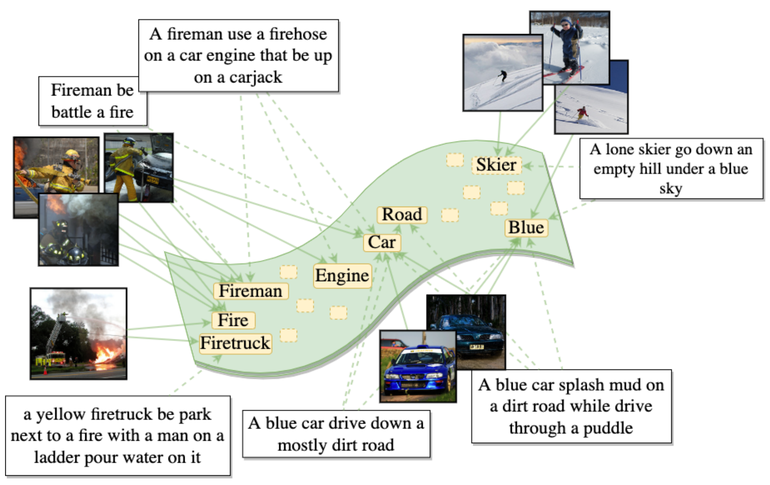
NAPReg: Nouns As Proxies Regularization For Semantically Aware Cross-Modal Embeddings

Bhavin Jawade*, Deen Dayal Mohan*, Naji Mohamed Ali, Srirangaraj Setlur, Venu Govindaraju
{bhavinja, dmohan, najimoha, setlur, govind}@buffalo.edu

WACV WAIKOLOA HAWAII JAN 3-7 • 2023 | CVF | IEEE COMPUTER SOCIETY | CUBSLab

Motivation

- **Text-to-image matching** is the most common form of cross-modal retrieval.
- Existing methods use dual encoders with an attention mechanism and a ranking loss to learn embeddings for retrieval.
- These methods **do not have explicit supervision to enforce semantic alignment** between visual regions and textual words
- We propose **NAPReg, a regularization** formulation that projects high-level semantic entities into the embedding space as **shared learnable proxies**.
- This allows the attention mechanism to learn better **word-region alignment** and build a more generalized latent representation for semantic concepts.
- Our method outperforms existing methods in cross-modal metric learning for text-image and image-text retrieval tasks.



Problem Statement

- Consider, visual features of an image $\mathbf{V=\{v_1,v_2..v_n\}}$ and textual features $\mathbf{T=\{t_1,t_2..t_m\}}$
- Fine-grained similarity between image and text can be given as:

$$S(V,T) = f(\Phi(V;\theta_i), \Psi(T;\theta_j))$$

Stacked Cross Attention (Lee et.al [2])

For each visual location, an attended combination of word representation a (i.e., the attended sentence vector a_t^i, with respect to the i^{th} image region a_t^i) is constructed as defined below:

$$s_{ij} = \frac{v_i^T t_j}{||v_i||.||t_j||}, i \in [1,n], j \in [1,m]$$

$$w_{ij} = \frac{exp(\tau.\bar{s}_{ij})}{\sum_{j=1}^{n} exp(\tau.\bar{s}_{ij})}$$

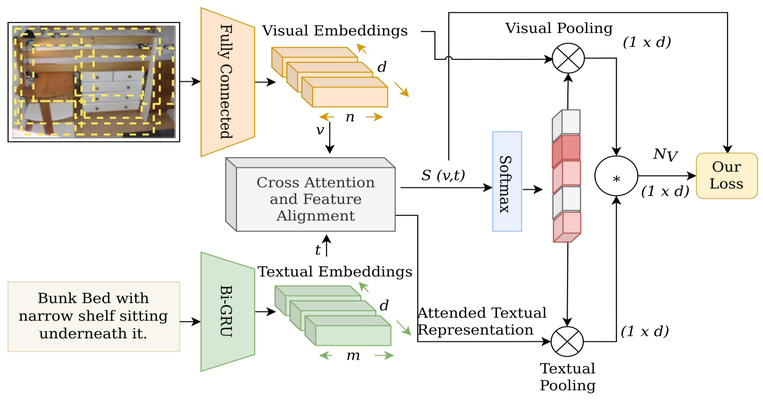
$$a_i^t = \sum_{j=1}^{m} w_{ij} * t_j$$

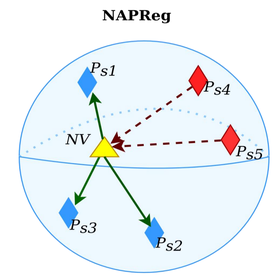
The overall cosine similarity between the image-text pair is given by:

$$S(V,T) = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_i \cdot \hat{a}_i{}^t$$

* Equal contribution authors

Method (Nouns As Proxies)





$$S = \{\hat{v}_i \cdot \hat{a}_i^t\} \forall i \in n \qquad s = \text{softmax}(S)$$

$$\mathcal{N}_\mathcal{V} = (\sum_{i=1}^{n} s_i * a_i^t) \odot (\sum_{i=1}^{n} s_i * v_i)$$

- If NS is a set of nouns in a text T_1, then P^+ is the set of **positive proxies** for those nouns, and all other proxies in N are considered **negatives**.
- $\hat{P}$ is a **learnable proxy vector** with length M where M are all unique nouns in the dataset

For $S_{np} = \hat{P} \cdot \hat{\mathcal{N}_\mathcal{V}}$

$$\mathcal{L}_{nap} = \sum_{\mathcal{X}} \left\{ \frac{1}{\alpha_1} \log\left(1 + \sum_{p \in P^+} e^{-\alpha_1 (S_{np} - \lambda_1)}\right) + \frac{1}{\beta_1} \log\left(1 + \sum_{p \notin P^+} e^{\beta_1 (S_{np} - \lambda_1)}\right) \right\}$$

$$\mathcal{L}_{pair} = \sum_{\mathcal{X}} \left\{ \frac{1}{\alpha_2} \log\left(1 + \sum_{(v,t) \in D^+} e^{-\alpha_2 (\bar{S} - \lambda_2)}\right) + \frac{1}{\beta_2} \log\left(1 + \sum_{(v,t) \notin D^+} e^{\beta_2 (\bar{S} - \lambda_2)}\right) \right\}$$

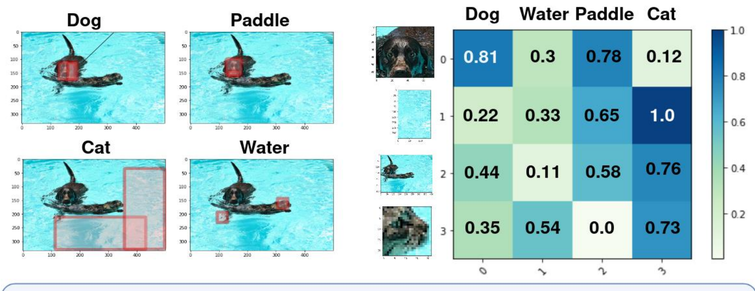
$$\mathcal{L} = \mathcal{L}_{pair} + \gamma \mathcal{L}_{nap}$$

Qualitative Results

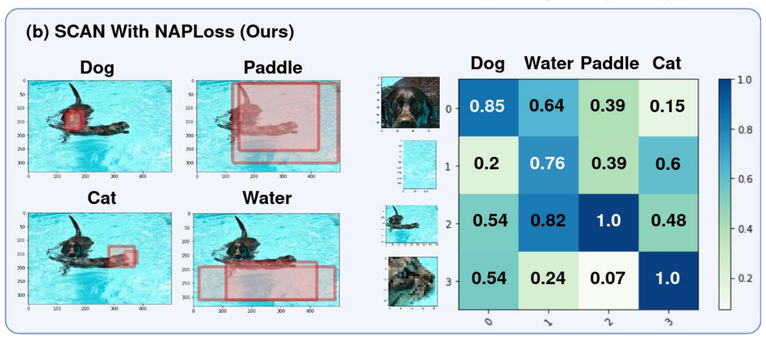
Proxies: Dog | Cat | Paddle | Water

A black dog paddle behind a black cat in a body of water .

(a) SCAN Without NAPReg



(b) SCAN With NAPLoss (Ours)



- Shows the **top 2 regions attended by each proxy word** in the image on Left and **heatmap between the similarity of selected visually relevant regions and the word proxies** on the right
- The similarity score of the visual region containing the cat and the dog is highest for the corresponding word in the text
- The magnitude of the scores has also increased in comparison to the model without the proposed regularization

Ablation Study

	Text-to-Image		Image-to-Text	
Gamma	**R@1**	**Rsum**	**R@1**	**Rsum**
0	37.7	184.3	52.1	226.4
0.1	37.6	184.9	54.4	227
0.2	38.1	186.4	54.5	228.4
0.3	**39.2**	**188**	**56.2**	**229.7**
0.4	38.3	186.5	54.8	228.7

Quantitative Results

Recall@K(%) performance on **MSCOCO dataset**

Method	Loss	Text-to-Image			Image-to-Text		
		R@1	R@5	R@10	R@1	R@5	R@10
MSCOCO - 1K Evaluation							
IMRAM (Full)	Triplet	61.7	89.1	95	76.7	95.6	98.5
GSMN (Sparse)	Triplet	60.4	88.7	95	76.1	95.6	98.3
PFAN (i2t)	Triplet	53.0	84.5	92.6	70.7	94.1	97.8
SCAN (i2t) [2]	Triplet	54.4	86	93.6	69.2	93.2	97.5
SHAN	Triplet	62.6	89.6	95.8	76.8	96.3	98.7
VSE∞	Triplet	61.7	90.3	95.6	78.5	96.0	98.7
UWML (i2t) [1]	Polyloss	56.8	86.7	93	71.1	93.7	98.2
NAAF (BiGRU)	Triplet	61.3	90.6	96.0	76.8	95.2	98.2
SGRAF (SGR) [3]	Triplet	61.4	89.3	95.4	78	95.8	98.2
SCAN (i2t)	**Ours**	**58.6**	**87.5**	**93.8**	**71.6**	**94.5**	**98.2**
SGRAF (SGR)	**Ours**	**63.3**	**90**	**95.6**	**78.7**	**96.2**	**98.8**
SCAN (i2t+t2i)	**Triplet**	58.8	88.4	94.8	72.7	94.8	98.4
SGRAF (SGR+SAF)	**Triplet**	63.2	90.7	96.1	79.6	96.2	98.5
SGRAF (SGR+SAF)	**Ours**	**66.9**	**91.6**	**96.5**	**81.9**	**97.5**	**99.2**
MSCOCO-5K Evaluation							
IMRAM (Full)	Triplet	39.7	69.1	79.8	53.7	83.2	91
SCAN (i2t) [2]	Triplet	34.4	64.2	75.9	46.4	77.4	87.6
UWML (i2t) [1]	Polyloss	34.4	64.2	75.9	46.9	77.7	87.6
SGRAF (SGR) [3]	Triplet	40.2	-	79.8	56.9	-	90.5
SCAN (i2t)	**Ours**	**36.5**	**66**	**77.6**	**48**	**78.6**	**88.3**
SGRAF (SGR)	**Ours**	**41.7**	**71.2**	**81.5**	**58**	**85.1**	**91.6**
SCAN (i2t+t2i)	Triplet	38.6	69.3	80.4	50.4	82.2	90.0
SGRAF (SGR+SAF)	Triplet	41.9	-	79.8	57.8	-	91.6
SGRAF (SGR+SAF)	**Ours**	**43**	**72.1**	**82.4**	**59.8**	**86**	**92.6**

Recall@K(%) performance on **Flickr30K dataset**

Method	Loss	Text-to-Image			Image-to-Text		
		R@1	R@5	R@10	R@1	R@5	R@10
BFAN	Triplet	50.8	78.4	85.8	68.1	91.4	95.9
IMRAM	Triplet	53.9	79.4	87.2	74.1	93	96.6
GSMN (Sparse)	Triplet	53.9	79.7	87.1	71.4	92	96.1
PFAN (i2t)	Triplet	45.7	74.7	83.6	67.6	90.0	93.8
SCAN (i2t) [2]	Triplet	43.9	74.2	82.8	67.9	89	94.4
SMFEA	Triplet	54.7	82.1	88.4	73.7	92.5	96.1
SHAN	Triplet	55.3	81.3	88.4	74.6	93.5	96.9
VSE∞	Triplet	56.4	83.4	89.9	76.7	94.2	97.7
UWML (i2t) [1]	Polyloss	47.5	75.5	83.1	69.4	89.4	95.4
NAAF (BiGRU)	Polyloss	55.5	81.0	87.9	75.9	93.6	97.7
SGRAF (SGR) [3]	Triplet	56.2	81	86.5	75.2	93.3	96.6
SCAN (i2t)	**Ours**	**51.4**	**77.6**	**85.7**	**70.8**	**90.9**	**95.3**
SGRAF (SGR)	**Ours**	**58.3**	**83.1**	**89.2**	**79.2**	**95.3**	**97.7**
SGRAF (SGR+SAF)	Triplet	58.5	83.0	88.8	77.8	94.1	97.4
SCAN (i2t+t2i)	Triplet	48.6	77.7	85.2	67.4	90.3	95.8
SGRAF (SGR+SAF)	**Ours**	**60**	**84.1**	**90.2**	**79.6**	**95.6**	**98**

References

1. Wei, Jiwei, et al. "Universal weighting metric learning for cross-modal matching." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
2. Lee, Kuang-Huei, et al. "Stacked cross attention for image-text matching." Proceedings of the European conference on computer vision (ECCV). 2018.
3. Diao, Haiwen, et al. "Similarity reasoning and filtration for image-text matching." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 2. 2021.

University at Buffalo The State University of New York
Department of Computer Science and Engineering

Github
github.com/bhavinjawade/NAPReq