

Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation

Supplementary Material

Sara Sarto¹ Manuele Barraco¹ Marcella Cornia¹ Lorenzo Baraldi¹ Rita Cucchiara^{1,2}

¹University of Modena and Reggio Emilia, Modena, Italy ²IIT-CNR, Pisa, Italy

{name.surname}@unimore.it

In the following, we present additional experiments and qualitative results about the proposed PAC-S metric.

1. Additional Experimental Results

Correlation with MID score. In addition to the experiments presented in the main paper, we conducted further comparisons with the MID metric [3]. Since it exploits CLIP-based features as CLIP-S [2] and our proposal, in Table 1 we compare the results of the original MID score with a re-implemented version that uses our embeddings in place of those of CLIP. In particular, we conduct this analysis on the Flickr8k-Expert, Flickr8k-CF, and FOIL datasets and show that using our embeddings can further improve the results of the MID score in the majority of the considered settings, thus further demonstrating the appropriateness of our positive-augmented contrastive learning approach.

Reference-based results using ViT-based backbones. As a complement to Table 8 of the main paper, in Table 2 we report the reference-based results using different cross-modal features. In particular, we experiment with different ViT-based backbones of CLIP [4] and OpenCLIP [6] models. From these results, we confirm the effectiveness of PAC-S also in the reference-based setting on both image and video captioning datasets. Both ViT-L/14 models outperform the others even in this case, still confirming that using more powerful features can lead to better results.

Analyzing ResNet-based backbones. In Table 3, we conduct the same analysis in both reference-free and reference-based settings but using visual features extracted from a ResNet backbone [1]. Specifically, we use the following CLIP-based models: ResNet-50, ResNet-101, and ResNet-50×4, which employ an EfficientNet-style architecture scaling. For these experiments, we finetune the last attention pooling of the visual backbone and the final projection of the textual branch using the same settings described in the main paper. Also in this case, our metric

Features	Flickr8k-Expert		Flickr8k-CF		Pascal-50S		FOIL	
	Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c	Accuracy	Accuracy		
MID [3]	CLIP	-	54.9	37.3	-	85.2	90.5	
MID [†]	CLIP	54.3	54.6	36.5	18.7	84.6	93.2	
MID [†]	PAC (ours)	54.7	55.1	36.7	18.8	85.0	93.3	

Table 1. Performance of MID with CLIP and PAC ViT-B/32 features. The [†] marker indicates our re-implementation.

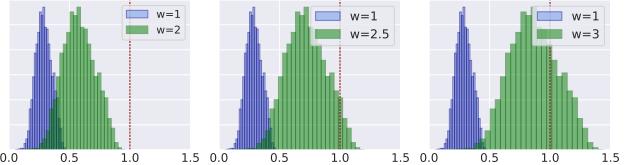


Figure 1. Distribution of PAC scores using different w (Eq. 1 of the main paper).

achieves the best results in almost all datasets, with the only exception of VATEX-EVAL in which the EMScore obtains slightly better correlation scores.

Choice of hyperparameters. The scaling factor, denoted by w in Eq. 1 of the main paper, is utilized to adjust the scale of the final metric to improve its numerical readability, without affecting the ranking of the results. CLIP-S also employs a comparable technique, where w is assigned the value of 2.5. To provide additional clarification, we present in Fig. 1 the impact of varying values of w . The raw PAC-S scores with $w = 1$ lie between 0 and 0.5 on all datasets. Therefore, we decide to use a scaling factor w equal to 2 which stretch the PAC-S scores between 0 and 1.

2. Generated Samples and Qualitatives

Fig. 2 shows additional image-text generated examples used for the presented positive-augmented contrastive learning strategy. As it can be seen, both image and text generated samples are realistic and plausible and can be effectively used as an additional source of supervision.

		Flickr8k-Expert		Flickr8k-CF		VATEX-EVAL		PASCAL-50S	FOIL	ActivityNet-FOIL
		Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c	Kendall τ_b	Spearman ρ	Accuracy	Accuracy	Accuracy
CLIP ViT-B/16	RefCLIP-S [2]	53.6	54.0	36.7	19.0	-	-	84.0	94.8	-
	EMScoreRef [5]	-	-	-	-	37.1	47.5	-	-	92.2
	RefPAC-S	56.0 (+2.4)	56.4 (+2.4)	37.5 (+0.8)	19.4 (+0.4)	38.8 (+1.7)	49.6 (+2.1)	84.8 (+0.8)	95.1 (+0.3)	92.6 (+0.4)
CLIP ViT-L/14	RefCLIP-S [2]	54.0	54.4	36.5	18.9	-	-	85.0	94.9	-
	EMScoreRef [5]	-	-	-	-	37.0	47.4	-	-	93.5
	RefPAC-S	56.7 (+2.7)	57.1 (+2.7)	37.7 (+1.2)	19.5 (+0.6)	38.6 (+1.6)	49.3 (+1.9)	85.0 (+0.0)	95.3 (+0.4)	94.2 (+0.7)
OpenCLIP ViT-B/32	RefCLIP-S [2]	53.9	54.3	36.8	19.0	-	-	84.7	94.7	-
	EMScoreRef [5]	-	-	-	-	38.4	49.1	-	-	93.0
	RefPAC-S	54.8 (+0.9)	55.2 (+0.9)	37.4 (+0.6)	19.3 (+0.3)	38.8 (+0.4)	49.5 (+0.4)	84.5 (-0.2)	94.1 (-0.6)	93.6 (+0.6)
OpenCLIP ViT-L/14	RefCLIP-S [2]	55.7	55.8	37.5	19.4	-	-	85.3	95.9	-
	EMScoreRef [5]	-	-	-	-	39.4	50.3	-	-	94.0
	RefPAC-S	56.5 (+0.8)	56.9 (+1.1)	38.0 (+0.5)	19.7 (+0.3)	40.3 (+0.9)	51.4 (+1.1)	84.9 (-0.4)	95.8 (-0.1)	94.4 (+0.4)

Table 2. Captioning evaluation results in a reference-based setting on both image and video captioning datasets using different cross-modal features.

		Flickr8k-Expert		Flickr8k-CF		VATEX-EVAL		PASCAL-50S	FOIL	ActivityNet-FOIL
		Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c	Kendall τ_b	Spearman ρ	Accuracy	Accuracy	Accuracy
CLIP RN50	CLIP-S [2]	51.0	51.4	34.0	17.6	-	-	80.6	87.9	-
	EMScore [5]	-	-	-	-	22.0	28.6	-	-	87.0
	RefPAC-S	52.6 (+1.6)	52.9 (+1.5)	34.6 (+0.6)	17.9 (+0.3)	19.4 (-2.6)	25.4 (-3.2)	81.7 (+1.1)	87.1 (-0.8)	87.7 (+0.7)
CLIP RN50	RefCLIP-S [2]	52.5	52.8	35.9	18.5	-	-	83.4	93.4	-
	EMScoreRef [5]	-	-	-	-	36.6	46.9	-	-	91.8
	RefPAC-S	54.1 (+1.6)	54.5 (+1.7)	36.4 (+0.5)	18.8 (+0.3)	36.4 (-0.2)	46.7 (-0.2)	83.8 (+0.4)	93.1 (-0.3)	92.7 (+0.9)
CLIP RN101	CLIP-S [2]	50.5	50.9	33.5	17.3	-	-	80.5	89.1	-
	EMScore [5]	-	-	-	-	21.6	28.2	-	-	89.6
	RefPAC-S	53.4 (+2.9)	53.7 (+2.8)	34.4 (+0.9)	17.8 (+0.5)	20.4 (-1.2)	26.6 (-1.6)	81.8 (+1.3)	89.0 (-0.1)	88.9 (-0.7)
CLIP RN101	RefCLIP-S [2]	52.2	52.6	35.6	18.4	-	-	83.3	95.2	-
	EMScoreRef [5]	-	-	-	-	36.6	46.9	-	-	91.7
	RefPAC-S	55.5 (+3.3)	55.9 (+3.3)	36.6 (+1.0)	18.9 (+0.5)	37.1 (+0.5)	47.5 (+0.6)	84.8 (+1.5)	95.4 (+0.2)	92.1 (+0.4)
CLIP RN50×4	CLIP-S [2]	50.7	51.0	34.0	17.6	-	-	80.7	89.5	-
	EMScore [5]	-	-	-	-	22.0	28.8	-	-	88.8
	RefPAC-S	53.9 (+3.2)	54.3 (+3.3)	35.9 (+1.9)	18.6 (+1.0)	21.9 (-0.1)	28.6 (-0.2)	82.5 (+1.8)	90.5 (+1.0)	87.7 (-1.1)
CLIP RN50×4	RefCLIP-S [2]	52.3	52.7	36.1	18.7	-	-	83.3	95.3	-
	EMScoreRef [5]	-	-	-	-	36.7	45.0	-	-	91.5
	RefPAC-S	56.2 (+3.9)	56.6 (3.9)	37.3 (+1.2)	19.3 (+0.6)	37.4 (+0.7)	47.7 (+2.7)	84.8 (+1.5)	95.8 (+0.5)	91.9 (+0.4)

Table 3. Additional human correlation and accuracy scores on both image and video captioning datasets using different cross-modal ResNet-based backbones.

We report in Fig. 3 some additional qualitative comparisons between PAC-S and well-known metrics on the Pascal-50S dataset. These qualitative results show that in the majority of cases PAC-S is more aligned with the human judgments than other metrics. Finally, in Fig. 4 and 5, we report sample results comparing our metric with CLIP-S [2] on FOIL, Flickr8k-Expert, and Flickr8k-CF datasets. As it can be observed, PAC-S can correctly identify hallucinated objects and better correlates with human judgments,

demonstrating its effectiveness compared to CLIP-S also from a qualitative point of view.

References

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation



Figure 2. Additional real and generated image-text samples used to augment the training set for positive-augmented contrastive learning.

Image	Candidate Captions	Evaluation Scores				Image	Candidate Captions	Evaluation Scores			
		METEOR	CIDEr	CLIP-S	PAC-S			METEOR	CIDEr	CLIP-S	PAC-S
	A blue bird being held by a handler.	35.2	96.3	80.1	80.0		A passenger train in the snow.	26.8	89.7	83.5	83.1
	A blue bird perched on a gloved hand.	18.6	39.0	76.1	82.1		A red train driving through a snow-covered city.	27.2	72.6	81.4	85.7
	A black boxer dog with a white underbelly and brown collar looks at the camera.	35.1	26.6	77.5	82.3		A dog pokes its head out from under a pile of stuff.	25.8	60.5	67.5	75.6
	A close up of a black pug.	11.6	21.1	71.0	83.5		A dog underneath a wooden beam.	22.0	38.9	63.9	81.6
	Trains amble by the rail yard.	26.2	68.8	81.9	75.4		A large green coach with a bridge in the background.	28.3	32.0	87.1	76.7
	The red train and the yellow train on the tracks.	14.7	28.3	79.8	81.6		Green bus and tan truck on a city street with a man waiting to cross the street.	34.0	17.8	79.2	79.4

Figure 3. Additional comparisons of existing metrics for captioning with respect to PAC-S on the Pascal-50S dataset. The candidate caption highlighted in green is the one preferred by humans.

Metric for Image Captioning. In *EMNLP*, 2021. 1, 2

[3] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual Information Divergence: A Unified Metric for Multimodal Generative Models. In *NeurIPS*, 2022.

1

[4] 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 *ICML*, 2021.

1

[5] Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In *CVPR*, 2022. 2

[6] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 1

Image	Candidate Captions	Evaluation Scores		Image	Candidate Captions	Evaluation Scores	
	A silver knife containing many carrots with long, green stems.	CLIP-S 0.942	PAC-S 0.854		A boy at a playground, sitting on a bench and reading a scissors .	CLIP-S 0.867	PAC-S 0.877
	A silver bowl containing many carrots with long, green stems.	CLIP-S 0.912	PAC-S 0.893		A boy at a playground, sitting on a bench and reading a book .	CLIP-S 0.847	PAC-S 0.893
	A person tries to catch a ball on a beach.	CLIP-S 0.781	PAC-S 0.798		A person riding a snowboard on a big wave.	CLIP-S 0.734	PAC-S 0.768
	A person tries to catch a frisbee on a beach.	CLIP-S 0.759	PAC-S 0.828		A person riding a surfboard on a big wave.	CLIP-S 0.733	PAC-S 0.780
	A baby horse is seen standing in between another elephant's legs.	CLIP-S 0.782	PAC-S 0.793		A large polar cat stands on rock with an open mouth.	CLIP-S 0.890	PAC-S 0.849
	A baby elephant is seen standing in between another elephant's legs.	CLIP-S 0.769	PAC-S 0.820		A large polar bear stands on rock with an open mouth.	CLIP-S 0.877	PAC-S 0.860
	Different kinds of food on a plate with a cup .	CLIP-S 0.682	PAC-S 0.758		A passenger bus is riding down the tracks.	CLIP-S 0.701	PAC-S 0.738
	Different kinds of food on a plate with a fork .	CLIP-S 0.676	PAC-S 0.789		A passenger train is riding down the tracks.	CLIP-S 0.699	PAC-S 0.777

Figure 4. Sample images from the FOIL hallucination detection dataset and corresponding evaluation scores generated by our proposed metric in comparison with CLIP-S. Captions with hallucinated objects are highlighted in red.

Image	Candidate Captions	Evaluation Scores		Image	Candidate Captions	Evaluation Scores	
	Two white dogs running.	CLIP-S 0.530	PAC-S 0.500		A man and young girl eat a meal on a city street.	CLIP-S 0.769	PAC-S 0.764
	A man riding a motorbike kicks up dirt.	CLIP-S 0.486	PAC-S 0.542		A small brown and white dog running through tall grass.	CLIP-S 0.752	PAC-S 0.820
	Little girl in bare feet sitting in a circle.	CLIP-S 0.524	PAC-S 0.431		A man jumps while snow skiing.	CLIP-S 0.512	PAC-S 0.503
	A white dog runs in the grass.	CLIP-S 0.426	PAC-S 0.456		A man is hiking on a snow-covered trail.	CLIP-S 0.464	PAC-S 0.567
	Four women wearing formal gowns pose together and smile.	CLIP-S 0.700	PAC-S 0.730		Two girls walking down the street.	CLIP-S 0.583	PAC-S 0.556
	A man in a wetsuit surfs.	CLIP-S 0.613	PAC-S 0.762		A dog lies down on a cobblestone street.	CLIP-S 0.550	PAC-S 0.562
	Boy with a red crown in a shopping cart.	CLIP-S 0.385	PAC-S 0.467		A woman is signaling to traffic, as seen from behind.	CLIP-S 0.753	PAC-S 0.767
	People stand outside near a concrete wall and a window.	CLIP-S 0.359	PAC-S 0.509		A man rides a bike through a course.	CLIP-S 0.714	PAC-S 0.800

Figure 5. Sample images from both Flickr8k-Expert and Flickr8k-CF datasets associated with the corresponding CLIP-S and PAC-S scores. The preferred caption accordingly to the human ratings is highlighted in green.