for nocaps uses the priming of annotators to mention one or more of displayed fine-grained ground-truth object classes (e.g., "red panda") present in the image [1], the large-scale and natural fine-grainedness of CC12M succeeds in correctly learning to generate captions containing such concepts, in spite of being textually out-of-domain.

Following [1], we also report results of our best model on the COCO Captions val2017 split, see Table 5, with 5K and 10K fine-tuning steps. We note that, since we do not rely on techniques such as constrained beam search (CBS) [4, 1] that constrain the model outputs, we do not suffer from the large performance trade-offs seen with the previous solutions (degradation on in-domain performance as out-of-domain performance increases, see each model vs. "reference"). Our result on out-of-domain data, as we vary the number of fine-tuning steps (last two rows), suggests that over–fine-tuning on COCO Captions may incur a cost in terms of poor generalization.

A second set of results is reported in Table 6. We observe that, even when the task requires the generation of much longer captions for LocNar, CC12M achieves superior performance (as measured by CIDEr) compared to CC3M as pretraining data. However, the gain is smaller compared to

Method	in-domain CIDEr	in-domain SPICE	near-domain CIDEr	near-domain SPICE	out-of-domain CIDEr	out-of-domain SPICE	overall CIDEr	overall SPICE
	nocaps val							
UpDown [1]	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1
UpDown + CBS [1]	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
UpDown + ELMo + CBS [1]	79.3	12.4	73.8	11.4	71.7	9.9	74.3	11.2
Oscar$_L$ [37]	79.9	12.4	68.2	11.8	45.1	9.4	65.2	11.4
Oscar$_L$ + CBS [37]	78.8	12.2	78.9	12.1	77.4	10.5	78.6	11.8
Oscar$_L$ + SCST + CBS [37]	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4
VIVO [22]	88.8	12.9	83.2	12.6	71.1	10.6	81.5	12.2
VIVO + CBS [22]	90.4	13.0	84.9	12.5	83.0	10.7	85.3	12.2
VIVO + SCST + CBS [22]	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4
pretrain `ic` *on CC12M*	88.3	12.3	86.0	11.8	91.3	11.2	87.4	11.8
pretrain `ic` *on CC3M+CC12M*	**92.6**	12.5	**88.3**	12.1	94.5	11.9	**90.2**	12.1
Human	84.4	**14.3**	85.0	**14.3**	**95.7**	**14.0**	87.1	**14.2**
	nocaps test							
UpDown [1]	74.3	11.5	56.9	10.3	30.1	8.1	54.3	10.1
UpDown + ELMo + CBS [1]	76.0	11.8	74.2	11.5	66.7	9.7	73.1	11.2
VIVO + SCST + CBS [22]	**89.0**	12.9	**87.8**	12.6	80.1	11.1	86.6	12.4
pretrain `ic` *on CC12M*	82.9	11.9	85.7	12.0	85.3	11.3	85.3	11.8
pretrain `ic` *on CC3M+CC12M*	87.2	12.3	87.4	12.1	87.2	11.4	**87.3**	12.0
Human	80.6	**15.0**	84.6	**14.7**	**91.6**	**14.2**	85.3	**14.7**

Table 4: Comparison between our best model (in *italics*, pre-trained on CC12M with `ic` and fine-tuned on COCO Captions) and existing models, on the nocaps val (top) and test (bottom) splits. Bold indicates best-to-date, underline indicates second-best.

Method	COCO val2017	nocaps val
	CIDEr	CIDEr
UpDown (reference)	116.2	55.3
UpDown + CBS	97.7	73.1
UpDown + ELMo + CBS	95.4	74.3
no pretrain (reference)	108.5	54.7
pretrain `ic` *on CC12M* (5K)	108.1	**87.4**
pretrain `ic` *on CC12M* (10K)	**110.9**	87.1

Table 5: Performance on the in-domain COCO Captions val2017 split along with the nocaps val split. Our methods are in *italics* with the number of fine-tuning steps in the parentheses.

Pretraining data	Finetuning data	LocNar COCO val	LocNar OID val
		CIDEr	CIDEr
None	LocNar COCO	29.6	33.8
CC3M	LocNar COCO	29.1	35.7
CC12M	LocNar COCO	**30.0**	**38.6**

Table 6: Novel object captioning on LocNar.

the one observed for nocaps. We attribute this to the fact that injecting novel concepts into longer texts is harder, and also the fact that LocNar does not use priming in their annotation process, leading to more generic terms in their annotation ("musical instruments" vs. "trumpets").

Finally, we fine-tune our best pre-trained model (`ic` on CC12M) using CC3M in Table 7, and then evaluate on the dev split. We find that we improve the CIDEr score on the dev split from 100.9 to 105.4 (+4.5 CIDER points). We note that the model trained on CC12M and evaluated directly on the CC3M dev set (without fine-tuning on the CC3M train

Method	CC3M dev	CC3M test
	CIDEr	CIDEr
FRCNN [10]	89.2	94.4
TTIC+BIU (single model)	-	98.0
Ultra [10]	93.7	98.4
no pretrain	100.9	-
pretrain `ic` *on CC12M (no ft)*	39.3	-
pretrain `ic` *on CC12M*	**105.4**	-

Table 7: Performance on the Conceptual Captions (CC3M) benchmark. Our methods are in *italics*. "ft" stands for fine-tuning. The top two CC3M test CIDEr baseline scores are from the Conceptual Captions Leaderboard as of Nov 15, 2020.

split) obtains a low dev CIDEr of 39.3. This again indicates that the additional processing steps done for CC3M (e.g., hypernimization) result in captions that are different enough from the ones in CC12M to require a fine-tuning step.

4.2. Vision-and-Language Matching

Table 8 reports zero-shot and default IR performance on **Flickr30K** as well as default IR performance on **LocNar Flickr30K**. The results are consistent with those in vision-to-language generation. First, both CC3M and CC12M are beneficial, improving over "from-scratch" training (Pretraining data as "None") by at least 8.6% and 6.6% in R1 on Flickr30K and LocNar Flickr30K, respectively. Additionally, CC12M significantly outperforms CC3M in all cases. Finally, combining the two datasets (CC3M+CC12M) results in even better performance. We provide qualitative results and additional discussion in the supplementary ma-

Pretraining data	Finetuning data	`Flickr30K test`		
		R1	R5	R10
None	Flickr30K	43.7	74.8	84.1
CC3M	None	35.4	65.2	76.2
CC12M	None	42.5	73.1	83.4
CC3M+CC12M	None	**47.1**	**76.4**	**83.4**
CC3M	Flickr30K	52.3	81.7	88.4
CC12M	Flickr30K	58.5	86.6	92.1
CC3M+CC12M	Flickr30K	**61.5**	**87.5**	**92.8**
Pretraining data	Finetuning data	`LocNar Flickr30K test`		
		R1	R5	R10
None	LocNar Flickr30K	54.5	85.0	91.0
CC3M	LocNar Flickr30K	61.1	88.2	93.7
CC12M	LocNar Flickr30K	70.2	92.1	95.6
CC3M+CC12M	LocNar Flickr30K	**71.0**	**93.0**	**97.0**

Table 8: Image retrieval on Flickr30K and LocNar Flickr30K

terial.

Our zero-shot IR results (the three rows in Table 8 with fine-tuning data as "None") are also competitive to the state-of-the-art, despite the fact that our model is much smaller (6 layers of transformers of hidden layer size 512 with 8 attention heads vs. 12 layers of size 768 with 12 attention heads) and uses late fusion instead of early fusion. In particular, our zero-shot IR on CC3M outperforms the one in ViLBERT [42] (35.4 vs. 31.9 in R1), while the CC12M performance goes up by +7.1% R1 to 42.5, and an additional +4.6% R1 to 47.1 when using CC3M+CC12M, surpassing the "from-scratch" setting.

5. Related Work

V+L Pre-training. V+L pre-training research makes use existing large-scale datasets with image-text pairs. A majority of these resources are image captioning datasets. CC3M [54] has been the most popular for pre-training [42, 43, 2, 56, 66, 35, 12, 37]. Smaller but less noisy SBU Captions [46] ($\tilde{1}$M) and COCO Captions [11] (106K) datasets are also of high interest. Some work [58, 12, 37] use V+L resources collected for dense captioning or visual question answering (VQA), such as VG [31], VQA2 [17], and GQA [24]. In contrast, CC12M is not collected for specific target tasks, and thus it is order-of-magnitude larger than those datasets.[4] Furthermore, it is much more visually diverse, especially given the fact that COCO Captions, VG, VQA2, GQA are built on top of COCO images [39] or its subsets.

Objectives in V+L pre-training research are largely influenced by BERT [15]. Masked language modeling has been extended to visual region inputs, while the next sentence prediction is analogous to `vlm`. Based directly upon BERT, V+L pre-training research has largely been focused on V+L *understanding* [42, 36, 12, 58, 2, 56, 35, 43], with classification or regression tasks that do not involve generation. One exception is UnifiedVL [66], which pre-trains a unified architecture for both image captioning (generation) and VQA (understanding). Our work focuses on simpler objectives and consider one at a time. This allows for a "clean" study of the effect of pre-training data sources. At the same time, we also pre-train vision-to-language generation and encoder-decoder jointly as opposed to an encoder-only setup. Our work also shows that `ic` is a strong objective for vision-to-language generation with respect to the widely-used masking-based objectives. Consistent with our results, `ic` is successfully adopted for learning visual representations for lower-level vision tasks [14].

Long-tail Visual Recognition in V+L. Addressing long-tail distributions of visual concepts is an important component of V+L systems that generalize, as long and free-form texts exhibit a large number of compositional, fine-grained categories [67, 41, 10]. Our work focuses on downstream testbeds for V+L research that require this adaptation ability. For example, the train-test distribution discrepancy in nocaps exists in both visual (COCO vs. Open Images) and textual domains (80 to object classes vs. 600 classes). The same can be said for zero-shot image retrieval [42], in which the model must generalize visually and textually from the pre-training data sources of CC3M or CC12M to Flickr30K. Our work identifies pre-training with large-scale noisy data as a promising solution. In addition, for the task noval object captioning, our approach works more robustly across in- and out-of- domain scenarios and is simpler than the state-of-the-art techniques that utilize constrained beam search (CBS) [4], finite state machine construction plus CBS [5], generating slot-filling templates [44, 61], and copying mechanisms [63].

6. Conclusion

We introduce the new V+L pre-training resource CC12M, obtained by extending the pipeline in [54]. We show that the scale and diversity of V+L pre-training matters on both generation and matching, especially on benchmarks that require long-tail recognition such as nocaps. Our results indicate leveraging noisy Web-scale image-text pairs as a promising direction for V+L research.

Acknowledgments. We thank Peter Anderson for his feedback on earlier version of the draft, Bo Pang, Zhenhai Zhu for helpful discussions, Sebastian Goodman and Ashish V. Thapliyal for help with model implementation, Chris Alberti for help with the data collection pipeline, and Harsh Agrawal for detail on nocaps.

[4]Recently appearing after we submitted our paper, ALIGN [26], CLIP [50], WIT [55], WenLan [25] all explore enlarging Web-scale data for V+L pre-training with success (albeit with different focuses), further confirming our intuition that scale is a critical factor.

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.

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *EMNLP-IJCNLP*, 2019.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *ECCV*, 2016.

[4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017.

[5] Peter Anderson, Stephen Gould, and Mark Johnson. Partially-supervised image captioning. In *NeurIPS*, 2018.

[6] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[7] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[8] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshops*, 2005.

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

[10] Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In *EMNLP-IJCNLP*, 2019.

[11] 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.

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Learning UNiversal Image-TExt Representations. In *ECCV*, 2020.

[13] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *CVPR*, 2020.

[14] Karan Desai and Justin Johnson. VirTex: Learning visual representations from textual annotations. In *CVPR*, 2021.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[16] Sebastian Goodman, Zhenzhong Lan, and Radu Soricut. Multi-stage pretraining for abstractive summarization. *arXiv preprint arXiv:1909.10599*, 2019.

[17] 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 *CVPR*, 2017.

[18] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *ECCV*, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Proceedings of NeurIPS workshop*, 2015.

[21] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020.

[22] 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. *arXiv preprint arXiv:2009.13682*, 2020.

[23] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *CVPR*, 2019.

[24] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 2019.

[25] 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.

[26] 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. *arXiv preprint arXiv:2102.05918*, 2021.

[27] Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. Graph-RISE: Graph-regularized image semantic embedding. *arXiv preprint arXiv:1902.10814*, 2019.

[28] Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. Ultra fine-grained image semantic embedding. In *WSDM*, 2020.

[29] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://g.co/dataset/openimages*, 2017.

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and

Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[31] 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. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018.

[33] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*, 2020.

[34] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[35] 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.

[36] 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.

[37] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, 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 *ECCV*, 2020.

[38] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

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

[40] 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. *arXiv preprint arXiv:1907.11692*, 2019.

[41] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

[42] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[43] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.

[44] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018.

[45] Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

[46] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2Text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002.

[48] 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 *ICCV*, 2015.

[49] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020.

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

[51] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[53] 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*, 115(3):211–252, 2015.

[54] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[55] 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.

[56] 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 *ICLR*, 2020.

[57] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017.

[58] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2019.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[60] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015.

[61] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *MM*, 2018.

[62] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[63] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, 2017.

[64] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

[65] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *arXiv*, 1905.07841, 2019.

[66] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pretraining for image captioning and VQA. In *AAAI*, 2020.

[67] Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *CVPR*, 2014.

[68] 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 *ICCV*, 2015.