ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang♢, Seong Joon Oh♣

NAVER AI Lab
♢ Now at Google Research ♣ Now at University of Tübingen

Abstract. Image-Text matching (ITM) is a common task for evaluating the quality of Vision and Language (VL) models. However, existing ITM benchmarks have a significant limitation. They have many missing correspondences, originating from the data construction process itself. For example, a caption is only matched with one image although the caption can be matched with other similar images and vice versa. To correct the massive false negatives, we construct the Extended COCO Validation (ECCV) Caption dataset by supplying the missing associations with machine and human annotators. We employ five state-of-the-art ITM models with diverse properties for our annotation process. Our dataset provides ×3.6 positive image-to-caption associations and ×8.5 caption-to-image associations compared to the original MS-COCO. We also propose to use an informative ranking-based metric mAP@R, rather than the popular Recall@K (R@K). We re-evaluate the existing 25 VL models on existing and proposed benchmarks. Our findings are that the existing benchmarks, such as COCO 1K R@K, COCO 5K R@K, CxC R@1 are highly correlated with each other, while the rankings change when we shift to the ECCV mAP@R. Lastly, we delve into the effect of the bias introduced by the choice of machine annotator. Source code and dataset are available at https://github.com/naver-ai/eccv-caption

1 Introduction

Image-caption aligned datasets (*e.g.*, MS-COCO Caption [13,40], Flickr30k [49], Conceptual Caption [10,56]) have become *de-facto* standard datasets for training and evaluating Vision-Language (VL) models. Particularly, Image-to-Text Matching (ITM) tasks [5,11,12,15,18,20–22,25,26,33,37,39,59,64–66,69] are widely used benchmarks for evaluating a VL model. The existing ITM benchmark datasets are built by annotating captions (by alt-texts [10,50,56], web crawling [17], or human annotators [13]) for each image without considering possible associations with other images in the dataset. The collected image-caption pairs are treated as the only positives in the dataset, while other pairs are considered the negatives. However, in practice, there exists more than one caption to describe one image. For example, the description "The man is playing tennis with a racket" may describe multiple images with tennis players equally well

The man is playing tennis with a racket
A man taking a swing at a tennis ball
A man on a court swinging a tennis racket
A tennis player swinging a racket at a ball
A person hitting a tennis ball with a tennis racket
The man is playing tennis on the court
A man with a tennis racket swings at a tennis ball

Fig. 1: **Inherent multiplicity of correspondences in MS-COCO Caption.** While any image-caption pair above makes sense (positive pair), only red and blue image-caption pairs are marked as positive in MS-COCO Caption.

(Figure 1). We have observed that the number of missing positives is tremendous; there exist $\times 3.6$ positive image-to-caption correspondences and $\times 8.5$ caption-to-image correspondences than the original MS-COCO dataset.

While the huge number of false negatives (FNs) in VL datasets is potentially sub-optimal for training VL models, it is downright detrimental for evaluation. For example, the small number of positive correspondences of image-caption-aligned datasets limits the evaluation metrics.[1] In other tasks, such as image retrieval [34, 41, 46, 63], the positives and negatives are defined by class labels; hence, the number of possible matched items is large enough to measure precision or mean average precision (mAP) metrics. On the other hand, because existing ITM benchmarks only have one positive correspondence for each item, they are only able to use recall-based metrics (*e.g.*, Recall@k) that are known to be less informative than the precision- or ranking-based evaluation metrics [44]. In this paper, we focus on correcting the FNs in the evaluation dataset and the recall-based evaluation metrics to make a fair comparison of VL models.

As our first contribution, we correct the FNs in MS-COCO Caption by constructing Extended COCO Validation (ECCV) Caption dataset. We annotate whether each MS-COCO image-caption pair is positive with human workers. The labor cost for this process scales quadratically with the size of the dataset (*e.g.*, MS-COCO has 76B possible image-caption pairs, while the number of images is only 123K). Since verifying every possible image-text pair is not scalable, we subsample the queries in the dataset and reduce the number of candidates for positive matches with the machine-in-the-loop (MITL) annotation process. MITL lets a model reduce the number of candidate positives; then human annotators evaluate the machine-selected candidates. We employ five state-of-the-art ITM models with distinct properties as machine annotators; CLIP [50], ViLT [32], VSRN [39], PVSE [59], and PCME [15]. After post-processing, ECCV Caption contains 1,261 image queries (originally 5,000) but with 17.9 positive captions per image query on average (originally 5). It also contains 1,332 caption queries (originally 25,000) with 8.5 positive images per caption (originally 1).

[1] In MS-COCO Caption, a caption is only matched to one image, and an image is matched to five captions. Other datasets usually have one caption for each image.

While the use of a machine annotator is inevitable for the sake of scalability, the choice of a particular model may bias the dataset towards the specifics of the model. This can be problematic because different models show different filtered results to the human annotators, which brings the impartialness of the annotated dataset towards any particular model to the surface. In other words, the MITL annotations are not stable across model choices. Our studies show that the underlying ML model conditions the annotated dataset towards favoring certain models over the others. Therefore, this practice could lead to the danger of biased evaluation results using such datasets. We show that the rankings among the VL models can be arbitrarily shifted by modifying the underlying ML model. Our study also shows that using multiple machine annotators can alleviate machine bias in dataset construction. We note that the findings are applicable to a wide range of tasks in which users put labels on samples from a long list of candidate classes; our task is a special case of such a framework.

A similar MITL approach for expanding the positive matches was also employed by Parekh *et al*. [47], resulting in the dataset CrissCrossed Caption (CxC). However, CxC focuses on scoring the text-to-text similarities, resulting in many missing positives in the text-to-image relationship. Furthermore, CxC only employs one language-based machine annotator, which can lead to a biased dataset as our observation. Our ECCV Caption focuses on the inter-modality relationship and utilizes five ITM methods to avoid biased dataset construction. As another attempt to correct COCO benchmark, Chun *et al*. [15] annotate pseudo-positives by using the COCO instance classes, called Plausible Match (PM). For example, both images in Figure 1 contain the same object class, "tennis racket". Hence, the red and blue captions are considered positives for both red and blue images. Although PM items can detect most of the false negatives, it also introduces many false positives. Compared to PM [15] which relies on noisy proxies for correspondence, we correct the missing false negatives with "human ground truths" with the help of machine annotations. All in all, our dataset results in a higher recall than CxC and high precision than PM.

We not only fix FNs but also evaluation metrics. We argue that R@1 can overestimate the model performance by focusing only on the accuracy of the top-1 item rather than the rest of the items. Instead, we propose to use better ranking-based evaluation metrics, mAP@R [44]. Our human study shows that mAP@R is more aligned to humans than Recall@k. Now that the FNs are corrected in the evaluation sets and the evaluation metric is fixed, we re-examine the known ranking of 25 state-of-the-art VL models evaluated in the COCO Caption. We have observed that COCO 5K R@1 & R@5, and CxC R@1 are highly correlated (larger than 0.87 Kendall's rank correlation τ). On the other hand, we observe that the rankings across methods measured by mAP@R on ECCV Caption and COCO 1K R@1 are less well-correlated (τ=0.47). This confirms the observation by Musgrave *et al*. [44] and Chun *et al*. [15] on class-based datasets.

Our contributions are as follows. (1) We discover the false negative (FN) problem and quantify the exact number of wrong labels in MS-COCO. There exist $\times$3.6 positive image-to-caption associations and $\times$8.5 caption-to-image associa-

tions compared to the original MS-COCO. (2) We construct a corrected ITM test dataset, **ECCV Caption**, to avoid a wrong evaluation by FNs. We employ the machine-in-the-loop (MITL) annotation process to reduce the amount of human verification, resulting in saving 99.9% cost compared to the full exhaustive verification. ECCV Caption shares the same images and captions as the original MS-COCO; therefore, the existing methods can be evaluated on our dataset without additional training. We fix not only the annotations but also the evaluation metric. We propose to use mAP@R, a more human-aligned metric than R@1 for comparing model performances as shown in our human study. (3) We re-evaluate 25 state-of-the-art VL models on our ECCV Caption dataset based on mAP@R instead of Recall@k. In Table 4 and Figure 4, we can observe that focusing on MS-COCO R@1 will mislead the true ranking between the models (MS-COCO R@1 and ECCV mAP@R show a low correlation). Our observation aligns with Musgrave *et al*. [44] and Chun *et al*. [15]; focusing on R@1 can mislead the true rankings between models. (4) We provide a detailed analysis of the constructed dataset and the model bias. In particular, we focus on avoiding potential model biases in the proposed dataset by employing multiple models. Our analysis shows that our design choice is effective in solving the model bias.

2 Related Works

2.1 Noisy many-to-many correspondences of image-caption datasets

There have been a few attempts to introduce many-to-many or noisy correspondences for VL datasets. Parekh *et al*. [47] construct a CrissCrossed Caption (CxC) dataset by employing a similar MITL approach to ours. However, CxC focuses on intra-modality similarity, particularly text-to-text. They employed the Universal Sentence Encoder [8] and average bag-of-words (BoW) based on GloVe embeddings [48], while we directly focus on the inter-modality relationships and utilizes powerful ITM methods [15,32,39,50,59] to select candidates for validation by humans. CxC contains human ratings for 89,555 image-to-caption associations, among which 35,585 are positive, $\times$1.4 more positive relationships than 25,000 in COCO Caption. We show that the additional positives by CxC are precise, but their annotations still have many missing positives (*i.e*., high precision but low recall), resulting that R@1 on CxC perfectly preserves the rankings of VL models on COCO 5K R@1. On the other hand, our ECCV Caption has $\times$4.4 positives ($\times$3.6 image-to-caption correspondences and $\times$8.5 caption-to-image correspondences) compared to COCO Captions and roughly three times more positives compared to CxC. Furthermore, it is possible to measure mAP on our dataset due to the abundance of positive pairs, unlike for CxC.

Another attempt by Chun *et al*. [15] focused on precision rather than R@1 by annotating the pseudo-positives in a fully algorithmic approach. The authors defined "plausible matching (PM)" items that have the same instance classes with the query image (or the image corresponding to the query caption) to annotate pseudo-positives. For example, both images in Figure 1 contain the same instance class, "tennis racket", leading to the conclusion that the red and blue captions are

marked as positives for both red and blue images. More precisely, two instances are PM if $y_1, y_2 \in \{0,1\}^d$ differ at most ζ positions, where d is the number of instance classes (*e.g.*, for COCO, $d = 80$). Using the class-based pseudo-positives, Chun *et al.* propose Plausible-Match R-Precision (PMRP) metric, an R-Precision [44] metric based on the PM policy. The authors propose to use multiple ζ (*e.g.*, $\zeta \in \{0,1,2\}$) and report the average precision value. PM items can detect many missing false positives in the dataset, but we observe that most PM pseudo-positives are not actual positives (*i.e.*, high recall but low precision) — See Table 2. We also observe that PMRP shows a low correlation to other evaluation metrics; PMRP is a noisy metric compared to others.

2.2 Machine-in-the-loop (MITL) annotation

Humans and machines complement each other in the annotation process as they have different comparative advantages. Humans are the ultimate source of true labels, but they are slow and prone to errors and biases [27, 57, 60]. Machines are highly scalable, but their generalizability to unseen samples is limited. Machines are also prone to their own versions of errors and biases [43, 54]. MITL annotations have been designed to take the best of both worlds [3, 6, 55, 68].

Depending on the required trade-off between annotation quality and efficiency, one may opt for either single-turn or multi-turn annotation pipeline. The latter serves for the maximal demand for annotation quality: humans and machines alternate to correct and learn from each other's annotations [3, 55]. This is a widely used technique, the applications ranging from building a dictionary of cooking vocabularies [9], to supporting real-time screen-reading for blind people [23] and characterizing system failures [45]. Here, we focus on *single-turn MITL annotations* to focus on the atomic building block for MITL pipelines in general. There are two types of the single-turn paradigm: machine-verified human annotations [62, 67] or human-verified machine annotations. We focus on the latter, which are highly relevant for dealing with huge sources of data.

Under the human-verification framework, machines make label proposals for each image, focusing more on recall than precision [2, 36]. Previous crowdsourcing research in human-computer interaction (HCI) had mainly focused on the annotation interface and its effects on the annotation [16, 28, 58], or building a crowdsourcing workflow that leverages microtask pipelines [4, 31]. We investigate the side effects of the model choice in the MITL annotation paradigm where machines provide candidate label proposals.

3 ECCV Caption Dataset Construction

In this section, we describe ECCV Caption construction details. We annotate image-caption pairs in MS-COCO to solve the multiplicity of MS-COCO. However, the number of candidates is too huge for an exhaustive verification by humans: 76B for the whole dataset and 125M for the test split only. To reduce the amount of judgment by humans, we employ a single-turn machine-in-the-loop

Table 1: **Overview of the machine annotators.** Differences among five ITM models in terms of architectures and training objectives are shown. ViLT and CLIP are trained on a massive amount of aligned VL data, while other methods only use COCO Caption.

Model	Text backbone	Visual backbone	Objective function
PVSE [59]	Bi-GRU [14]	ResNet-152 [24]	Multiple instance learning
VSRN [39]	Bi-GRU	Faster R-CNN [52]	Semantic reasoning matching
PCME [15]	Bi-GRU	ResNet-152	Probabilistic matching
ViLT [32]	Vision Transformer (ViT-B/32) [19]		Vision-language pre-training
CLIP [50]	Transformer [51]	ViT-B/32	Contrastive learning

(MITL) annotation pipeline, containing three stages: (1) Filtering by machine annotators. (2) Judging the filtered relationships by MTurkers and additional verification by internal workers. (3) Post-processing and merging with CxC.

3.1 Model candidates for machine annotators

We choose five VL models with diverse properties to cover both diversity and practical relevance. The models use different text backbones (Bi-GRU [14], Transformer [61]), visual backbones (ResNet-152 [24], Faster R-CNN [52], ViT [19]), training objective functions, and training datasets as shown in Table 1. We use the officially released pre-trained weights by the authors. Specifically, we use the CutMix [70] pre-trained version for PCME to match the retrieval performances with others, and CLIP ViT-B/32, the largest model at the time of our data construction. We describe more details of each method in Appendix A.1.

We quantify the diversity of the models by measuring the differences in their retrieved items. We first retrieve the top 25 images for each model on the captions of the COCO Caption test split. We measure the similarities of the models in two different metrics. First, for every pair of models, we measure the Kendall rank correlation [30] between the two rankings of the retrieved items by the models. We observe that the models usually have low similarity ($\tau < 0.3$), except for PVSE and PCME. We additionally measure, for each pair of model i and j, the average ranking of model i's top-1 ranked item by model j. The top-1 items retrieved by the models are usually not included in the top-3 items by the others. These analyses show that the chosen models are diverse and the retrieved items do not correlate that much. The full results are shown in Appendix A.2.

3.2 Crowdsourcing on Amazon Mechanical Turk

We crowdsource image-caption matches on Amazon Mechanical Turk (MTurk) platform. For the sake of scalability, we subsample 1,333 caption queries and 1,261 image queries from the COCO Caption test split. Since the number of all possible matches is still prohibitive (40M), we employ the filtering strategy to reduce the number of candidates for human verification. We pre-select top-5

Table 2: **Precision and recall of the existing benchmarks measured by our human verified positive pairs.** A low Prec means that many positives are actually negatives, and a low Recall means that there exist many missing positives.

Dataset	I2T Prec	I2T Recall	T2I Prec	T2I Recall
Original MS-COCO Caption	47.3	20.0	89.4	12.8
CxC [47]	39.6	22.0	81.4	15.0
Plausible Match [15]	8.3	74.6	10.5	69.0

Table 3: **The number of positive images and captions for each dataset.** We show the number of positive items for the subset of the COCO Caption test split. The number of query captions and images are 1,332 and 1,261, respectively.

Dataset	# positive images	# positive captions
Original MS-COCO Caption	1,332	6,305 (=1,261×5)
CxC [47]	1,895 (×1.42)	8,906 (×1.41)
Human-verified positives	10,814 (×8.12)	16,990 (×2.69)
ECCV Caption	11,279 (×8.47)	22,550 (×3.58)

captions and images retrieved by the five models. After we remove the duplicate pairs from the (1,261 + 1,333) × 5 × 5 = 64,850 pairs, 46,424 pairs remain.

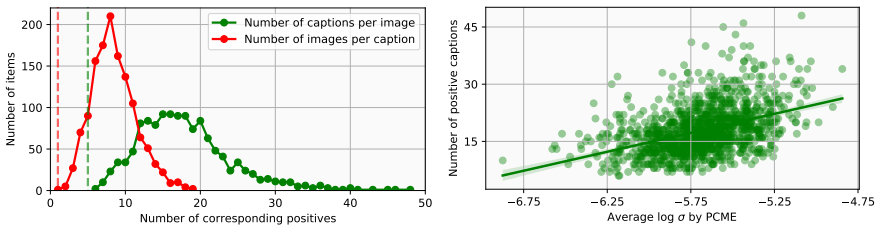
We package the task for human annotators into a series of Human Intelligence Tasks (HITs). Each HIT contains 18 machine-retrieved pairs, consisting of 1 true positive (*i.e.*, an *original positive pair*), 1 true negative (random pair, not in the top-25 of any model), and 16 pairs to be annotated. The golden examples are used for the qualification process; if a submitted HIT contains wrong answers to the golden examples, we manually verify the HIT. For each image-caption pair candidate, workers can choose an answer among the choices "100% YES", "Partially YES, but", "Mostly NO, because", and "100% NO". We use four choices instead three-level ("YES", "Not Sure", and "NO") to discourage workers from selecting "Not Sure" for all the questions. We have assigned 2,160 HITs, consisting of 43,200 pairs to be verified, to 970 MTurk workers. The crowdsourcing details, including an example HIT, compensation details, worker statistics, and detailed statistics for each machine annotator are in Appendix B.

3.3 Postprocessing MTurk annotations

We observe that 21,995 associations among 43,200 associations are annotated as positives ("Yes" or "Weak Yes"). We then filter out 18 meaningless captions (*e.g.*, "I am unable to see an image above"), 14 wrong captions found by workers (*e.g.*, "A group of birds flying above the beach" for the image with many kites), and 1 duplicate image found in the training set. The full list is in Appendix C.1.

Using the 21,995 human-verified positives, we report precision and recall of the existing benchmarks. Let t_i be the set of human-annotated positives for the

Fig. 2: **ECCV Caption examples**. The given caption query: "A herd of zebras standing together in the field". Red: original positive. Green: annotated as "100% Yes". Blue: annotated as "Weak Yes". More examples are in Appendix C.2.



(a) Number of positive pairs. (b) Multiplicity by positive items.

Fig. 3: **Multiplicity in ECCV Caption.** (a) The number of positive pairs in ECCV Caption. Dashed lines denote the number of the original COCO positives (1 image for each caption, and 5 captions for each image). ECCV Caption contains plenty of positive items per each modality. (b) PCME-predicted multiplicity against the number of positive captions for each image. There exists a positive correlation.

query i in Section 3.2 and r_i be the set of positives for i in the target dataset. We define precision and recall of a dataset as $Prec = \frac{1}{N}\sum_{i=1}^{N}\frac{|r_i \cap t_i|}{|r_i|}$ and $Recall = \frac{1}{N}\sum_{i=1}^{N} 1 - \frac{|t_i \setminus r_i|}{|t_i|}$. Table 2 shows precision and recall of COCO Caption, CxC [47], and Plausible Match (PM) pseudo-positives [15]. While COCO and CxC show high precisions, we observe that their recall is significantly low, around or less than 20%. Evaluating models on such a low-recall dataset with the R@1 metric can be highly misleading. A model may be able to retrieve good enough positive items which are not captured in the dataset, resulting in erroneously low R@1 scores. On the other hand, more than 70% of the positives can be captured by PM, but only about 10% of pseudo-positives are correct.

We consider the CxC positives as the additional sixth machine-human verified annotations, and extend our human-verified positives with CxC positives to construct the final ECCV Caption. Table 3 shows the detailed statistics of CxC, human-verified positives, and our ECCV Caption. Overall, ECCV Caption has $\times$8.47 positive images and $\times$3.58 positive captions than the original dataset. Figure 3a shows the number of positive images and captions per each item; there exist many positives beyond the original COCO associations. We illustrate example image-caption pairs from ECCV Caption in Figure 2 and Appendix C.2.

We additionally analyze the multiplicity of ECCV Caption by PCME [15] that produces a degree of multiplicity (uncertainty) for each query. Figure 3b

shows that more uncertain images correspond to more captions in our dataset. In other words, our new annotations capture the hidden FNs in COCO well.

4 Re-evaluation of ITM models on ECCV Caption

In this section, we re-evaluate the existing VL models on our new dataset and previous benchmarks. We first introduce the evaluation metrics and comparison methods (§4.1). We compare the performances and analyze the results (§4.2).

4.1 Evaluation metrics and comparison methods

Evalution metrics. The existing ITM benchmarks (*e.g.*, COCO Caption) use Recall@k metrics, particularly Recall@1 (R@1). Specifically, previous works measure R@1 for 5-fold validation splits (*i.e.*, each split has 1K images), and for the full test split [29]. The former is called COCO 1K R@k and the latter is called COCO 5K R@k, respectively. Previous studies separately report image-to-text, text-to-image retrieval R@1, R@5 and R@10 scores. However, as shown by Musgrave *et al.* [44], R@k is not an informative metric; embedding spaces with nearly 100% R@1 can have different properties. The problem becomes even worse for the ITM benchmarks, whose queries only have very few (usually only one) references: Even if a model correctly retrieves plausible items that are not among the set of original positives, the current benchmark cannot evaluate the model correctly. It is common to use larger values of k to less penalize wrong yet plausible predictions. However, as shown in Figure 3a, the actual number of plausible positives can be larger than the typical choice of k (*e.g.*, 5 or 10). Instead, we suggest using mAP@R [44], a modified mAP measured by retrieving R items where R is the number of positives for the query. Previous ITM benchmarks cannot employ mAP@R because R is too small (*i.e.*, 1). Thanks to our human-verified ground-truth positives, we can reliably measure mAP@R on ECCV Caption.

We additionally conduct a human study to confirm that mAP@R is more aligned to humans than R@k. We collect 3,200 pairwise preferences of human annotators among (A) only top-1 is wrong (B) only top-1 is correct (C) top-1 to 5 are wrong (D) only top-5 is correct, and (E) all items are wrong. For example, if the number of positives is 8, then (A) shows 0 R@1, 100 R@5 and 66.0 mAP@R, (B) shows 100 R@k and 12.5 mAP@R, (C) shows 0 R@k and 10.3 mAP@R, and (D) shows 0 R@1, 100 R@5 and 2.5 mAP@R. We compute user preference scores using Bradley–Terry model [7]. We observe that mAP@R is exactly aligned to the human preference score: (A: 70.85, B: 10.66, C: 13.15, D: 4.89, E: 0.44). We provide the details of the human study in Appendix D.1.

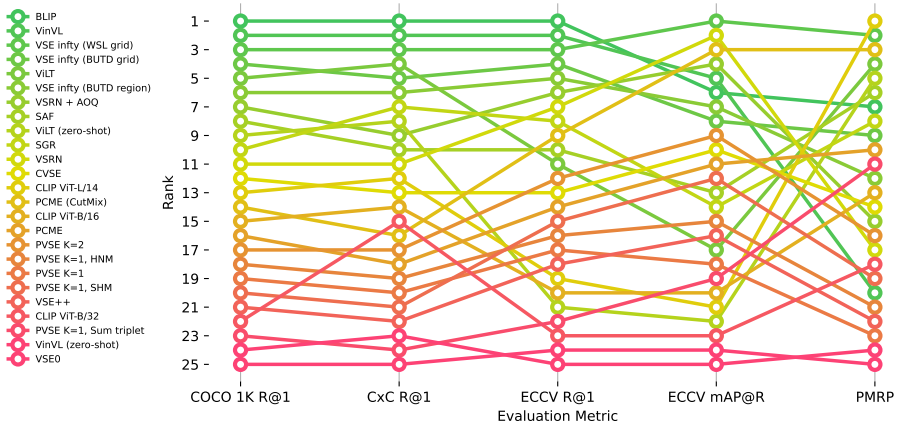
We also report modified Plausible Match R-Precision (PMRP) scores by changing R to $\min(R, 50)$, because the number of pseudo-positives R can be very large (*e.g.*, larger than 10,000) but most of them are not actual positive (Table 2). While Chun *et al.* [15] proposed to use the average R-Precision for three different thresholds, (*e.g.*, $\zeta = \{0, 1, 2\}$), we only report PMRP when $\zeta = 0$. We additionally compute R@1, R@5, and PMRP scores on the original COCO

Table 4: **Re-evaluating VL models.** ECCV Caption mAP@R, R-Precision (R-P), Recall@1 (R@1), CxC R@1, COCO 1K R@1, 5K R@1, and PMRP are shown. The numbers are the average between the image-to-text retrieval and text-to-image retrieval results. Full numbers for each modality and COCO R@5, R@10 results are in Appendix D.3. † denotes our re-implemention and "zero-shot" for VinVL and ViLT denotes VL pre-trained models without fine-tuning on the COCO Caption for the retrieval task.

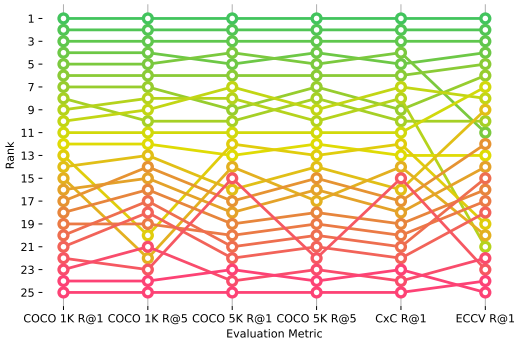
	ECCV Caption			CxC	COCO		
	mAP@R	R-P	R@1	R@1	1K R@1	5K R@1	PMRP
ResNet-152 [24] image encoder + Bi-GRU [14] text encoder							
VSE0† [20]	22.67	33.27	55.55	24.24	34.14	22.27	46.95
VSE++† [20]	35.01	45.50	73.11	37.95	48.46	35.79	54.26
PVSE K=1 [59]	33.98	44.49	73.25	38.38	48.67	36.20	53.56
PVSE K=2 [59]	40.26	49.92	76.74	40.18	50.29	38.13	55.52
PCME [15]	37.11	47.82	74.79	40.09	50.29	38.03	56.71
PCME (CutMix [70] pre-trained)† [15]	**41.74**	**51.45**	**78.67**	**41.70**	**51.35**	**39.51**	**57.65**
Region features based on Bottom-up Attention [1] and SCAN [37]							
VSRN [39]	**42.28**	**51.84**	81.51	48.85	58.33	46.74	55.44
VSRN + AOQ [12]	40.94	50.65	81.53	50.10	59.32	48.14	56.41
CVSE [64]	37.35	47.51	76.70	45.82	55.37	43.80	56.49
SGR [18]	35.80	46.04	78.77	50.60	58.87	48.86	56.91
SAF [18]	35.96	46.19	78.36	49.58	59.09	47.80	57.21
VSE infty (BUTD region) [11]	40.46	49.97	82.52	52.40	61.03	50.38	56.64
VSE infty (BUTD grid) [11]	40.40	50.09	83.01	53.47	62.26	51.60	56.87
VSE infty (WSL grid) [11]	42.41	51.43	**86.44**	**60.79**	**68.07**	**59.01**	**57.65**
Large-scale Vision-Language pre-training							
CLIP ViT-B/32 [50]	26.75	36.91	67.08	41.97	49.84	40.28	55.32
CLIP ViT-B/16 [50]	29.25	38.99	71.05	44.26	52.32	42.69	56.58
CLIP ViT-L/14 [50]	27.98	37.80	72.17	48.14	55.38	46.44	**57.70**
VinVL (zero-shot) [71]	22.18	32.93	55.19	33.74	43.51	32.07	47.26
VinVL [71]	**40.81**	**49.55**	87.77	67.76	**82.38**	66.39	54.72
ViLT (zero-shot) [32]	26.84	36.81	69.00	50.35	58.83	48.63	57.38
ViLT [32]	34.58	44.27	77.81	53.72	61.81	52.18	57.63
BLIP [38]	40.52	48.43	**90.99**	**74.30**	78.30	**73.11**	57.17
Different negative mining (NM) strategies							
PVSE K=1, No NM†	33.34	44.44	67.99	32.69	43.28	30.65	**56.67**
PVSE K=1, Semi-hard NM† [53]	**36.63**	**47.36**	**73.97**	38.17	48.49	36.00	55.15
PVSE K=1, Hardest NM† [20]	35.76	46.50	73.68	**39.02**	**49.12**	**36.88**	54.37

Caption, R@1 on CxC, and R@1 and R-Precision on ECCV Caption to analyze the correlation between each evaluation metric to ECCV mAP@R.

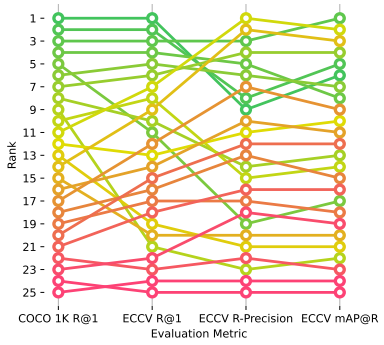
Evaluated methods. We compare 25 state-of-the-art VL models, whose trained weights are publicly accessible, categorized into four groups: (1) visual semantic embedding (VSE) methods with the ResNet-152 [24] image encoder, and Bi-GRU [14] text encoder, including VSE0, VSE++ [20], PVSE [59] (K=1 & K=2), and PCME [15] (the official model and the CutMix pre-trained version); (2) VSE methods with region features extracted by Visual Genome [35] pre-trained Faster R-CNN [52] based on the implementation by Anderson *et al.* [1] and Lee *et al.* [37], including VSRN [39], VSRN + AOQ [12], CVSE [64], SGR, SAF [18],



(a) Comparison of COCO, CxC, ECCV and PMRP.



(b) Comparison of Recall@1 metrics.



(c) Comparison of ECCV metrics.

Fig. 4: **Ranking correlation between different evaluation metrics.** Ranking of methods is largely perserved between COCO and CxC Recall@1, while it is rarely preserved among COCO Recall@1, ECCV mAP@R and PMRP.

and VSE∞ with BUTD region, grid and WSL grid features [11][2]. (3) Large-scale VL pre-training (VLP) methods, including pre-trained CLIP with ViT-B/32, ViT-B/16, and ViT/L14 backbones [50], pre-trained and fine-tuned ViLT [32], pre-trained and fine-tuned VinVL [71], and fine-tuned BLIP [38]. Here, "pre-trained" signifies that the model is trained with a massive image-text aligned dataset, but is not specifically trained for COCO Caption; "fine-tuned" signifies that the model is fine-tuned on COCO Caption for the ITM task. We note that VL transformers except CLIP need $O(|C| \times |I|)$ forward operations to compute the full pairwise ranks between $|C|$ number of captions and $|I|$ number of images,

[2] Techinally speaking, VSE∞ (WSL grid) does not use region features, but CNN features extracted from Instagram-trained ResNext [42]. This study treats all VSE∞ variants as region feature-based models for convenience.

Table 5: **Rank correlations between evaluation metrics.** Higher τ denotes two rankings are highly correlated, while τ values near zero denotes two rankings are barely correlated. We highlight the highly correlated pairs ($\tau > 0.8$) with **red** text.

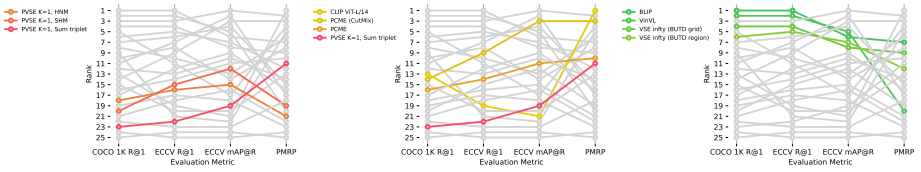
	COCO 1K			COCO 5K			CxC	ECCV			COCO
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@1	R-P	mAP@R	PMRP
COCO 1K R@1	-	**0.87**	**0.86**	**0.89**	**0.97**	**0.92**	**0.89**	0.72	0.39	0.47	0.45
COCO 1K R@5	**0.87**	-	**0.97**	0.79	**0.88**	**0.93**	0.79	**0.81**	0.49	0.58	0.39
COCO 1K R@10	**0.86**	**0.97**	-	0.77	**0.86**	**0.91**	0.77	0.79	0.49	0.57	0.43
COCO 5K R@1	**0.89**	0.79	0.77	-	**0.89**	**0.83**	**1.00**	0.65	0.30	0.39	0.45
COCO 5K R@5	**0.97**	**0.88**	**0.86**	**0.89**	-	**0.95**	**0.89**	0.75	0.41	0.50	0.43
COCO 5K R@10	**0.92**	**0.93**	**0.91**	**0.83**	**0.95**	-	**0.83**	**0.80**	0.47	0.55	0.38
CxC R@1	**0.89**	0.79	0.77	**1.00**	**0.89**	**0.83**	-	0.65	0.30	0.39	0.45
ECCV R@1	0.72	**0.81**	0.79	0.65	0.75	**0.80**	0.65	-	0.65	0.74	0.29
ECCV R-P	0.39	0.49	0.49	0.30	0.41	0.47	0.30	0.65	-	**0.90**	0.17
ECCV mAP@R	0.47	0.58	0.57	0.39	0.50	0.55	0.39	0.74	**0.90**	-	0.20
PMRP	0.45	0.39	0.43	0.45	0.43	0.38	0.45	0.29	0.17	0.20	-

while other methods only need $O(|I|) + O(|C|)$ forward operations to compute the full pairwise ranks based on the cosine similarity. For example, VinVL takes 25 hours to compute the full pairwise ranks for the COCO Caption test split by a single A100 GPU core, while VSE++ only takes 1 minute in the same environment. (4) PVSE models with different negative mining (NM) methods, including no NM, semi-hard NM (SHM) [53], and hardest NM (HNM) [20].

We use the official trained weights for each model with a few exceptions. We re-implement VSE0, VSE++, PCME with CutMix pre-trained ResNet, and PVSE models with various NM strategies. The training details are in Appendix D.2

4.2 Re-evaluation of ITM methods

Table 4 and Figure 4 shows the full comparisons of 25 VL models with different evaluation metrics. We report the Kendall's rank correlations (tau-b) between metrics in Table 5; larger τ denotes two metrics are more correlated. We report the full table including modality-wise results, R@5 and R@10 scores in Appendix D.3. We first observe that R@k scores across different datasets have high correlations among themselves (Figure 4b and Appendix D.3). In terms of the ranking correlation, we observe that COCO 1K R@1 shows almost τ=0.9 with the ranking yielded by R@5 (0.87), COCO 5K R@1 (0.89) and R@5 (0.97), or CxC R@1 (0.89). This implies that measuring Recall@k on different benchmarks, such as the original COCO Caption, CxC, and ECCV Caption are not more informative than only measuring Recall@k on COCO 1K or 5K. On the other hand, the rankings by COCO 1K are not preserved well to PMRP (0.45), ECCV R@1 (0.72), ECCV R-Precision (0.39) and ECCV mAP@R (0.47) in Kendall's τ. This implies that enlarging K of R@k (*e.g.*, using R@5, R@10 instead of R@1) cannot be an alternative of mAP@R because R@k metrics are highly correlated each other as shown in Table 5. We also observe that the rankings by PMRP are relatively less correlated to the other metrics, such as COCO R@1 (0.45), ECCV R@1 (0.29) or ECCV mAP@R (0.20) in Kendall's τ.



(a) Triplet mining strategies. (b) Contrastive methods. (c) Best R@1 models.

Fig. 5: **Rankings of different VL models.** Ranking of (a) PVSE models with diverse triplet mining strategies (b) contrastive methods (c) the best models are shown.

Our re-evaluation shows that existing ITM evaluation benchmarks can overestimate the VL model performance by focusing only on COCO R@1, where the rankings between COCO R@1 and ECCV mAP@R are not largely preserved. For example, we observe that the hardest negative mining technique [20], previously deemed useful for ITM tasks, is actually selectively effective for R@1, rather than for the actual task itself. Under our new metrics like ECCV mAP@R, we observe that the milder strategy of semi-hard negative mining is more effective – See Figure 5a. Chun *et al*. [15] also observed a similar pattern in the CUB Caption dataset [63] by using the class-defined positives. Our finding is the first observation in the practical large-scale VL dataset. Similarly, we observe that many large-scale VL pre-training methods with high R@1 scores show inferior ECCV mAP@R scores compared to other visual semantic embedding techniques. For example, CLIP ViT-L/14 shows superior COCO 1K R@1 than PCME (55.4% and 40.1%, respectively). However, in terms of ECCV mAP@R, CLIP shows inferior performances than PCME (28.0% and 37.1%, respectively).

Similarly, we observe that PMRP shows different behaviors compared to other metrics. Especially, we observe that the contrastive models without a negative mining strategy are specialized to PMRP metric – Figure 5b. We presume that it is because the contrastive learning strategy enforces the features with similar objects to be mapped to a similar embedding space. In contrastive the best models on COCO and ECCV (*e.g.*, BLIP, VinVL, and VSE∞) show inferior PMRP scores – Figure 5c. We presume that it is because PMRP only captures the existence or absence of the objects, while an optimal retrieval also should consider the plausibility between matched image-caption pairs.

5 Discussion and Limitations

Potential machine biases in our dataset. Our dataset construction process contains the MITL annotation process, where the choice of machine annotators can potentially harm the dataset quality. The positives in our dataset are the retrieved items by the machine annotators. If the machines are biased towards undesired patterns (*e.g.* favoring certain items over the others), future methods built on our benchmark will overfit those patterns. In this work, we employ five

diverse machine annotators to reduce the potential biases by models. In Appendix E, we explore and quantify the effect of the choice of multiple machine annotators on the dataset quality. From the study, we can conclude that our strategy (using more models) is effective to mitigate biases by a specific model.

Scale of ECCV Caption. In this work, we subsample 1,333 caption queries (5.3% of the full caption queries) and 1,261 image queries (25.2% of the full image queries) to reduce the scale of annotations. Note that without subsampling, we need to verify $(25{,}000 + 5{,}000) \times 5 \times 5 = 750$K pairs, which costs 16 times more than our current version, almost \$60K. Because we only subsample queries, not limiting the gallery samples, our dataset is an unbiased subset of the original COCO Caption. To scale up ECCV Caption, we have to reduce the human verification costs by reducing the total number of human verification. This can be achievable by applying a multi-turn MITL annotation process that alternatively repeats training machine annotators with human-annotated associations and verifying machine annotations by human workers. After enough iterations of the multi-turn MITL annotation process, we can automatically scale up our annotations by using the high-quality machine annotators while only low confident associations are verified by humans.

Noisy annotations. Despite our additional verification process to keep the quality of the annotations, there can be noisy annotations (*i.e.*, false positives) in ECCV Caption due to the noisy nature of crowdsourcing annotations. The noisy annotations can also occur because we use both "100% YES" and "Partially YES" to build positive pairs. However, we still encourage to use ECCV Caption for evaluating VL models, because the existing datasets are noisier; they usually have only one positive item per each query and they have tremendously many FNs. On the other hand, noisy annotations of our dataset are still "plausible" rather than "wrong". We provide more discussion in Appendix F. Finally, we expect that a multi-turn MITL process can improve not only the labeling cost but also the annotation quality as shown by Benenson *et al.* [3].

6 Conclusion

MS-COCO Caption is a popular dataset for evaluating image-text matching (ITM) methods. Despite its popularity, it suffers from a large number of missing positive matches between images and captions. Fully annotating the missing positives with human labor incurs prohibitive costs. We thus rely on machine annotators to propose candidate positive matches and let crowdsourced human annotators verify the matches. The resulting ITM evaluation benchmark, Extended COCO Validation (ECCV) Caption dataset, contains $\times 8.47$ positive images and $\times 3.58$ positive captions compared to the original MS-COCO Caption. We have re-evaluated 25 ITM methods on ECCV Caption with mAP@R, resulting in certain changes in the ranking of methods. We encourage future studies on ITM to evaluate their models on ECCV mAP@R that not only focuses on the correctness but also on the diversity of top-k retrieved items.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proc. CVPR (2018) 10
2. Andriluka, M., Uijlings, J.R., Ferrari, V.: Fluid annotation: a human-machine collaboration interface for full image annotation. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1957–1966 (2018) 5
3. Benenson, R., Popov, S., Ferrari, V.: Large-scale interactive object segmentation with human annotators. In: Proc. CVPR. pp. 11700–11709 (2019) 5, 14
4. Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., Panovich, K.: Soylent: a word processor with a crowd inside. In: Proceedings of the 23nd annual ACM symposium on User interface software and technology. pp. 313–322 (2010) 5
5. Biten, A.F., Mafla, A., Gómez, L., Karatzas, D.: Is an image worth five sentences? a new look into semantics for image-text matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1391–1400 (2022) 1
6. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In: Proc. ICCV. vol. 1, pp. 105–112. IEEE (2001) 5
7. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952) 9
8. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al.: Universal sentence encoder. In: Proc. EMNLP (2018) 4
9. Chang, M., Guillain, L.V., Jung, H., Hare, V.M., Kim, J., Agrawala, M.: Recipescape: An interactive tool for analyzing cooking instructions at scale. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2018) 5
10. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proc. CVPR. pp. 3558–3568 (2021) 1
11. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: Proc. CVPR (2021) 1, 10, 11
12. Chen, T., Deng, J., Luo, J.: Adaptive offline quintuplet loss for image-text matching. In: Proc. ECCV (2020) 1, 10
13. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 1
14. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014) 6, 10
15. Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: Proc. CVPR (2021) 1, 2, 3, 4, 6, 7, 8, 9, 10, 13
16. Chung, J.J.Y., Song, J.Y., Kutty, S., Hong, S., Kim, J., Lasecki, W.S.: Efficient elicitation approaches to estimate collective crowd answers. Proceedings of the ACM on Human-Computer Interaction **3**, 1 – 25 (2019) 5
17. Desai, K., Kaul, G., Aysola, Z., Johnson, J.: RedCaps: Web-curated image-text data created by the people, for the people. In: NeurIPS Datasets and Benchmarks (2021) 1

18. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proc. AAAI (2021) 1, 10
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. ICLR (2021), https://openreview.net/forum?id=YicbFdNTTy 6
20. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving visual-semantic embeddings with hard negatives. In: Proc. BMVC (2018) 1, 10, 12, 13
21. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Proc. NeurIPS. pp. 2121–2129 (2013) 1
22. Gu, J., Cai, J., Joty, S.R., Niu, L., Wang, G.: Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7181–7189 (2018) 1
23. Guo, A., Chen, X., Qi, H., White, S., Ghosh, S., Asakawa, C., Bigham, J.P.: Vizlens: A robust and interactive screen reader for interfaces in the real world. Proceedings of the 29th Annual Symposium on User Interface Software and Technology (2016) 5
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR (2016) 6, 10
25. Huang, Y., Wu, Q., Song, C., Wang, L.: Learning semantic concepts and order for image and sentence matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6163–6171 (2018) 1
26. Huang, Z., Niu, G., Liu, X., Ding, W., Xiao, X., hua wu, Peng, X.: Learning with noisy correspondence for cross-modal matching. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Proc. NeurIPS (2021), https://openreview.net/forum?id=S9ZyhWC17wJ 1
27. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on amazon mechanical turk. In: Proceedings of the ACM SIGKDD workshop on human computation. pp. 64–67 (2010) 5
28. Kaplan, T., Saito, S., Hara, K., Bigham, J.P.: Striving to earn more: A survey of work strategies and tool use among crowd workers. In: HCOMP (2018) 5
29. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proc. CVPR. pp. 3128–3137 (2015) 9
30. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938) 6
31. Kim, J., Nguyen, P., Weir, S.A., Guo, P.J., Miller, R., Gajos, K.Z.: Crowdsourcing step-by-step information extraction to enhance existing how-to videos. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2014) 5
32. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: Proc. ICML (2021) 2, 4, 6, 10, 11
33. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014) 1
34. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proc. CVPR Worshops. pp. 554–561 (2013) 2
35. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV **123**(1), 32–73 (2017) 10

36. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. IJCV **128**(7), 1956–1981 (2020) 5
37. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proc. ECCV (2018) 1, 10
38. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation (2022) 10, 11
39. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proc. ICCV. pp. 4654–4662 (2019) 1, 2, 4, 6, 10
40. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proc. ECCV (2014) 1
41. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proc. CVPR. pp. 1096–1104 (2016) 2
42. Mahajan, D.K., Girshick, R.B., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proc. ECCV (2018) 11
43. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (Jul 2021). https://doi.org/10.1145/3457607, https://doi.org/10.1145/3457607 5
44. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Proc. ECCV (2020) 2, 3, 4, 5, 9
45. Nushi, B., Kamar, E., Horvitz, E.: Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In: HCOMP (2018) 5
46. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proc. CVPR. pp. 4004–4012 (2016) 2
47. Parekh, Z., Baldridge, J., Cer, D., Waters, A., Yang, Y.: Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. arXiv preprint arXiv:2004.15020 (2020) 3, 4, 7, 8
48. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proc. EMNLP. pp. 1532–1543 (2014) 4
49. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015) 1
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proc. ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), http://proceedings.mlr.press/v139/radford21a.html 1, 2, 4, 6, 10, 11
51. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 6
52. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proc. NeurIPS. pp. 91–99 (2015) 6, 10
53. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proc. CVPR. pp. 815–823 (2015) 10, 12

54. Scimeca, L., Oh, S.J., Chun, S., Poli, M., Yun, S.: Which shortcut cues will dnns choose? a study from the parameter-space perspective. In: International Conference on Learning Representations (ICLR) (2022) 5
55. Settles, B.: Active learning literature survey (2009) 5
56. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL. pp. 2556–2565 (2018) 1
57. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.: Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. pp. 254–263. Association for Computational Linguistics, Honolulu, Hawaii (Oct 2008), https://aclanthology.org/D08-1027 5
58. Song, J.Y., Fok, R., Lundgard, A., Yang, F., Kim, J., Lasecki, W.S.: Two tools are better than one: Tool diversity as a means of improving aggregate crowd performance. 23rd International Conference on Intelligent User Interfaces (2018) 5
59. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: Proc. CVPR. pp. 1979–1988 (2019) 1, 2, 4, 6, 10
60. Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8 (2008). https://doi.org/10.1109/CVPRW.2008.4562953 5
61. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. NeurIPS. pp. 5998–6008 (2017) 6
62. Verma, Y., Jawahar, C.: Image annotation by propagating labels from semantic neighbourhoods. IJCV **121**(1), 126–148 (2017) 5
63. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011) 2, 13
64. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: Proc. ECCV (2020) 1, 10
65. Wehrmann, J., Souza, D.M., Lopes, M.A., Barros, R.C.: Language-agnostic visual-semantic embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5804–5813 (2019) 1
66. Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., Ma, W.Y.: Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6609–6618 (2019) 1
67. Wu, W., Yang, J.: Smartlabel: An object labeling tool using iterated harmonic energy minimization. In: Proceedings of the 14th ACM international conference on Multimedia. pp. 891–900 (2006) 5
68. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.S.: Deep interactive object selection. In: Proc. CVPR. pp. 373–381 (2016) 5
69. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. ACL **2**, 67–78 (2014) 1
70. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proc. ICCV (2019) 6, 10
71. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Making visual representations matter in vision-language models. In: Proc. CVPR (2021) 10, 11