Revisiting Image Captioning Training Paradigm via Direct CLIP-based Optimization

Nicholas Moratelli*, Davide Caffagni*, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara
University of Modena and Reggio Emilia, Italy
{name.surname}@unimore.it * Equal contribution

BMVC Oral 2024

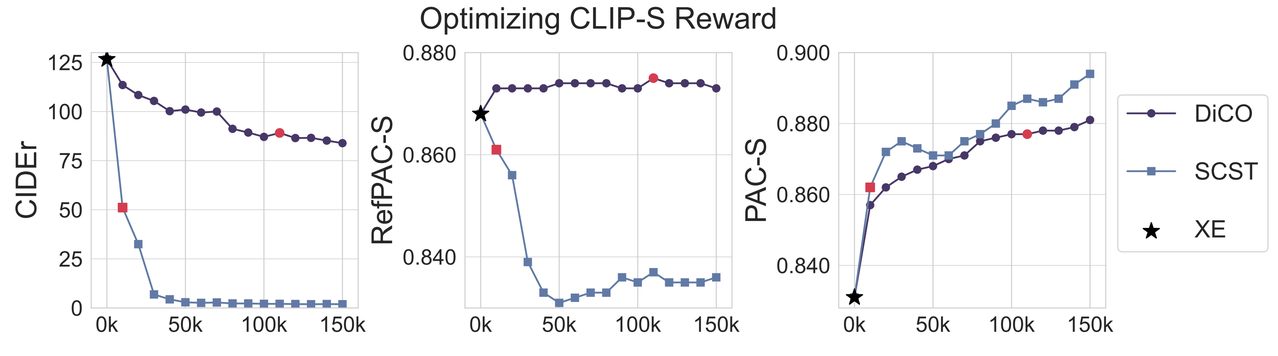
UNIMORE AImage Lab

Overview and Motivations

- The **discrepancy** between the XE objective and the metrics employed for evaluation makes SCST [1] the standard for fine-tuning captioning models.
- New evaluation metrics (*e.g.* CLIP-S [2], PAC-S [3]) and training algorithms (*e.g.* RLHF [4], DPO [5]) are being studied to better reflect **human preferences** for the output of a generative model.
- RLHF is **impracticable** due to the lack of preference annotations.

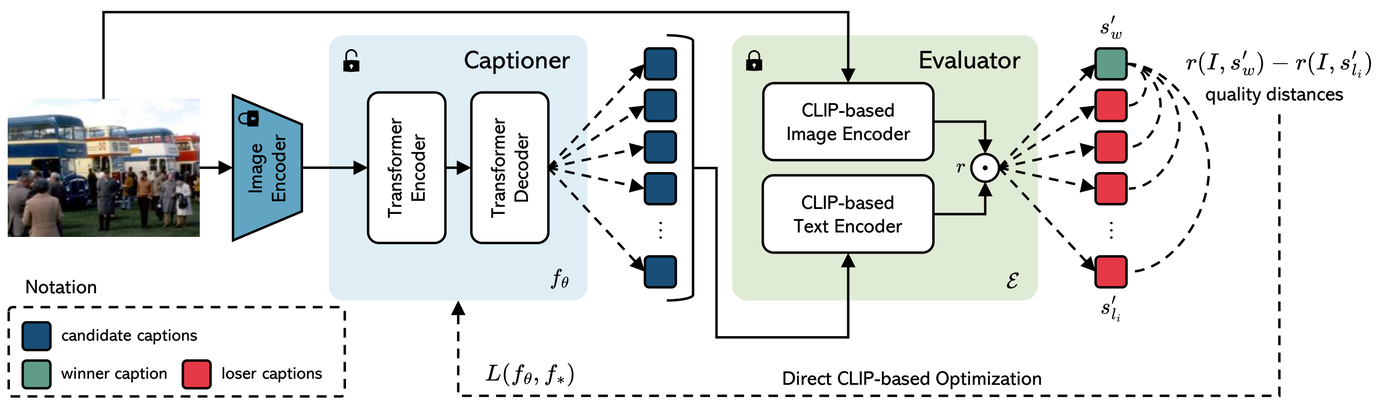
Can we apply the well-known SCST to optimize modern reference-free captioning metrics? Yes, but it is not trivial.
*We propose **Di**rect **C**LIP-based **O**ptimization as a better alternative.*

DiCO jointly learns and optimizes a reward model that is distilled from a CLIP-based learnable captioning evaluator with high human correlation. This is done by solving a weighted classification problem directly inside the captioner.



SCST (after 10k iters): A group of young boys kicking around a soccer ball on a soccer field with other young boys running around with net in background.
SCST (after 50k iters): Young boys kicking soccer ball around soccer goal kicking grass underneath a goal on grass behind background behind surface surface with trees in background behind surface surface.
SCST (after 100k iters): Young boys soccer teams chasing after after soccer soccer goalie in background with green leaves on grass behind background surface court with young boy.
DiCO (Ours): A group of young children kicking a soccer ball in a field.

Direct CLIP-based Optimization (DiCO)



1. We let the captioner generate $k + 1$ candidate captions (*e.g.* through beam search).

2. We select the caption with the highest score according to a captioning evaluator $\mathcal{E}$ (*e.g.* CLIP-S) and denote it as s'_w (*i.e.* "winner"). The others, instead, are denoted as $\{s'_{l_i}\}_{i=1}^{k}$ (*i.e.* "losers").

3. To make the reward model more robust and accurate, we also impose that it can predict the *relative quality distances* between the winner and the loser captions. It is computed as a normalized probability distribution between score distances, as follows:

$$\gamma_i = \text{softmax}_{s'_{l_1},\ldots,s'_{l_k}}\left(\frac{\mathcal{E}(I, s'_w) - \mathcal{E}(I, s'_{l_i})}{\tau}\right)$$

4. The final loss of **DiCO** is: $L(f_\theta, f_*) = -E\left[\log\sigma\left(\beta\log\frac{f_\theta(s'_w|I)}{f_*(s'_w|I)} - \beta\sum_{i=1}^{k}\gamma_i\log\frac{f_\theta(s'_{l_i}|I)}{f_*(s'_{l_i}|I)}\right)\right]$

CLIP-based Optimization Results

Model			Reference-based			Reference-free			
			B-4	M	C	CLIP-S	PAC-S	R@1	MRR
Standard Captioners		**Backbone**							
CLIP-VL		RN50×4	40.2	29.7	134.2	0.770	0.826	24.0	34.8
COS-Net		RN101	42.0	30.6	141.1	0.758	0.832	25.8	37.1
PMA-Net		ViT-L/14	43.0	30.6	144.1	0.755	0.821	-	-
LLM-based Captioners		**Backbone**							
ZeroCap		ViT-B/32	2.3	10.1	15.1	0.810	0.816	-	-
ClipCap		ViT-B/32	32.3	28.1	108.5	0.766	0.833	27.1	38.3
SmallCap		ViT-B/32	37.0	27.9	119.7	0.748	0.826	23.1	33.7
MiniGPT-v2		ViT-g/14	18.8	24.6	80.4	0.752	0.818	27.4	37.9
BLIP-2		ViT-g/14	43.7	32.0	145.8	0.767	0.837	31.4	42.7
SCST	CLIP-S	ViT-B/32	11.4	23.1	1.1	**0.851**	0.846	43.4	55.1
SCST	PAC-S	ViT-B/32	20.3	27.1	40.7	0.810	0.870	50.0	61.8
DiCO	CLIP-S	ViT-B/32	22.6	27.0	81.7	0.825	0.858	46.3	58.0
DiCO	PAC-S	ViT-B/32	**23.7**	**27.3**	**84.8**	0.814	**0.882**	**52.9**	**64.8**

CIDEr-based Optimization Results

Model	Reference-based					Reference-free	
	B-4	M	C	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S
Up-Down	36.3	27.7	120.1	0.787	0.848	0.723	0.803
SGAE	39.0	28.4	129.1	0.796	0.855	0.734	0.812
AoANet	38.9	29.2	129.8	0.797	0.857	0.737	0.815
$\mathcal{M}^2$ Transformer	39.1	29.8	131.3	0.793	0.852	0.734	0.813
COS-Net	42.0	30.6	141.1	0.814	0.870	**0.758**	**0.832**
PMA-Net	43.0	30.6	144.1	0.814	0.869	0.755	0.821
SCST	43.6	30.8	143.3	0.809	0.866	0.750	0.826
DiCO (w/ β = 0.05)	43.2	31.2	**144.2**	0.815	0.871	0.756	0.831
DiCO (w/ β = 0.1)	**43.7**	31.2	143.8	**0.817**	**0.872**	0.757	**0.832**
DiCO (w/ β = 0.2)	**43.7**	**31.3**	143.5	0.816	**0.872**	0.756	0.831

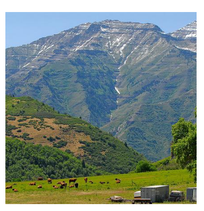
Considerations and Qualitative Results

- **DiCO** showcases SoTA performance on reference-free metrics, outscoring larger multimodal LLMs such as MiniGPT-v2 and BLIP-2.
- In striking contrast to SCST, it maintains competitive results on reference-based metrics, which allows the generation of correct and fluent captions.
- However, **DiCO** also demonstrates SoTA results when optimizing a reference-based metric (*e.g.* CIDEr). β can be tuned to control the deviation from the reference model pre-trained with XE.

Training	Reward	Reference-based			Reference-free			
		B-4	M	C	CLIP-S	PAC-S	R@1	MRR
XE	-	37.3	30.4	126.6	0.758	0.831	27.7	38.5
RLHF	HF	21.4	27.8	57.9	0.745	0.819	24.7	34.9
RLHF	CLIP-S	12.9	24.2	2.3	0.732	0.794	19.5	29.0
RLHF	PAC-S	12.4	23.7	2.0	0.726	0.790	18.1	27.5

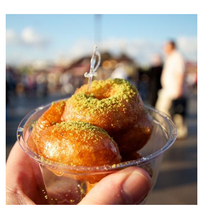
- RLHF degrades the reference model because there is not enough data to train a good reward model for captioning.

SCST: A rusted rusty rusty rusty rusty rusty rusty scissors in a garden with plants in the background.
DiCO (Ours): A rusted scissors sticking out of a metal fence with plants in the background.



SCST: Several cows grazing in a field on a mountain range with a mountain in distance under mountain range in background.
DiCO (Ours): A herd of cattle grazing on a lush green hill next to a large mountain range.

SCST: A little boy with headphones sitting in front of a computer with headphones on a desk in corner with headphones on background.
DiCO (Ours): A little boy with headphones sitting at a desk using a computer.



SCST: A person holding up a clear plastic plastic container filled with sugar covered donuts with people in background on background.
DiCO (Ours): A hand holding a plastic container filled with sugar covered donuts.

References

[1] Rennie et al. *Self-Critical Sequence Training for Image Captioning*. CVPR, 2017
[2] Hessel et al. *CLIPScore: A Reference-free Evaluation Metric for Image Captioning*. EMNLP, 2021.
[3] Sarto et al. *Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation*. CVPR, 2023.
[4] Ouyang et al. *Training language models to follow instructions with human feedback*. NeurIPS, 2022.
[5] Rafailov et al. *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. NeurIPS, 2023.
[6] Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. ICML, 2021.