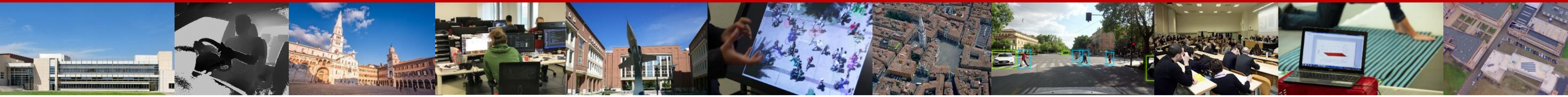
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

HIGHLIGHT PAPER, TUE-PM-266

Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation



Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara

{name.surname}@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImageLab

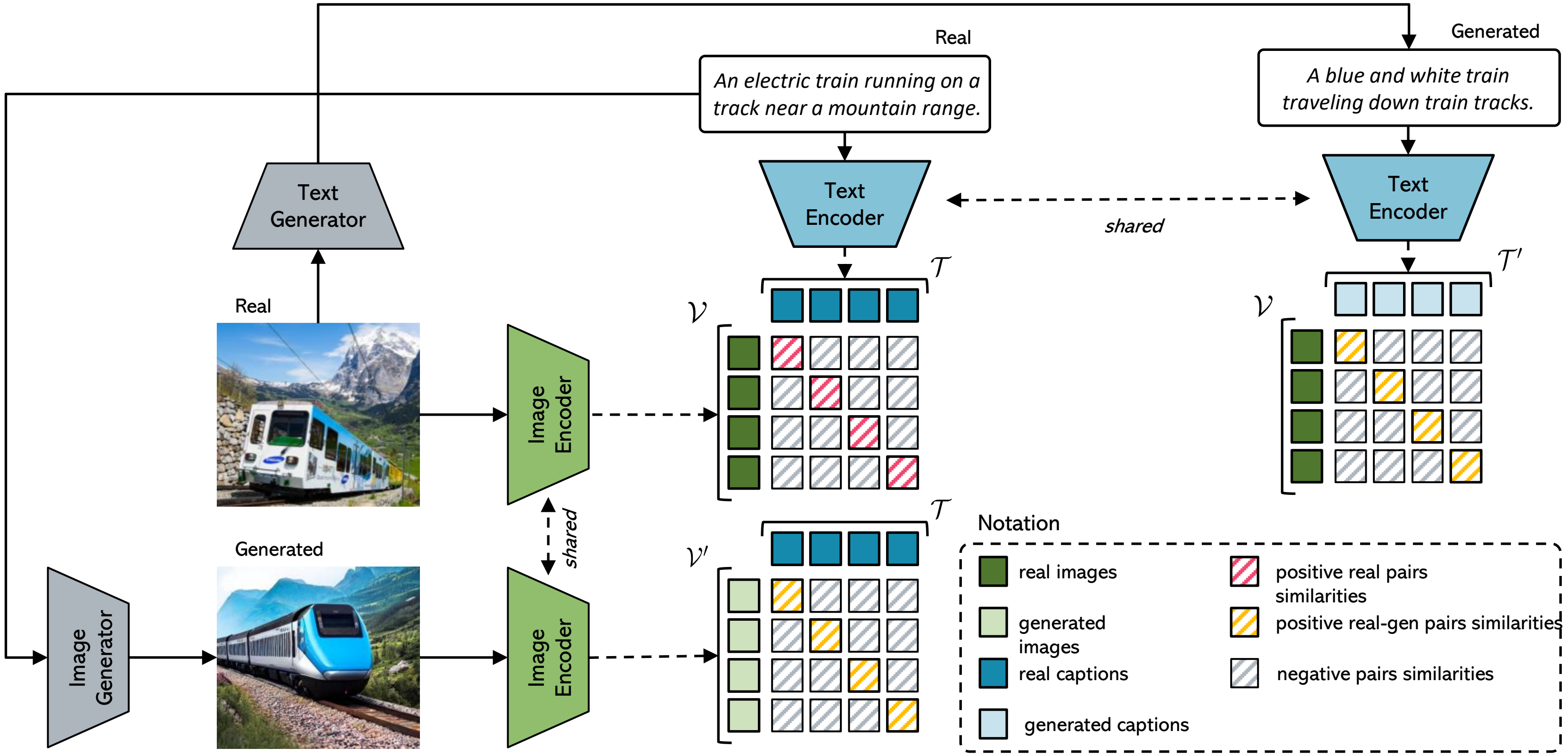
Main Contributions

- Existing metrics for image-text correspondence are either only based on **(few) human references** or multi-modal embeddings trained on **noisy data**.
- We propose a **learnable metric** for video and image captioning, which employs pre-training on **web-collected data**, **generated data for data augmentation** and the power of **human annotations**.
- Based on a ***positive-augmented training*** of a multimodal embedding space.
- Our metric outperforms previous reference-free and reference-based metrics in terms of *correlation with human judgment*.

Image	Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
	*A black cow by a person.*	9.67	14.9	0.766	0.676
	A cow walking through a field.	15.0	17.2	0.754	0.775
	*A silver bicycle is parked in a living room.*	23.1	68.6	0.686	0.853
	A silver bicycle leaning up against a kitchen table and chairs.	32.4	63.7	0.637	0.862
	*A yellow bus passes through an intersection.*	42.7	167.0	0.816	0.836
	A yellow bus is traveling down a city street just past an intersection.	33.9	94.5	0.813	0.844

Evaluation Scores

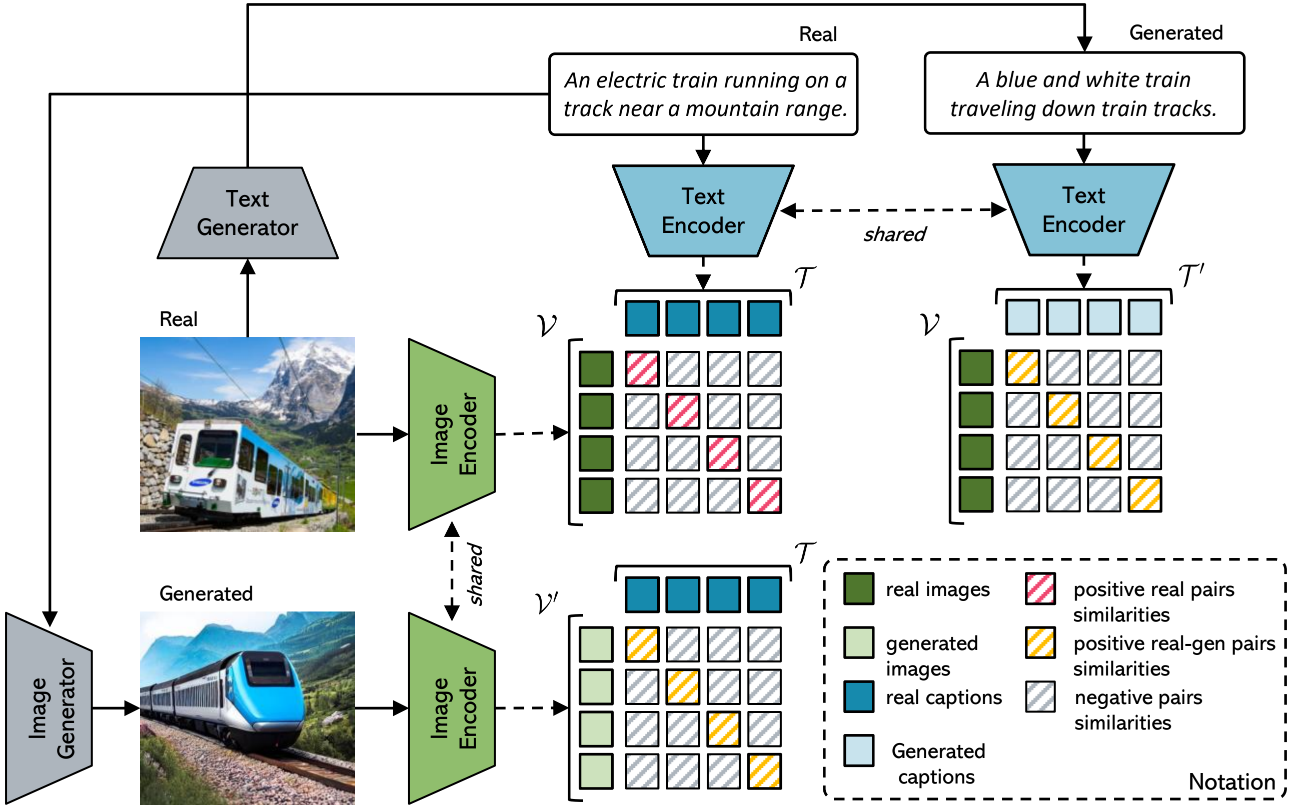
Positive-Augmented Contrastive Learning



- ***Dual-encoder architecture*** comparing the visual and textual inputs via cosine similarity.
- Usage of ***synthetic generators*** of both visual and textual data

Fine-tuning on human annotated data by taking into account ***contrastive relationship*** between real and generated matching image-caption pairs.

Positive-Augmented Contrastive Learning



$$L = L_{\mathcal{V},\mathcal{T}} + \lambda_v L_{\mathcal{V}',\mathcal{T}} + \lambda_t L_{\mathcal{V},\mathcal{T}'}$$

- A batch of N real images *V* and their corresponding captions *T*

$$\mathcal{V} = [v_1, v_2, \dots, v_N] \quad \mathcal{T} = [t_1, t_2, \dots, t_N]$$

- We adopt a symmetric infoNCE loss.

$$L_{\mathcal{V},\mathcal{T}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(v_i,t_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(v_i,t_j)/\tau)} + -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(v_i,t_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(v_j,t_i)/\tau)}$$

- We generate images thanks to Stable Diffusion[1].

$$\mathcal{V}' = [v'_1, v'_2, \dots, v'_N]$$

- We generate texts thanks to BLIP[2].

$$\mathcal{T}' = [t'_1, t'_2, \dots, t'_N]$$

1. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022

2. Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In ICML, 2022.

Image Captioning Scores[1]

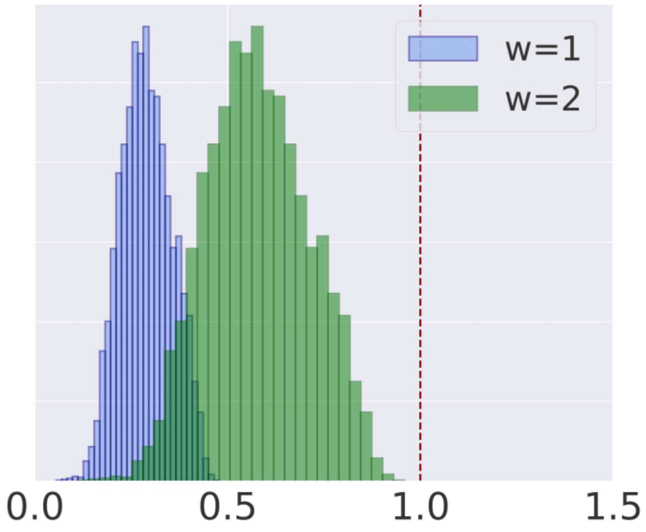
The new positive-augmented CLIP is used to compute either the image captioning score and the video score.

$$\text{PAC-Score}(t, v) = w \cdot \max(\cos(t, v), 0),$$

$$\text{RefPAC-Score}(t, v, R) = \text{H-Mean}(\text{PAC-Score}(t, v), \max(0, \max_{r \in R} \cos(c, r)))$$

In these formulas, $\cos(t, v)$ indicates the cosine similarity computed inside of the embedding space and w is a scaling factor to enhance numerical readability.

For PAC-S the value of the scaling factor is set to w=2, without affecting the ranking of the results.



1. *Hessel Jack et al. CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in EMNLP 2021*

Video Captioning Scores[1]

The new positive-augmented CLIP is used to compute either the image captioning score and the video score.

$$\text{PAC-Score}(c, V) = \frac{\text{Score}(c, V)_c + \text{Score}(c, V)_f}{2}$$

Two granularity levels:

- Coarse-grained level → $\text{Score}(c, V)_c$
- Fine-grained level → $\text{Score}(c, V)_f$

$$\text{RefPAC-Score}(c, V, r) = \frac{\text{PAC-Score}(c, V) + \max_{r \in R} \text{Score}(c, r)}{2}$$

1. *Shi, Yaya et al. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching, in CVPR 2022*

Image Captioning Correlation with Human Judgment

PAC score achieves the **best correlation with human judgment** and accuracy on all the considered image datasets, demonstrating its ***effectiveness*** compared to previously proposed metrics.

	Flickr8k-Expert		Flickr8k-CF	
	Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c
BLEU-1	32.2	32.3	17.9	9.3
BLEU-4	30.6	30.8	16.9	8.7
ROUGE	31.1	32.3	19.9	10.3
METEOR	41.5	41.8	22.2	11.5
CIDEr	43.6	43.9	24.6	12.7
SPICE	51.7	44.9	24.4	12.0
BERT-S	-	39.2	22.8	-
LEIC	46.6	-	29.5	-
BERT-S++	-	46.7	-	-
UMIC	-	46.8	-	-
TIGEr	-	49.3	-	-
ViLBERTScore	-	50.1	-	-
MID	-	54.9	37.3	-
CLIP-S	51.1	51.2	34.4	17.7
PAC-S	**53.9** (+2.8)	**54.3** (+3.1)	**36.0** (+1.6)	**18.6** (+0.9)
RefCLIP-S	52.6	53.0	36.4	18.8
RefPAC-S	**55.4** (+2.8)	**55.8** (+2.8)	**37.6** (+1.2)	**19.5** (+0.7)

	Composite	
	Kendall τ_b	Kendall τ_c
BLEU-1	29.0	31.3
BLEU-4	28.3	30.6
ROUGE	30.0	32.4
METEOR	36.0	38.9
CIDEr	34.9	37.7
SPICE	38.8	40.3
BERT-S	-	30.1
BERT-S++	-	44.9
TIGEr	-	45.4
ViLBERTScore	-	52.4
FAIEr	-	51.4
CLIP-S	49.8	53.8
PAC-S	**51.5** (+1.7)	**55.7** (+1.9)
RefCLIP-S	51.2	55.4
RefPAC-S	**52.8** (+1.6)	**57.1** (+1.7)

	Pascal-50S				
	HC	HI	HM	MM	Mean
length	51.7	52.3	63.6	49.6	54.3
BLEU-1	64.6	95.2	91.2	60.7	77.9
BLEU-4	60.3	93.1	85.7	57.0	74.0
ROUGE	63.9	95.0	92.3	60.9	78.0
METEOR	66.0	97.7	94.0	66.6	81.1
CIDEr	66.5	97.9	90.7	65.2	80.1
BERT-S	65.4	96.2	93.3	61.4	79.1
BERT-S++	65.4	98.1	96.4	60.3	80.1
TIGEr	56.0	99.8	92.8	74.2	80.7
ViLBERTScore	49.9	99.6	93.1	75.8	79.6
FAIEr	59.7	99.9	92.7	73.4	81.4
MID	67.0	99.7	97.4	76.8	85.2
CLIP-S	55.9	**99.3**	96.5	72.0	80.9
PAC-S	**60.6** (+4.7)	**99.3** (+0.0)	**96.9** (+0.4)	**72.9** (+0.9)	**82.4** (+1.5)
RefCLIP-S	64.9	**99.5**	95.5	73.3	83.3
RefPAC-S	**68.2** (+3.3)	**99.5** (+0.0)	**95.6** (+0.1)	**75.9** (+2.6)	**84.8** (+1.5)

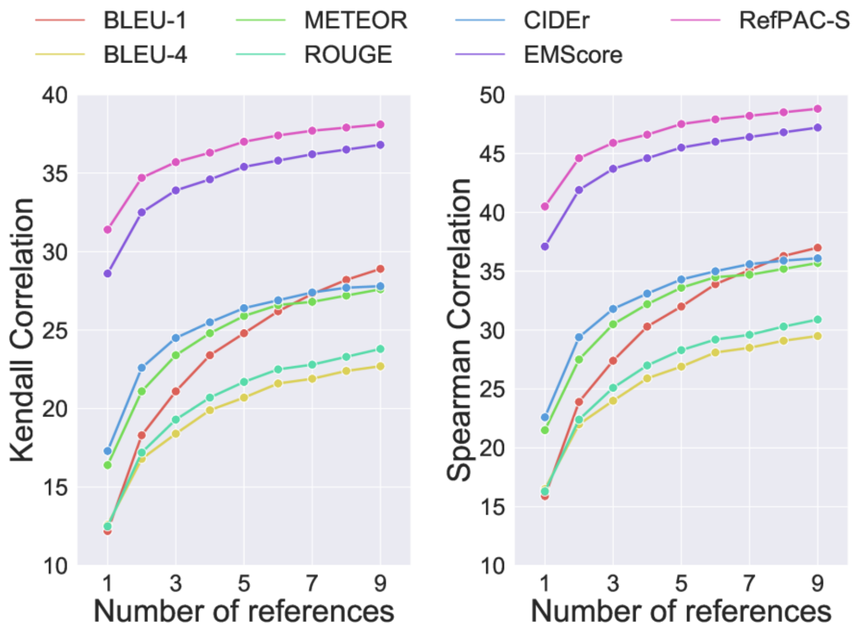
1. *Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR, 47:853–899, 2013*
2. *Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge. 2015*
3. *Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In CVPR, 2015*

Video Captioning Correlation with Human Judgment

It works well on videos too.

	No Ref		1 Ref		9 Refs	
	Kendall τ_b	Spearman ρ	Kendall τ_b	Spearman ρ	Kendall τ_b	Spearman ρ
BLEU-1	-	-	12.2	15.9	28.9	37.0
BLEU-4	-	-	12.6	16.4	22.4	29.5
ROUGE	-	-	12.5	16.3	23.8	30.9
METEOR	-	-	16.4	21.5	27.6	35.7
CIDEr	-	-	17.3	22.6	27.8	36.1
BERT-S	-	-	18.2	23.7	29.3	37.8
BERT-S++	-	-	15.2	19.8	24.4	31.7
EMScore	23.2	30.3	28.6	37.1	36.8	47.2
PAC-S / RefPAC-S	**25.1** (+1.9)	**32.6** (+2.3)	**31.4** (+2.8)	**40.5** (+3.4)	**38.1** (+1.3)	**48.8** (+1.6)

Human judgment correlation scores on the VATEX-EVAL[1] dataset. We show Kendall τ_B correlation score at varying of the number of reference captions.

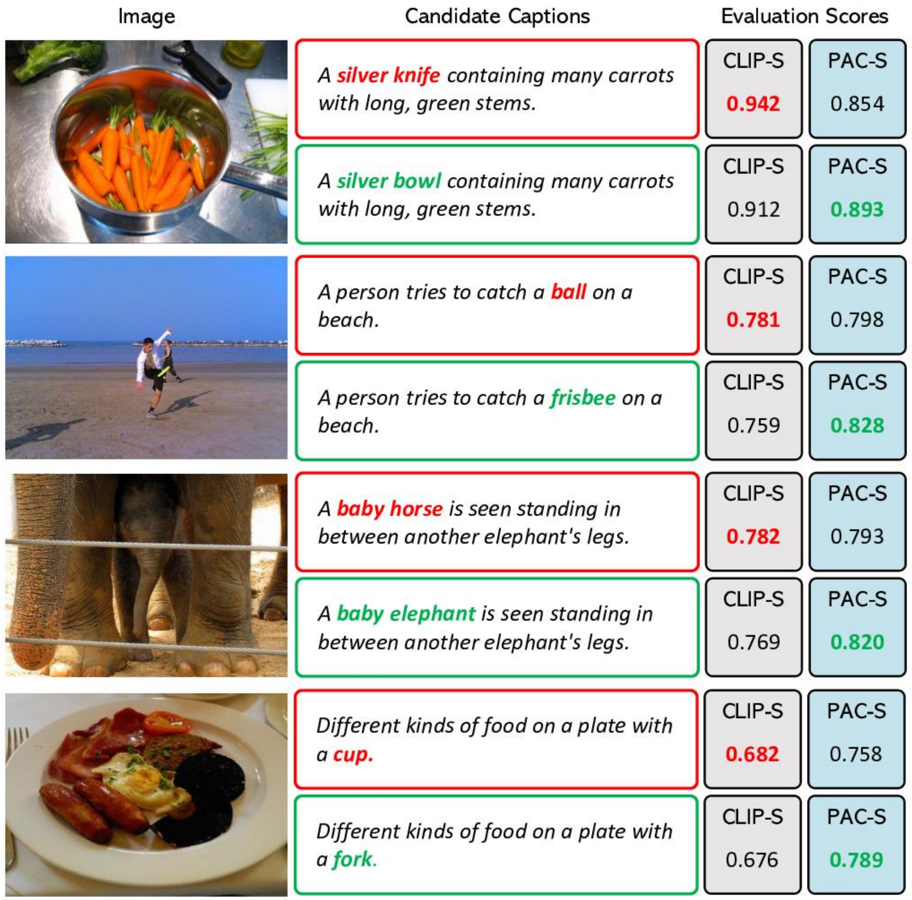


1. *Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In CVPR, 2022*

And it hallucinates less than previous metrics.

	FOIL		ActivityNet-FOIL
	Acc. (1 Ref)	Acc. (4 Refs)	Accuracy
BLEU-1	65.7	85.4	60.1
BLEU-4	66.2	87.0	66.1
ROUGE	54.6	70.4	56.7
METEOR	70.1	82.0	72.9
CIDEr	85.7	94.1	77.9
MID	90.5	90.5	-
CLIP-S	87.2	87.2	-
EMScore	-	-	89.5
PAC-S	**89.9** (+2.7)	**89.9** (+2.7)	**90.1** (+0.6)
RefCLIP-S	91.0	92.6	-
EMScoreRef	-	-	92.4
RefPAC-S	**93.8** (+2.8)	**95.2** (+2.6)	**93.5** (+1.1)

We extend our analysis to two datasets for detecting hallucinations in textual sentences, namely FOIL[2] and ActivityNet[1].



1. *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. *Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aur´elie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! Find One mismatch between Image and Language caption. In ACL, 2017.*

Different Cross-Modal Features

PAC-S achieves the best results across ***all cross-modal backbones*** and almost all datasets, overcoming correlation and accuracy scores of other metrics by a large margin.

		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	CLIP-S	51.7	52.1	34.9	18.0	-	-	81.1	90.6	-
	EMScore	-	-	-	-	24 1	31.4	-	-	90.0
	PAC-S	**54.5** (+2.8)	**54.9** (+2.8)	**35.9** (+1.0)	**18.5** (+0.5)	**26.8** (+2.7)	**34.7** (+3.3)	**82.9** (+1.8)	**91.1** (+0.5)	**90.7** (+0.7)
CLIP ViT-L/14	CLIP-S	52.6	53.0	35.2	18.2	-	-	81.7	90.9	-
	EMScore	-	-	-	-	26.7	34.7	-	-	89.0
	PAC-S	**55.4** (+2.8)	**55.8** (+2.8)	**36.8** (+1.6)	**19.0** (+0.8)	**28.9** (+2.2)	**37.4** (+2.7)	**82.0** (+0.3)	**91.9** (+1.0)	**91.2** (+2.2)
OpenCLIP ViT-B/32	CLIP-S	52.3	52.6	35.4	18.3	-	-	81.2	88.9	-
	EMScore	-	-	-	-	24.8	32.2	-	-	88.2
	PAC-S	**53.6** (+1.3)	**53.9** (+1.3)	**36.1** (+0.7)	**18.6** (+0.3)	**25.4** (+0.6)	**33.1** (+0.9)	**82.4** (+1.2)	**90.1** (+1.2)	**89.5** (+1.3)
OpenCLIP ViT-L/14	CLIP-S	54.4	54.5	36.6	18.9	-	-	82.5	92.2	-
	EMScore	-	-	-	-	27.0	35.0	-	-	90.7
	PAC-S	**55.3** (+0.9)	**55.7** (+1.2)	**37.0** (+0.4)	**19.1** (+0.2)	**27.8** (+0.8)	**36.1** (+1.1)	**82.7** (+0.2)	**93.1** (+0.9)	**91.2** (+0.5)

Qualitative Results

Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
A blue bird being held by a handler.	**35.2**	**96.3**	**80.1**	80.0
A blue bird perched on a gloved hand.	18.6	39.0	76.1	**82.1**

Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
A black boxer dog with a white underbelly and brown collar looks at the camera.	**35.1**	**26.6**	**77.5**	82.3
A close up of a black pug.	11.6	21.1	71.0	**83.5**

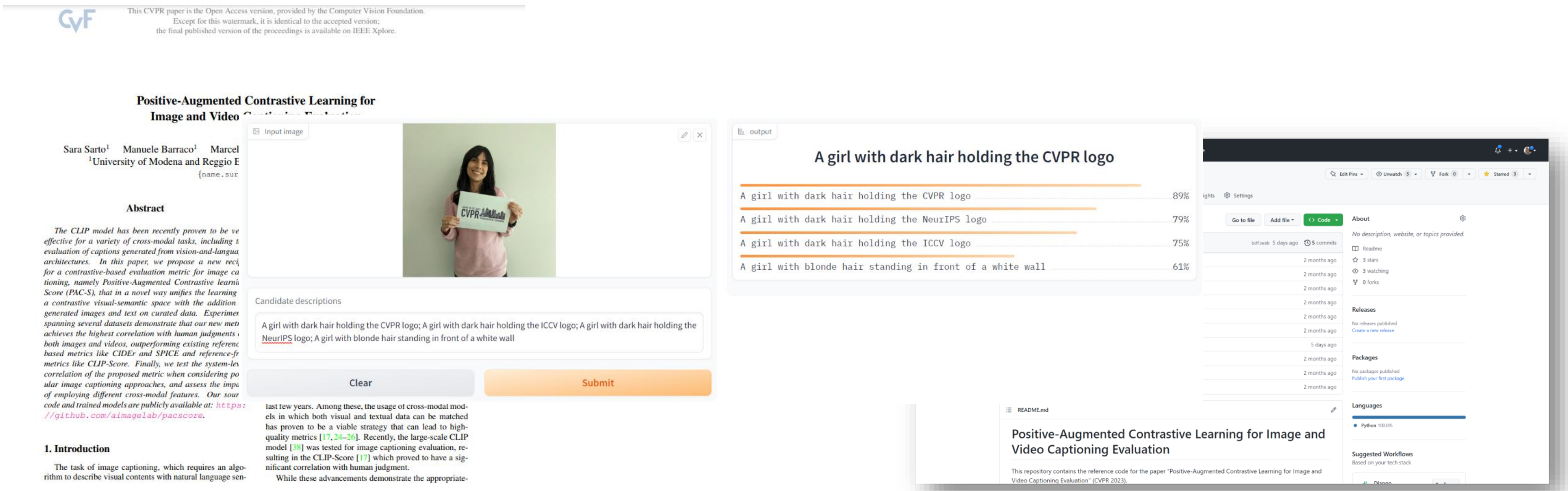
Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
Trains amble by the rail yard.	**26.2**	**68.8**	**81.9**	75.4
The red train and the yellow train on on the tracks.	14.7	28.3	79.8	**81.6**

Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
A passenger train in the snow.	26.8	**89.7**	**83.5**	83.1
A red train driving through a snow covered city.	**27.2**	72.6	81.4	**85.7**

Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
A dog pokes it's head out from under a pile of stuff.	**25.8**	**60.5**	**67,5**	75.6
A dog underneath a wooden beam.	22.0	38.9	63.9	**81.6**

Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
A large green coach with a bridge in the background	**28.3**	**32.0**	**87.1**	76.7
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**

Want to Know More?



Read the paper

Use it in your projects ☺

https://github.com/aimagelab/pacscore

TUE-PM-266

Thank you for your attention

Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara

{name.surname}@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab