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

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

A. Broader Impact

Our publicly-available V+L pre-training resource CC12M has the potential to positively impact multiple vision-and-language tasks. One main aspect that we have identified is a much higher degree of coverage of long-tail visual concepts than previous resources, including CC3M. As a result, we expect the models (pre-)trained on our data to be more robust in the wild than before.

In addition, our work could benefit the design of new setups for the downstream tasks that shift away from in-domain (e.g., COCO/Visual Genome) to out-of-domain/in-the-wild (e.g., OID), similar to nocaps that our work focuses heavily on. The setups could also avoid the use of in-domain data during *pre-training* that in some cases resulting in transfer learning between (almost) identical sets of images, e.g., COCO, Visual Genome (VG), VQA2, VG QA, Visual7W, GQA, GuessWhat, and RefCOCO*.

At the same time, datasets curated from the Web could come with risks such as unsuitable content (adult content, profanity) and unintended privacy leakage [27, 10, 11]. We take the steps in Sect. 2.2 of the main text to mitigate both of these risks by applying the necessary image and text filtering steps and replacing each person name (celebrities' included) with the special <PERSON> token.

Less specific to the Web data are the unwanted dataset biases [6, 33, 37] that are prone to amplification by machine learning models [5, 39]. Our preliminary analysis in Sect. 2.3 of the main text shed light on the degree to which our data exhibits some aspects of these inherent biases, and we suspect that the better coverage of the tail in fact makes this issue less severe. Nevertheless, the users of this data and the systems trained on it shall be aware of such risks and other ones that might arise.

B. Additional analyses of CC12M

B.1. Out-of-domain (OOD) visual concepts on an expanded list of datasets

We use the 394 nocaps' out-of-domain classes as a proxy for OOD visual concepts and analyze popular vision-and-language datasets, in addition to CC3M and CC12M that we focus in the main text. These datasets span a wide range of use cases, both in terms of tasks (image-to-text generation, image-and-text matching, visual question answering (VQA), referring expression comprehension, and multimodal verification), and in terms of the stage during which they are used (pre-training, fine-tuning/evaluation, or both.)

- CC3M [31] An instance of text is the caption associated with each image url of the training split.
- CC12M (ours) An instance of text is the caption associated with each image url. It has been used and is currently the most popular V+L pre-training dataset [24, 2, 13, 34, 40, 25, 23].
- COCO Captions [12] An instance of text comes from the caption associated with each image of the 2017 training split (five captions per image). This dataset is designed for the task of image captioning, and has been used for caption-based image retrieval as well. It has been used for V+L pre-training [36, 22, 13, 23].
- Visual Genome [20] An instance of text comes from the caption of each region in images of the training split. This dataset aims to connect vision and language through scene graphs and is used for multiple tasks that include but not limited to dense image captioning, visual relationship detection and scene graph parsing, image retrieval and generation, and visual question answering. It has been used for V+L pre-training [36, 13].
- SBU Captions [28] An instance of text is the caption associated with each image url of the "preferred" version of the dataset. This dataset is designed for the

Dataset	Freq		Freq (per 1M)	
	median	mean	median	mean
CC3M	462	2325.7	139.2	700.8
CC12M	**3110**	**13455.8**	**250.3**	**1083.1**
COCO Captions	37	248.6	62.3	417.1
Visual Genome	133	1114.47	40.7	341.4
SBU Captions	121	798.6	121.0	798.6
VQA2	37	242.0	63.8	417.2
RefCOCOg	1	21.2	8.8	186.4
NLVR2	4	79.9	11.6	245.5

Table 1: Statistics of the (normalized) frequency of nocaps' out-of-domain visual concepts in the texts of popular vision-and-language datasets.

task of image captioning. It has been used for V+L pre-training [36, 13, 21, 23].

- VQA2 [16] An instance of text is the question and the answers in each image-question-answers triplet of the train2014 + val2train2014 splits. This dataset is designed for the task of visual question answering (VQA) [3]. It has been used for V+L pre-training [36, 23].
- RefCOCOg [26] An instance of text is the referring expression in each region in images of the training split. This dataset is designed for the task of referring expression comprehension [17].
- NLVR2 [35] An instance of text comes from the caption associated with each pair of images of the training split. This dataset is used for the task called multimodal verification in [25], but designed for the general task of visual reasoning.

Table 1 summarizes the number of instances whose texts contain OOD visual concepts for all selected datasets. We use both the absolute frequency and the normalized one (per 1M text instances). Essentially, these numbers indicate the degree of OOD coverage. We find that CC12M has many more OOD instances than all other datasets by a large margin (6.7x median and 5.8x mean vs. the second best CC3M). Moreover, CC12M still prevails *even after normalization* to account for its size. In other words, CC12M covers these OOD classes better in both absolute and relative senses.

Fig. 1 provides a more complete picture of the normalized frequency of OOD classes in these datasets, at different thresholds. It shows the number of OOD classes (y-axis) with at least K per 1M captions (x-axis). Evidently, other datasets experience sharper drops as K increases than CC12M (black solid curve). We also find that captioning datasets (solid curves) generally provide better coverage than non-captioning datasets: VQA2, RefCOCOg, and NLVR2 (dashed curves).

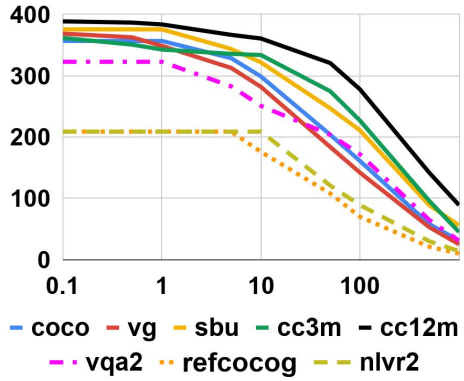

Figure 1: Comparison of nocaps' out-of-domain coverage degree among captioning (solid) and 3 other tasks' (dashed) datasets (see text for details).

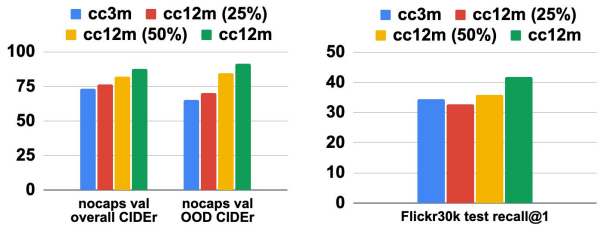

Figure 2: Performance with sub-sampled CC12M (25% & 50%) on novel object captioning (left, CIDEr's on nocaps val) and zero-shot IR (right, recall@1 on Flickr30K test).

B.2. The impact of the dataset size

We experiment with pre-training on randomly subsampled CC12M, 25% (3.1M) and 50% (6.2M) and evaluate the pre-trained models on novel object captioning on nocaps and zero-shot IR on Flickr30K. Fig. 2 shows the larger, the better trend, with 25% of CC12M gives rise to similar performance as CC3M.

C. Qualitive Results for Image Retrieval

Fig. 3 provides qualitative image retrieval results on the Flickr30K dataset, top-3 images retrieved by the from-scratch model trained on Flickr30K, as well as by two models pre-trained on CC3M and CC12M and then fine-tuned on Flickr30K. We report three cases in which CC12M pre-training helps correct the rankings from the other two models, which we suspect due to the model getting more familiar with the rare words, highlighted in blue.

D. Pre-Training: Data and Method Variants

D.1. Vision-to-Language Pre-Training on LocNar Open Images

Table 2 considers pre-training on LocNar Open Images for the nocapsbenchmark. We observe inferior performance to both CC3M and CC12M. We attribute this to the long

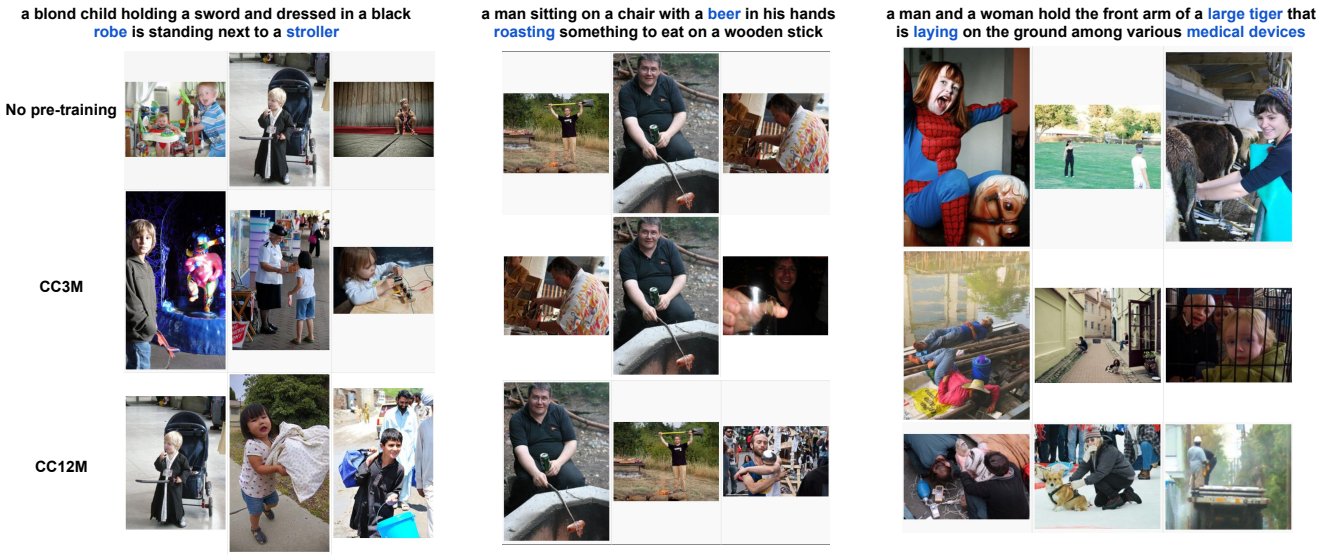

Figure 3: **Qualitative results for the image retrieval task** on Flickr30K given the query text (very top) when the model is not pre-trained (top), pre-trained on CC3M (middle), and pre-trained on CC12M (bottom).

Pre-training data	nocaps val											
	in-domain		near-domain		out-of-domain		overall					
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	BLEU1	BLEU4	METEOR	ROUGE	CIDEr	SPICE
LocNar Open Images	76.0	11.6	65.9	10.9	48.9	9.3	73.3	17.4	23.5	50.7	63.9	10.7
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

Table 2: Comparison between pre-training data. LocNar Open Images's images are from the same visual domain as nocaps. All approaches use the `ic` pre-training objective.

narratives in LocNar having drastically different styles from those from COCO Captions and nocaps. Furthermore, the data collection protocol in nocaps does not involve priming the annotator to mention object names present to the user, resulting in more generatic terms (instrument vs. guitar). This again highlights the natural fine-grainedness inherent in noisy Web data, especially in the case of non-hypernymized data source (CC12M).

D.2. Pre-Training Strategies

In the main text, we focus on the image captioning (`ic`) and the visual-linguistic matching (`vlm`) learning objectives both during pre-training and fine-tuning stages. Our motivation here is to keep the setup for evaluating pre-training data as "clean" as possible. However, other pre-training strategies exist in the literature and we describe and test the effectiveness of them in this section.

D.2.1 Masked Vision-to-Language Generation

Given the training image-text pairs, the `ic` objective predicts the text from the image. The following objectives predict (all or part of) the text from the image *and* (all or part of) the text. In order to *encode* both the image and the text, we concatenate the sequence of image feature vectors and the sequence of text token feature vectors, and use the Transformer encoder to encode them [22, 13, 34]. This vanilla fusion is effective, shown to consistently outperform the co-attentional transformer layer [24, 25], in which the "query" comes from the other modality than that of "key" and "value" (see Sect. 2 and Fig. 2 in [24] for details).

Masked Language Modeling (`mlm`). We mask a percentage of the input text tokens at random, and predict the target text sequence using the decoder. Following BERT [15], we use a mixed strategy for masking: for each selected token, we replace it with the mask token [MASK] 80% of the time, replace it with a random token 10% of the time, and leave it as is 10% of the time.

Masked Sequence to Sequence Modeling (`mass`). We apply the mixed masking strategy as in `mlm` to the input text tokens, but require that the mask is applied to consecutive tokens (i.e., a contiguous segment). The task is to sequentially predict the masked segment using the decoder. This approach is inspired by MASS [32] and PEGASUS [38].

Results. Table 3 compares `ic`, `mlm`, and `mass` pre-training

Pre-training objective	nocaps val											
	in-domain		near-domain		out-of-domain		overall					
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	BLEU1	BLEU4	METEOR	ROUGE	CIDEr	SPICE
`ic`	88.3	12.3	86.0	11.8	91.3	11.2	78.5	23.4	25.9	54.5	87.4	11.8
`mlm[.1]`	76.4	11.5	68.4	10.8	57.6	9.6	73.0	18.1	23.5	50.6	67.4	10.6
`mlm[.2]`	79.8	11.3	76.3	10.9	76.2	10.2	76.2	20.5	24.1	52.4	76.8	10.8
`mlm[.4]`	86.5	12.3	82.7	11.5	86.3	11.3	78.0	22.7	25.2	53.7	84.0	11.6
`mlm[.8]`	89.3	12.5	87.5	11.9	91.1	11.3	78.7	23.8	25.9	54.4	88.5	11.9
`mass[.1]`	86.0	12.1	74.8	11.1	71.7	10.1	75.8	20.5	24.6	52.5	75.8	11.0
`mass[.2]`	84.9	12.0	78.1	11.2	78.6	10.5	76.0	20.8	24.7	52.7	79.2	11.2
`mass[.4]`	85.7	11.7	83.7	11.5	88.5	10.9	77.3	22.8	25.1	53.6	85.0	11.4
`mass[.8]`	88.8	12.2	85.1	11.7	87.8	10.6	78.1	23.7	25.5	54.2	86.2	11.5

Table 3: Comparison between the `ic` pre-training and masked V+L pre-training. We consider two masking schemes (`mlm` and `mass`) and four masking rates (.1, .2, .4, .8) and report their effects on the nocaps val set.

Pre-training objectives	nocaps val											
	in-domain		near-domain		out-of-domain		overall					
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	BLEU1	BLEU4	METEOR	ROUGE	CIDEr	SPICE
`ic`	88.3	12.3	86.0	11.8	91.3	11.2	78.5	23.4	25.9	54.5	87.4	11.8
`ic+vlm`	88.6	12.3	85.8	11.9	90.0	11.4	78.0	23.1	25.7	54.4	87.1	11.9
`ic+moc`	91.1	12.4	88.4	12.1	93.6	11.4	78.8	24.6	26.2	55.2	89.9	12.0

Table 4: Effect of visual linguistic matching (`vlm`) and masked object classication (`moc`) when combined with the `ic` objective on the nocaps val set.

objectives. Our main observation is that `ic` clearly outperforms masked vision-to-language pre-training when the masking rate is low. Overall, `ic` is competitive to `mlm` and `mass`, slightly below `mlm[.8]` in overall CIDEr, but higher on out-of-domain CIDEr.

In addition, the trend suggests that it is critical that the text masking rate is high enough such that the models become less and less reliant on text — that is, when `mlm` and `mass` become more similar to the `ic` task. Note that widely-used configurations in the VLP literature on vision-and-language understanding are the ones with low text masking rates (0.2 in most cases), which consistently underperform in our generation setup.

We attribute this result to the models' (over)reliance on text during pre-training, which hurts the quality of its *image* representations. Supporting evidence for this phenomenon is found in the recent work of [9], which observe that image+text pre-trained models exhibit a preference for attending text rather than images during inference (in image and text understanding task). Another supporting evidence is the issue of strong language priors (well-known in the VQA community), which led to interest in *adversarial* test sets and other methods to overcome strong language biases [1, 30, 14, 7]. The same pheonmenon has been reported for multi-modal machine translation, where models trained on image+text tend to ignore the image and primarily use the text input [8]. Based on these results, the design of V+L pre-training objectives that are capable of outperforming the image-only `ic` objective (i.e., overcoming the language through modeling) is an interesting venue for future work.

Another observation is that `mass` significantly works better than `mlm` for lower masking rates. When masking rates are high, the two objectives become more similar. This suggests the importance of bridging the gap between pre-training and fine-tuning (producing consecutive tokens).

D.2.2 Image Captioning with Visual-Linguistic Matching or Masked Object Classification

We explore adding auxiliary losses to the main `ic` objective. First, we define a pre-training task that does not require text.

Masked object classification (`moc`). We mask one of the visual regions (selected at random), and predict the cluster ID of that region [24, 36, 13]. We use a total of 8192 clusters, obtained via K-means over the training data.

Then, we either add the `vlm` loss (multipled by 0.1) or the `moc` loss (multipled by 0.1) to the main `ic` loss.

Results. Table 4 reports the effect of multi-task pre-training on the nocaps val set. We observe a slight improvement when adding `moc` but a slight drop when adding `vlm`. This again shows that `ic` is a good pre-training task to start with. We leave developing advanced auxiliary losses on top of it and multi-task pre-training strategies for future work.

E. Implementation Details

E.1. Data Preprocessing and Feature Embedding

- Text tokenizer: preprocesed with COCO tokenizer https://github.com/tylin/coco-caption. We then create a vocabulary of subtokens out of these.
- Text input embedding (during pre-training only): subtoken lookup embeddings of size E = 512 are randomly initialized, followed by Linear(512)-ReLU-Dropout(0.3)-Linear(512).
- Image's geometric features: two pairs of coordinates (top left and bottom right) and the relative area, represented by *relative* numbers between 0 and 1. Each of these 5 numbers is linearly projected into an embedding of size E = 512. We concatenate the result to get an embedding of size E x 5 = 2560, followed by Linear(512)-ReLU-Dropout(0.3)-Linear(512).
- Image's semantic features: each feature vector (a global image feature vector or one of the 16 box's image feature vector, followed by Linear(512)-ReLU-Dropout(0.3)-Linear(512).
- Image's combined geometric and semantic features: we first apply LayerNorm [4] to each of the geometric or the semantic features. We then add the two and apply Linear(512)-ReLU-Dropout(0.3)-Linear(512)-LayerNorm.
- Image's tag features: same as text input embedding.

For the `ic` objective, we have a bag of 1 + 16 visual feature vectors and up to 16 tag feature vectors, each of size 512. For the `vlm` objective, where text has to be encoded, we also have a sequence of text (sub)token feature vectors of size 512.

E.2. Model

The `ic`-based task uses a transformer encoder-decoder model. The `vlm`-based uses two transformer encoders, one for texts and the other for images.

- Transformer image encoder: number of layers L = 6.
- Transformer image encoder: vocab embedding size E = 512.
- Transformer image encoder: hidden embedding size H = 1024.
- Transformer image encoder: feedforward/filter size F = H x 4 = 4096, following [15].
- Transformer image encoder: number of attention heads A = H / 64 = 8, following [15].
- Transformer text encoder (for `vlm` only): L, E, H, F, A are the same as Transformer image encoder.
- Transformer decoder: L, E, H, F, A are the same as Transformer image encoder.
- Transformer decoder: beam search width = 5.
- Transformer decoder: beam search alpha = 0.6.
- Transformer decoder: maximum output length = 36 for all datasets except for LocNar which is set to 180.

E.3. Training

- Infrastructure: Google Cloud 32-core TPUs.
- Batch size per core: 128 (for a total of 4096)
- Optimizer: Adam [19] with default hyperparameters (except for the initial learning rate; see below).
- Learning rate — Initial: See Hyperparameter search below.
- Learning rate — Warm-up epochs: 20 for all pre-training and fine-tuning experiments.
- Learning rate — Decay rate: 0.95 for all pre-training and fine-tuning experiments.
- Learning rate — Decay epochs: 25 for all pre-training and fine-tuning experiments.
- Data augmentation: a set of input visual regions are permuted during training.
- Maximum number of steps: 2M for vision-to-language generation pre-training on both CC12M and CC3M (and CC3M+CC12M). For vision-and-language matching, 1M for CC3M instead. See Hyperparameter search below for fine-tuning experiments.

E.4. Evaluation

For nocaps evaluation, we submit inference results to the leaderboard https://evalai.cloudcv.org/web/challenges/challenge-page/464/overview. Code for all evaluation metrics can be found at https://github.com/nocaps-org/updown-baseline/blob/master/updown/utils/evalai.py. For in-depth discussions of these metrics see [18].

Participating in the default formulation of the nocaps challenge requires that one (i) does not use val and test Open Images's ground-truth object detection annotations, and (ii) does not use image-caption data collected via additional annotation protocols. We satisfy both requirements as we train our object detector on Visual Genome, and both CC3M and CC12M are automatically harvested from the web (alt-text) and belong to the category of noisy web data, therefore satisfying the second requirement. On the other hand, models that leverage the Open Images Localized Narratives dataset (LocNar) [29] for pre-training belong to the nocaps (XD) leaderboard rather than the default one.

Some of our results on the CC3M benchmark are taken from the leaderboard, which is located at https://ai.google.com/research/ConceptualCaptions/leaderboard?active_tab=leaderboard.

E.5. Hyperparameter search

For pre-training experiments, we do not conduct hyperparameter tuning besides an initial stage of exploration as we believe small changes would not considerably affect the downstream performance. For instance, we fix an initial

learning rate to 0.000032 and observe it works consistently well (on the validation set) across scenarios.

For fine-tuning experiments, we focus on tuning one hyperparamter: the initial learning rate. In the case of nocaps, we also lightly tune the maximum number of training steps as we observe the model overfitting on COCO Captions. In all cases, we make sure to allocate similar resources to any two settings that we make a comparison between, such as pre-training data sources of CC3M and CC12M.

For generation, the ranges for the initial learning rate are {3.2e-9, 3.2e-8, 3.2e-7} and the ranges for the maximum number of training steps are {5K, 10K}. For matching, the ranges for the initial learning rate are {3.2e-8, 3.2e-7, 3.2e-6} while the maximum number of training steps is fixed to 10K.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018. 4

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *EMNLP-IJCNLP*, 2019. 1

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 2

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, 2016. 1

[6] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 1

[7] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases in visual question answering. In *NeurIPS*, 2019. 4

[8] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *NAACL*, 2019. 4

[9] 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. 4

[10] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In {*USENIX*} *Security*, 2019. 1

[11] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020. 1

[12] 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

[13] 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. 1, 2, 3, 4

[14] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP-IJCNLP*, 2019. 4

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3, 5

[16] 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. 2

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2

[18] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *EACL*, 2017. 5

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[20] 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. 1

[21] 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

[22] 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, 3

[23] 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. 1, 2

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 3, 4

[25] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 1, 2, 3

[26] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 2

[27] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE SP*, 2019. 1

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

[29] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 5

[30] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*, 2018. 4

[31] 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. 1

[32] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. In *ICML*, 2019. 3

[33] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *ECCV*, 2018. 1

[34] 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. 1, 3

[35] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2018. 2

[36] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2019. 1, 2, 4

[37] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019. 1

[38] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*, 2020. 3

[39] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017. 1

[40] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, 2020. 1