Cross-Domain Image Captioning with Discriminative Finetuning

Roberto Dessì, Meta AI / UPF
Michele Bevilacqua, Samaya AI
Eleonora Gualdoni, UPF
Nathanaël Carraz Rakotonirina, UPF
Francesca Franzon, UPF
Marco Baroni, UPF / ICREA

Abstract

Neural captioners are typically trained to mimic human-generated references without optimizing for any specific communication goal, leading to problems such as the generation of vague captions. In this paper, we show that fine-tuning an out-of-the-box neural captioner with a self-supervised discriminative communication objective helps to recover a plain, visually descriptive language that is more informative about image contents. Given a target image, the system must learn to produce a description that enables an out-of-the-box text-conditioned image retriever to identify such image among a set of candidates. We experiment with the popular ClipCap captioner, also replicating the main results with BLIP. In terms of similarity to ground-truth human descriptions, the captions emerging from discriminative finetuning lag slightly behind those generated by the non-finetuned model, when the latter is trained and tested on the same caption dataset. However, when the model is used without further tuning to generate captions for out-of-domain datasets, our discriminatively-finetuned captioner generates descriptions that resemble human references more than those produced by the same captioner wihtout finetuning. We further show that, on the Conceptual Captions dataset, discriminatively finetuned captions are more helpful than either vanilla ClipCap captions or ground-truth captions for human annotators tasked with an image discrimination task.[1]

1. Introduction

The last decade has seen impressive progress on the task of automatically generating image descriptions with deep neural networks [5, 39, 44, 46]. Most of the proposed methods try to optimize the similarity of system-produced captions with ground-truth human references, either through a standard cross-entropy cost function [3], or by maximizing natural-language-generation (NLG) metrics such as CIDEr [31, 36] through a reward-based objective. While imitating human captions is a reasonable goal, it does not take into account that, in concrete applications, an image description is produced for a purpose [15, 21].

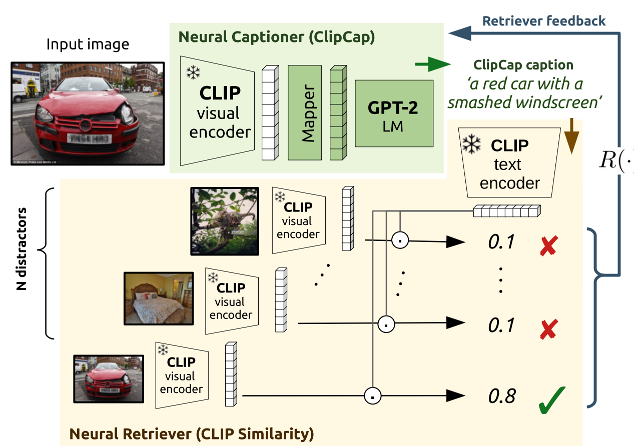

Figure 1. Setup of our discriminative finetuning method when applied to the ClipCap captioner [27]. All CLIP encoders are frozen, while the language generation modules (mapper and GPT-2) are updated based on reward values.

There are a multitude of context-dependent purposes a description might be produced for, but a fundamental one is to correctly characterize an object so that a hearer could *discriminate* it from other contextual elements [18]. This ability to discriminate between referents is a core purpose of communication, playing a fundamental role in its evolution and acquisition (e.g., [6, 38]). We study here what happens when we take an out-of-the-box image captioner that was trained to imitate human captions, and finetune its language components with a *discriminative* objective using

[1]Our code is available at https://github.com/facebookresearch/EGG/tree/main/egg/zoo/discriminative_captioner.

reinforcement learning. In particular, we let the captioner play a discrimination game with an out-of-the-box caption-based image retriever. The captioner generates a caption given a target image, and the retriever (whose weights are not updated) uses the caption to select the target among a set of candidates, as shown in Fig. 1. This finetuning technique does not require annotated data (only a set of images), and it's agnostic to the underlying captioner and retriever components.

We report two strong and novel results. First, we show that captions finetuned in this way lead to better 0-shot cross-domain caption generation.[2] Second, not only are the finetuned captions good for neural text-based image retrieval (both in- and across-domain), but they can also be more useful to human annotators, helping discriminate the target from distractors *more than human-generated ground-truth captions do*. We conclude the paper with an analysis of the finetuned captions, comparing them to human-generated and non-finetuned ones from the Conceptual Captions dataset. We find that discriminative finetuning undoes the more abstract language that the underlying system had learned from the ground-truth captions, leading to a more plainly descriptive style that we expect to be more useful in practical applications.

2. Related Work

In recent years, deep learning has led most progress in image captioning [3, 12]. While early approaches relied on supervised learning to maximize the likelihood of the model against human references [3, 44], other methods have tried to maximize a reward based on standard language generation metrics [34, 35]. Our work belongs to a third tradition, exploring the idea of caption learning or finetuning with a reward-based objective that is not (only) based on a comparison with reference captions. Among the approaches more closely related to ours, the system of Yu *et al.* [49] uses a ClipCap-like system for caption generation, and CLIP to measure image-caption similarity, focusing on generating captions in multiple styles. Cho *et al.* [8] use CLIP to finetune a pre-trained captioner. Like [49], they use the CLIP-Score image-caption similarity measure [16] as a reward signal, rather than a discriminative objective like the one we adopt. We show in Appendix B.2 that the discriminative objective outperforms an image-caption similarity objective similar to CLIPScore (*cosine similarity* in Table 9). Luo *et al.* [26] apply discriminative finetuning to basic MLE-trained models. Their method requires the use of ground-truth captions for concurrent CIDEr-based optimization. A similar approach was presented by Liu *et al.* [25]. They also finetune a MLE-pre-trained captioner with a mixture of discriminative and CIDEr-based rewards (requiring ground-truth captions for the latter). Intriguingly, their method produces a higher portion of novel and unique captions compared to a system trained only to imitate human captions, pointing to the potential for generalization that we confirm here. Finally, in another early paper using discrimination-based caption learning, Dai and Lin [9] propose a loss that explicitly pushes a model to make captions more discriminative with respect to those of a reference model.

Our study builds on this earlier work. Our main novelties lie in the development of a simple system to perform discriminative finetuning that only requires unannotated images, out-of-the-box generation/retrieval components and applying the vanilla REINFORCE algorithm, and in our out-of-domain and human-based evaluations.

3. Discriminative Self-Supervised Training

We finetune with reinforcement learning a pre-trained image captioner to generate a caption that is subsequently fed as input to a frozen discriminator. The discriminator is then tasked to retrieve the original image among a set of distractors. We refer to our tuning method as *DiscriTune*.

Captioner We employ a multimodal language model as neural captioner. We experiment with two different captioners, namely the ClipCap architecture from Mokady *et al*. [27], and the captioner network of the BLIP model [23].[3]

ClipCap is an image-conditioned publicly available GPT-2 model [33]. Importantly, Mokady *et al*. [27] showed that it performs comparably to other state-of-the-art captioning systems while having less parameters and being more efficient to train. Given an input image, ClipCap uses a frozen CLIP [32] visual encoder to extract visual features. These features are then projected through a trainable mapper network onto the GPT-2 embedding space, where they are used as a prefix conditioning the generation of an image description. In practice, they act as soft image prompts that are used to kickstart the caption generation process.

We use the publicly available ClipCap checkpoints where the language model weights were also updated during training. Such checkpoints were trained on Conceptual Captions [37] (*ClipCap-ConCap* henceforth), and MS COCO [24] (*ClipCap-COCO* henceforth), respectively. Given that these models were trained with a frozen ViT-B/32-based CLIP as visual feature extractor, we use the same visual encoder in our experiments.

To test whether our method generalizes to other models, we experiment with the BLIP system [23], a large captioner pre-trained with web-mined data which performed on par or

[2] We note that in the earlier literature, cross-domain captioning can rely on sets of unpaired images and captions from the target domain (e.g., [7, 47, 50]). We consider here a more challenging and realistic 0-shot transfer task in which *no* captions from the target domain are available.

[3] Additional results with the recently introduced CaMEL [4] captioner are provided in Appendix C.

better than other state-of-the-art systems on several benchmarks and has shown strong retrieval and captioning performance [23]. BLIP is made of a text Transformer [42] and a vision Transformer [13]. The text Transformer is trained to maximize the likelihood of reproducing a ground-truth reference by autoregressively generating a caption. Visual information is injected by cross-attending over the output of the vision Transformer. The model is trained with multi-task learning with two other multimodal alignment losses. We refer the reader to [23] for additional details on BLIP. We use the BLIP-base version,[4] pre-trained on a large dataset including both COCO and Conceptual Captions.

Retriever We used the standard CLIP model from Radford *et al*. [32] as our neural retriever. CLIP is a multimodal dual encoder model that embeds text and images and learns to maximize their similarity through a contrastive loss. It was trained on a dataset of 400M human-curated image-text pairs. We refer the reader to [32] for additional details about CLIP pre-training. We use the publicly available original implementation of CLIP with a ViT-B/32 backbone [13]. [5]

We perform retrieval by computing a matching score $\mathrm{match}(c, i)$ for an image i and a caption c as the dot product between the embedded representation of c (computed by the frozen CLIP textual encoder) and the embedded representation of i (computed by the frozen CLIP visual encoder). We consider an image i correctly retrieved against a set of distractors $D \subseteq I$ (taken from a larger image collection I) iff $i' \in D$, and $i \neq i'$, $\mathrm{match}(c, i) > \mathrm{match}(c, i')$.

Optimization Since the decoding process creates a discrete bottleneck, we cannot use a loss function to back-propagate end-to-end from the retriever output. Thus, we use REINFORCE [45] to optimize the captioner (ClipCap or BLIP) by using a reward that is based on the matching scores between the caption c and the set of candidates $D \cup \{i\}$. In this process, the CLIP retriever is kept frozen and, given the non-differentiable text generation step, there is no gradient flowing from the retriever to the captioner.

To compute a reward, we normalize the retriever scores to a probability distribution, and then use as reward the so-defined log probability of the original image i: $R(c, i, D) = \log \frac{e^{\mathrm{match}(c,i)}}{\sum_{i' \in D \cup \{i\}} e^{\mathrm{match}(c,i')}}$. We then compute the reward as the cross-entropy loss between such distribution and the position of the target image in the list of candidates. Our captioner, which, in essence, models the probability $P_\theta(c|i)$ of a caption c given an image i, is, in reinforcement learning terms, the action policy, and the actions taken are just the selected tokens. The policy is trained to minimize the negative expected reward, i.e., $\mathbb{E}_{c \sim P_\theta(\cdot|i)}[-R(c, i, D)]$ and we compute the gradient of this expectation as the expectation of the gradient. To reduce variance of the gradient estimator we use as a baseline a running mean of past rewards [41] (see Appendix B for more details).

In our experiments we let both our captioners freely generate for a maximum of 40 tokens, or until a full stop (or EOS token for BLIP) is produced. We set such maximum length after observing that most captions were much shorter and taking into account the fact that CLIP, our downstream retriever, does not process contexts larger than 75 tokens.

4. Experiments

4.1. Setup

In this section, we report experiments with our DiscriTune method where we finetune the pre-trained captioner with a self-supervised discriminative reward provided by a frozen downstream CLIP receiver. We apply the method to the ClipCap checkpoints introduced above, namely ClipCap-COCO and ClipCap-ConCap. We finetune them with CLIP-provided discriminative reward using COCO and Conceptual Captions data, respectively. We call the resulting finetuned models DiscriTune-COCO and DiscriTune-ConCap. We also repeat the main experiments using the BLIP checkpoint trained on COCO, which we further finetune on COCO with our DiscriTune method.[6]

We finetune both our captioners with the discriminative reward using the Adam [20] optimizer with a learning rate of 10^{-7} and a constant schedule. We use a batch size of 100, with in-batch distractors, so that each target image is mixed with 99 randomly sampled distractors for discriminative finetuning. All our experiments were performed on a single Tesla V100 GPU. For ClipCap on COCO, we finetune for 20 epochs, with around 1.2K batches per epochs and roughly 24K updates, whereas on Conceptual Captions we finetune for 2 epochs leading to around 28K batches per epoch and 56K updates. For BLIP, we finetune on COCO for a single epoch corresponding to 1.2K updates, as preliminary experiments showed that to be sufficient for convergence. We use greedy decoding at training time and beam search at test time.

Hyperparameter search We used the Flickr validation set [48], not used elsewhere in the paper, to tune learning rate, reward function (discriminative- vs accuracy-based), and REINFORCE baseline (see Appendix B).

Data We used the standard **MS COCO** [24] and **Conceptual Captions** [37] datasets to perform discriminative finetuning. MS COCO is on of the most commonly used

[4]The model is publicly available through the LAVIS library [22]

[5]https://github.com/openai/CLIP

[6]We use the EGG library [19] to perform all our experiments.

captioning and text-conditioned retrieval datasets, containing around 120K images, each provided with 5 human-generated captions. We use the Karpathy train and test set split [17]. Conceptual Captions is a collection of images mined from the web and aligned with their alt-text descriptions. It contains around 3M samples for training and around 16K images for validation, that we used as our test set. After filtering out corrupted images due to outdated download links, we are left with 2.8M samples in the train set and exactly 13K images for testing.

To test 0-shot cross-domain generalization, we use the **Flickr** [48], **nocaps** [1] and **Concadia** [21] datasets. We divided Flickr data according to the Karpathy split and evaluate on the test set (1K samples), where each image is aligned with 5 human captions. The nocaps dataset was introduced to test caption generalization performance of models trained on COCO. It is divided into three splits. We do not use the control *in-domain* split, since it contains the same object classes as COCO. The *near-domain* split has images from COCO categories as well as images from new categories. The *out-domain* split only contains pictures of objects that are not present in COCO. Following prior work [27], we use the validation set of nocaps for testing purposes. Concadia is a dataset recently introduced to test difference in text generation performance when producing captions compared to descriptions. The former are meant to accompany an image in order to provide additional context, as in books or newspapers, while the latter should be able to replace the image, an example being descriptions for visually impaired people.[7] We use the Concadia test split, which contains 9.6K images, each annotated with a caption and a description.

We want to emphasize again that, when finetuning our models, we did not use any human reference. Ground-truth captions were only used to compute NLG metrics.

4.2. Results

4.2.1 Text-Conditioned Image Retrieval

Table 1 shows that, as expected, ClipCap trained with DiscriTune greatly improves over vanilla ClipCap on the image retrieval tasks it was finetuned on. The result however also extends to cross-domain retrieval. DiscriTune-COCO (slightly) outperforms ClipCap-ConCap on Conceptual Captions, and DiscriTune-ConCap greatly outperforms ClipCap-COCO on COCO. Moreover, both versions of DiscriTune greatly outperform ClipCap on all other datasets. Interestingly, DiscriTune almost always outperforms human-generated captions, confirming a recent result by Dessì *et al*. [11] on how neural retrievers show better performance with neural captions. We performed additional experiments on a more challenging setup where hard distractors are either automatically mined or selected from adjacent frames in videos and confirm the superior performance of our DiscriTune method (Appendix D).

[7] Note that in this paper we follow instead the standard practice of using "caption" to refer to both captions and descriptions in the sense of Kreiss *et al*. [21].

4.2.2 Ground-Truth-Based Caption Quality Evaluation

The text-based caption retrieval results show that our approach is remarkably good at the task it was finetuned on, also in the 0-shot cross-domain setup. However, it is less clear that retrieval-based finetuning should improve the captions' faithfulness to human ground-truth image descriptions. Table 2 shows that, indeed, discriminative finetuning leads to some decrease in caption faithfulness, when testing on the dataset used for supervised pre-training. However, this performance drop (which, in the case of COCO is very small) is balanced by greater generalization performance, as shown in Table 3. DiscriTune-COCO is the best domain transfer model across-the-board. In addition, DiscriTune-ConCap is also consistently outperforming its vanilla counterpart, ClipCap-ConCap, in this 0-shot cross-domain generalization setup. Interestingly, for the Concadia dataset we have a (slightly) higher gain over plain ClipCap with the descriptions split rather than with the captions one, especially for DiscriTune-ConCap. Human references in the descriptions split were generated with the goal of replacing an image by describing its contents. This confirms that DiscriTune contributes to captions that are more suited for the "communicative" purpose of characterizing the crucial aspects of an image contents. Overall, the results suggest that, on the one hand, discriminative finetuning leads to captions that drift apart somewhat from the human descriptions that the model learnt to mimic. At the same time, though, this might allow the model to get away from the idiosyncrasies of a specific captioning style, leading to captions that better generalize to a wider set of unseen images and domains. As the experiment in Section 4.2.4 below shows, sometimes this drifting might be beneficial even when captioning images in the same domain.

4.2.3 Applying Discriminative Finetuning to BLIP

To verify the robustness of DiscriTune finetuning, we next apply it to BLIP. We finetune BLIP on COCO captions, one of the datasets it was pre-trained on. Tables 4 (retrieval) and 5 (captioning) show that Discrirune-BLIP outperforms its non-finetuned counterpart for both retrieval performance and caption quality. The results confirm that our captioner-agnostic discriminative finetuning is helpful even when applied to this latest-generation general-purpose vision-and-language model. DiscriTune-BLIP outperforms its vanilla counterpart on all out-of-domain datasets tested. Again, we note the greater performance boost on the Concadia Descriptions split compared to the Captions one, confirming

Model	*COCO*	*ConCap*	*Flickr*	*nocaps near*	*nocaps out*	*Concadia*
ClipCap-COCO	74.2	73.0	65.9	77.3	73.9	53.74
DiscriTune-COCO	**84.8**	83.6	79.4	86.0	82.5	64.79
ClipCap-ConCap	73.4	82.5	76.7	78.1	73.6	59.17
DiscriTune-ConCap	81.6	**94.4**	87.8	**89.1**	**88.7**	**80.49**
Human captions	76.3	81.6	**88.7**	85.5	87.7	73.96

Table 1. ClipCap and DiscriTune percentage accuracy (P@1) when retrieving a target image from a set of 100 candidates taken from the COCO, Concpetual Captions and Concadida test set, and nocaps validation sets.

COCO				
Model	B@4	M	C	S
ClipCap-COCO	**32.60**	**27.50**	**108.55**	**20.33**
DiscriTune-COCO	32.31	26.05	105.40	20.03
Conceptual Captions				
Model	B@4	M	C	S
ClipCap-ConCap	**7.32**	**10.81**	**87.22**	**18.07**
DiscriTune-ConCap	3.92	8.79	55.26	15.40

Table 2. NLG metrics (BLEU@4 [30], METEOR [10], CIDEr [43] and SPICE [2]) for ClipCap and DiscriTune captions tested in-domain on COCO and Conceptual Captions.

the results obtained with ClipCap and the benefits of our finetuning method to generate captions that are closer to human references produced with the communicative intent of replacing an image. Given that BLIP is not based on a CLIP encoder, this experiment also refutes the hypothesis that DiscriTune performance gains are simply due to using CLIP both as visual encoder for the captioner and as retriever, confirming the wider applicability of our method.

4.2.4 Human Text-Based Image Retrieval

The results in tables 1 and 2 show that, when testing on Conceptual Captions, DiscriTune-ConCap produces outputs that are less similar to human captions than those of ClipCap-ConCap, but it greatly outperforms the latter in text-based image retrieval accuracy. Recall that the ground-truth captions in Conceptual Captions come from the alt-texts associated to images harvested from the Web [37]. As discussed in more detail in the qualitative analysis below, when taken out of context, such captions are often non-informative, and thus it's not clear that learning to reproduce their style as closely as possible, like vanilla ClipCap-ConCap does, is a good idea. Consider for example Fig. 2(c) below. ClipCap-ConCap is perfectly reproducing the ground-truth description ("digital art selected for the #"), and yet this is not as informative as the caption produced by DiscriTune-ConCap ("a boy in a pond with a lot of stars"). We thus conjecture that, despite their lower faithfulness to the human ground truth, DiscriTune-ConCap captions are more informative than ClipCap captions, and possibly even more informative than the original human captions in this dataset.

COCO				
Model	B@4	M	C	S
ClipCap-ConCap	8.50	13.29	37.03	9.77
DiscriTune-ConCap	**13.99**	**16.97**	**53.20**	**12.07**
Conceptual Captions				
ClipCap-COCO	1.47	6.43	23.74	7.98
DiscriTune-COCO	**1.71**	**6.58**	**28.01**	**9.00**
Flickr				
ClipCap-COCO	17.21	18.43	41.65	12.04
DiscriTune-COCO	**18.48**	**18.61**	**44.78**	**12.68**
ClipCap-ConCap	8.28	12.24	27.57	7.81
DiscriTune-ConCap	13.01	15.16	36.35	9.44
nocaps-near				
ClipCap-COCO	30.47	**24.36**	69.66	10.89
DiscriTune-COCO	**32.87**	24.11	**70.63**	**10.98**
ClipCap-ConCap	10.41	13.25	30.47	5.72
DiscriTune-ConCap	18.93	17.66	46.45	7.53
nocaps-out				
ClipCap-COCO	20.32	20.22	51.74	8.55
DiscriTune-COCO	**24.10**	**20.49**	**57.06**	**8.83**
ClipCap-ConCap	10.69	13.15	36.57	5.71
DiscriTune-ConCap	16.76	17.03	54.03	7.56
Concadia-Descriptions				
ClipCap-COCO	1.94	5.57	14.99	6.44
DiscriTune-COCO	**2.03**	**5.65**	16.70	**7.15**
ClipCap-ConCap	0.62	3.69	12.77	5.82
DiscriTune-ConCap	1.12	4.81	**17.20**	7.13
Concadia-Captions				
ClipCap-COCO	0.24	2.45	4.35	2.11
DiscriTune-COCO	0.30	2.49	5.57	2.59
ClipCap-ConCap	**0.50**	2.74	9.22	3.65
DiscriTune-ConCap	0.30	**2.79**	**9.37**	**4.20**

Table 3. NLG metrics (BLEU@4, METEOR, CIDEr and SPICE) for ClipCap and DiscriTune captions tested across domain on COCO (Conceptual Captions-based models only), Conceptual-Captions (COCO-based models only), Flickr, nocaps, Concadia.

Model	*Flickr*	*nocaps near*	*nocaps out*	*Concadia*
BLIP	75.1	87.0	90.6	71.0
DiscriTune-BLIP	**80.8**	**90.2**	**92.6**	**74.9**

Table 4. BLIP and DiscriTune-BLIP percentage accuracy on the cross-domain retrieval task described in Section 4.2.1.

Model	B@4	M	C	S
Flickr				
BLIP	27.18	22.74	70.63	16.03
DiscriTune-BLIP	**28.19**	**23.61**	**74.77**	**16.9**
nocaps-near				
BLIP	44.91	29.52	107.13	14.52
DiscriTune-BLIP	**45.46**	**29.81**	**108.70**	**14.97**
nocaps-out				
BLIP	**38.42**	26.95	105.63	13.77
DiscriTune-BLIP	37.34	**27.13**	**106.23**	**14.14**
Concadia-Descriptions				
BLIP	2.71	6.90	26.02	9.75
DiscriTune-BLIP	**2.90**	**7.18**	**27.20**	**9.98**
Concadia-Captions				
BLIP	0.66	3.14	9.97	3.84
DiscriTune-BLIP	**0.74**	**3.35**	**10.84**	**3.96**

Table 5. NLG metrics (BLEU@4, METEOR, CIDEr and SPICE) for BLIP and DiscriTune-BLIP captions tested across domains.

To verify this hypothesis, we designed an experiment in which human annotators had to select a target image from a set of 10 candidates, based either on a ground-truth human description, or a ClipCap-ConCap or DiscriTune-ConCap-generated one. We sampled the data for this experiment from our Conceptual Captions test set. To make sure that the captions needed to be genuinely informative in order to allow successful target retrieval, we selected hard distractors among the nearest visual neighbours of the target. The latter were found by passing images through the CLIP ViT-B/32 visual encoder and computing the cosine of the resulting representations.[8] We collected human data for 500 target+distractor sets, for each of the three caption types. We recruited Amazon Mechanical Turk participants who saw 100 sets each.[9] Experimental details (including ethical approval) are in Appendix A.

Results are reported in Table 6. They confirm that the task, due to the hard distractors, is quite challenging, with the annotators failing to reach 50% accuracy independently of caption type. However, even with the least informative ClipCap-ConCap captions, humans are well above the

[8]We found that very high-similarity neighbours are often near duplicates of the target image and we excluded those with similarity above 0.8

[9]https://www.mturk.com/

Captions	accuracy
Human	42.8
ClipCap-ConCap	36.2
DiscriTune-ConCap	**47.6**

Table 6. Percentage accuracy of human annotators when selecting a target image among 9 distractors based on the target human-, ClipCap- or DiscriTune-generated captions.

10% chance level. Importantly, the DiscriTune-ConCap captions greatly outperform not only their ClipCap counterparts, but also the human ones, with a solid 5% accuracy boost. We thus confirm our conjecture that, given a relatively noisy dataset such as Conceptual Captions, our finetuning procedure can produce captions that are more informative than the original human-produced descriptions. In the next section, we explore the differences between human and DiscriTune-generated captions.

5. Caption Analysis

Fig. 2 shows 6 Conceptual Captions examples from the experiment with humans in which human annotators were only able to guess the target given the DiscriTune-generated captions (correct target image is on the left; image picked by subjects in response to the human-generated caption is shown on the right). These cases, selected among those where no human face is recognizable, were hand-picked to illustrate various phenomena. They are however representative of the data, which we manually inspected as a whole.

Starting with a comparison of DiscriTune and ClipCap captions,[10] we observe that the latter tend to be more vague (cf. examples (b), (c), (d) and (f)), where sometimes these vague captions are actually matching the human-generated ones very closely (as in (c)). The other issue with ClipCap captions is that they are often inaccurate, as in example (a), where people are playing in the water, not on the beach; and it's hard to believe the desolated landscape in (b) is a tourist attraction. Note that the use of a generic term such as *tourist attraction* in this last example is a peculiarity that ClipCap inherited from the Conceptual Captions style (the dataset was constructed by replacing place and people names with generic hypernyms such as this one). Clearly, learning to reproduce such Conceptual Captions-specific idiosyncrasies penalizes ClipCap when it's evaluated on other datasets. On the other hand, it's remarkable that discriminative finetuning, built on top of this very ClipCap system, was able to steer the captions back towards a more descriptive and precise language.

Comparing DiscriTune with the human captions, we see that the former tend to be more plainly descriptive and pre-

[10]As the focus of this analysis in on the Conceptual Captions-trained/finetuned models, we will drop the -ConCap suffix.

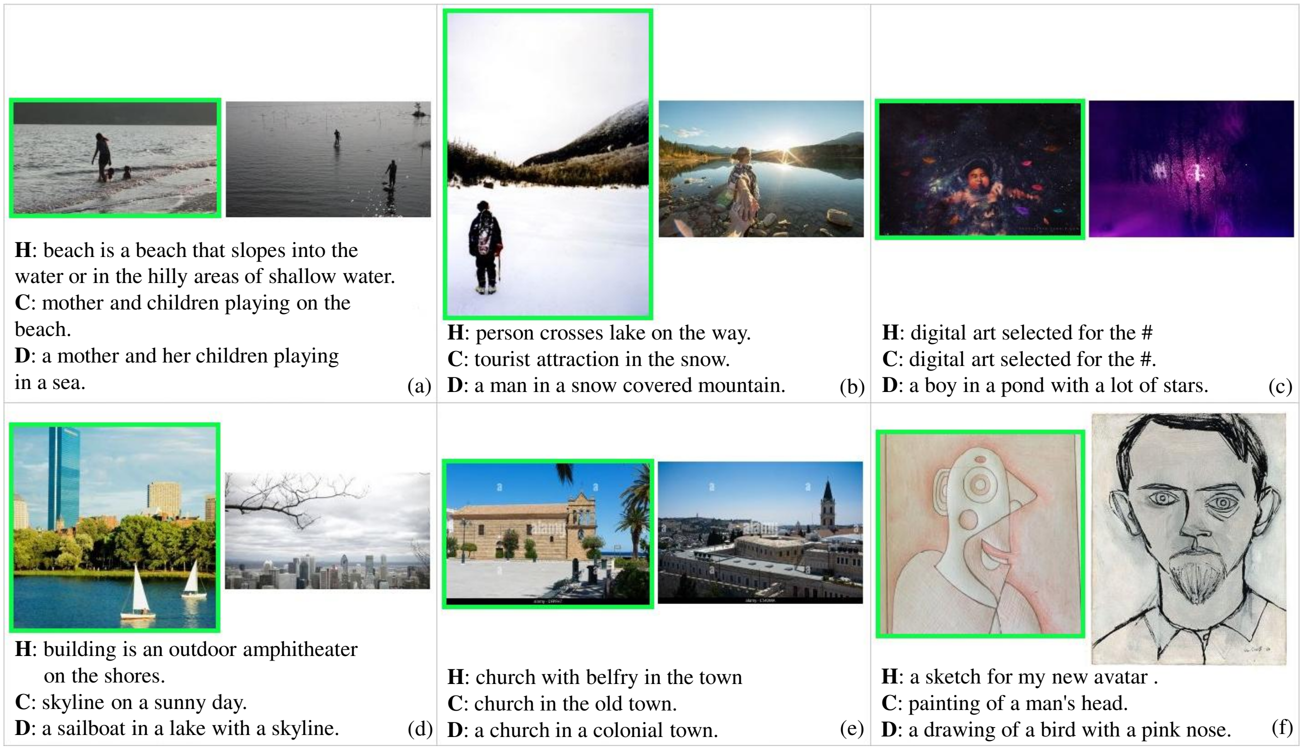

Figure 2. Example Conceptual Captions images with the corresponding captions produced by humans (**H**), ClipCap (**C**) and DiscriTune (**D**), respectively. In all cases, annotators were able to identify the target image (framed in green) only when receiving the DiscriTune caption as input. The image on the right of each pair is the distractor that was chosen by annotators when receiving the human caption as input.

cise than the latter. Often, human captions in the Conceptual Captions dataset contain non-discriminative "meta" information about an image that is not useful to identify it, or might have made sense in the context of the original web page, but becomes opaque once the image and its alt-text are extrapolated. In example (a), the human caption is just stating that we are on a beach with shallow water, so that the human discriminator picked another shallow-beach picture. The DiscriTune caption precisely reports that there are a woman and children in the water. The human caption of example (b) might be accurate, but without more context it's difficult to recognize the white plateau as a frozen lake. Consequently, the human discriminator wrongly chose a more stereotypical picture of a lake. Concerning examples (c) and (f), the human captions report that they are "digital art" and an "avatar", respectively, leading the annotators to pick other potential exemplars of digital art and avatars from the distractors. The DiscriTune caption for (f) is actually inaccurate, as the avatar is not a bird, but the mention of a pink nose nevertheless helped the subject identifying the right image. The human caption of example (d) refers to an outdoor amphitheater that is not visible in the picture. Again, DiscriTune was more helpful to the human discriminators by providing a plain description of the picture contents. Finally, example (e) is interesting because the human caption highlights the somewhat atypical belfry in the picture (and so the human discriminator picked a photo with a more prominent belfry), whereas the DiscriTune caption provides a more discriminative cue by mentioning the colonial style of the landscape.

human	*ClipCap*	*DiscriTune*
new	young	red
other	old	green
musical	biological	blue
outdoor	close	black
small	white	pink
long	trendy	colorful
big	digital	yellow
low	funny	denim
large	aerial	purple
beautiful	general	high

Table 7. Top 10 adjective lemmas most associated to a caption type (*human*, *ClipCap* or *DiscriTune*) according to the local Mutual Information association statistics computed on all captions generated for our full Conceptual Captions test set.

To get a more general sense of the language of the different caption types, we lemmatized and part-of-speech tagged the full Conceptual Captions test set caption corpora with Spacy.[11] We then computed the local Mutual Information statistics [14] across all possible lemma/caption-type pairs. We restricted the analysis to adjectives and nouns, as we found these two parts of speech to include the most visually descriptive terms. The adjective results are presented in Table 7 (noun analysis is in Appendix E). The difference between the adjectives in human and DiscriTune captions is striking: the latter are nearly all highly visual descriptive terms (in particular, colours), whereas the former contain several terms that are hardly providing any visual information (*new*, *musical*, *other*). Even the most concrete human-caption adjectives are not as specific as those strongly associated with DiscriTune (compare *small, big, large, beautiful* to *red, blue, purple, denim*). The ClipCap list also contains few adjectives that might be genuinely useful to discriminate a specific image (*white*, *aerial*), with most being either abstract or very generic (*biological*, *trendy*, *funny*, *general*). It is remarkable that self-supervised discriminative finetuning, probably by exploiting the multimodal knowledge encoded in the pre-trained ClipCap components, is able to recover a highly visually descriptive language, despite the fact that it operates on a system that has been trained on human captions that, as we have just seen, are not as plainly descriptive, and that the system is not exposed to any new language during finetuning.

6. Conclusion

We presented a simple finetuning method to make model-generated captions more discriminative. Given a pre-trained captioner, its text generation component is finetuned on the task of helping a black-box text-based image retriever picking a target image among distractors. The task only requires unannotated images, and we were able to make the system work with the basic REINFORCE algorithm. We leave the exploration of more sophisticated reinforcement learning techniques as an obvious direction for future work. Our results are reported using two captioners, ClipCap, a decoder-only model, and BLIP, an encoder-decoder trained with multitask learning on web-scale data.

We found that the discriminatively finetuned captions do not improve over the original ones in terms of similarity to human ground truth, when tested on the same dataset the captioner was trained on. However, for both models and on a variety of out-of-domain datasets, they consistently outperform those of the original captioner. Discriminative pressure might be a strong enough signal to "unlearn" some of the overfit on the caption style of the pre-training dataset, and instead better capture the semantic content of the image. What's more, for the noisily annotated Conceptual Captions data-set (where we observed the largest performance drop in terms of mimicking ground-truth descriptions when finetuning ClipCap), discriminatively finetuned captions are more helpful than ground-truth captions, not only to a neural retriever, but also for humans tasked with a challenging image identification task. This suggests that our system could be used as-is to generate Web image captions that are on average more informative to users than alt-text descriptions (which are the source of Conceptual Captions annotations).

Qualitatively, we find that, even when finetuning the Conceptual Captions-trained captioner (that has learned to reproduce the somewhat abstract style of alt-text descriptions), our discriminative finetuning procedure recovers a more precise and plainly descriptive language. Compared to those of the original captioner, it is also clear why these more descriptive captions, that shed the idiosyncrasies of alt-text, will generalize better to other datasets. We focused our analysis on ClipCap and Conceptual Captions, since this is the setup where we observed the largest discrepancy between human and discriminatively-generated captions, and a marked asymmetry in retrieval vs. generation performance. We leave a thorough investigation of how the nature of the captions used to train the backbone captioner affects our method to future work. Given that several pre-trained text-based image retrievers are publicly available, another interesting direction would be to alternate different retrievers during finetuning. This might help the model further generalize, as it would be less likely to overfit the quirks of one specific retriever.

Finally, there has been recent progress in training models to learn from human feedback through reinforcement learning [28, 29, 40]. Given the costly human annotation process required by this approach, our method could be seen as a cheaper alternative, exploiting "neural" feedback to guide the finetuning of an existing model. Future research directions should study the interplay between human and neural feedback to improve the capabilities of current systems.

Acknowledgements

We thank the area chairs and reviewers for feedback. We thank Corentin Kervadec for feedback and help in creating Figure 1, and Lorenzo Baraldi and Manuele Barraco for replying to a Sunday night email during the rebuttal. EG, NCR, FF and MB received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements No. 715154 and No. 101019291) and the Spanish Research Agency (ref. PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033). This paper reflects the authors' view only, and the funding agencies are not responsible for any use that may be made of the information it contains.

[11] https://spacy.io/

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 4

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 382–398, Cham, 2016. Springer International Publishing. 5

[3] 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 *Proceedings of CVPR*, pages 6077–6086, Salt Lake City, UT, 2018. 1, 2

[4] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. CaMEL: Mean Teacher Learning for Image Captioning. In *International Conference on Pattern Recognition*, 2022. 2

[5] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016. 1

[6] Paul Bloom. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA, 2000. 1

[7] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of ICCV*, pages 521–530, Venice, Italy, 2017. 2

[8] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with CLIP reward. In *Findings of NAACL*, pages 517–527, Seattle, WA, 2022. 2

[9] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Proceedings of NIPS*, pages 898–907, Long Beach, CA, 2017. 2

[10] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. 5

[11] Roberto Dessì, Eleonora Gualdoni, Francesca Franzon, Gemma Boleda, and Marco Baroni. Communication breakdown: On the low mutual intelligibility between human and neural captioning. In *Proceedings of EMNLP*, Abu Dhabi, United Arab Emirates, 2022. In press. 4

[12] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, 2017. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[14] Stephanie Evert. *The Statistics of Word Cooccurrences*. Ph.D dissertation, Stuttgart University, 2005. 8

[15] Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan Clark, and Regina Barzilay. CapWAP: Image captioning with a purpose. In *Proceedings of EMNLP*, pages 8755–8768, Online, 2020. 1

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of EMNLP*, pages 7514–7528, Punta Cana, Dominican Republic, 2021. 2

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of CVPR*, pages 3128–3137, Boston, MA, 2015. 4

[18] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of EMNLP*, pages 787–798, Doha, Qatar, 2014. 1

[19] Eugene Kharitonov, Roberto Dessì, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. EGG: a toolkit for research on Emergence of lanGuage in Games. `https://github.com/facebookresearch/EGG`, 2021. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[21] Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. Concadia: Towards image-based text generation with a purpose. In *Proceedings of EMNLP*, Abu Dhabi, United Arab Emirates, 2022. In press. 1, 4

[22] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022. 3

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. `https://arxiv.org/abs/2201.12086`, 2022. 2, 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of ECCV*, pages 740–755, Zurich, Switzerland, 2014. 2, 3

[25] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of ECCV*, 2018. Published online: `https://openaccess.thecvf.com/ECCV2018`. 2

[26] Ruotian Luo, Gregory Shakhnarovich, Scott Cohen, and Brian Price. Discriminability objective for training descriptive captions. In *Proceedings of CVPR*, pages 6964–6974, Salt Lake City, UT, 2018. 2

[27] Ron Mokady, Amir Hertz, and Amit Bermano. ClipCap: CLIP prefix for image captioning. https://arxiv.org/abs/2111.09734, 2021. 1, 2, 4

[28] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. 8

[29] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 8

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 5

[31] André Susano Pinto, Alexander Kolesnikov, Yuge Shi, Lucas Beyer, and Xiaohua Zhai. Tuning computer vision models with task rewards. *arXiv preprint arXiv:2302.08242*, 2023. 1

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763, virtual conference, 2021. 2, 3

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf, 2019. 2

[34] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732, 2016. 2

[35] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 2

[36] Ananya Sai, Akash Kumar Mohankumar, and Mitesh Khapra. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys*, 2022. In press. 1

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, pages 2556–2565, Melbourne, Australia, 2018. 2, 3, 5

[38] Brian Skyrms. *Signals: Evolution, learning, and information*. Oxford University Press, Oxford, UK, 2010. 1

[39] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. In press. 1

[40] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 8

[41] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 3

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008, Long Beach, CA, 2017. 3

[43] R. Vedantam, C. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society. 5

[44] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*, 2015. 1, 2

[45] Ronald Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 3

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML*, pages 2048–2057, Lille, France, 2015. 1

[47] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, and Kai Lei. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia*, 21(4):1047–1061, 2019. 2

[48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3, 4

[49] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, JaeSung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, and Yejin Choi. Multimodal knowledge alignment with reinforcement learning. https://arxiv.org/abs/2205.12630, 2022. 2

[50] Wentian Zhao, Xinxiao Wu, and Jiebo Luo. Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE Transactions on Image Processing*, 30:1180–1192, 2021. 2