Supplemental Material
NAPReg: Nouns As Proxies Regularization for Semantically Aware Cross-Modal Embeddings

1. Ablation on number of proxies

Table 1 shows an abaltion study on the effect of minimum frequency K used to filter proxies from the captions to the overall performance on Flickr8K. It can be observed that by just using all extracted nouns (*i.e.* minimum frequency 1, and proxy count 3874), we get a performance close to the most optimal proxy count (*i.e.* 2444). Filtering with a minimum frequency of 2 removes some redundant, misspelled, or rarer nouns giving a 3.1 points improvement over minimum frequency 1 on Rsum for image-to-text retrieval. The most important observation to note is that even while just using 1102 proxies (minimum frequency of 8), we get a significant improvement over vanilla SCAN, thereby substantiating the value of NAPReg regularization for cross-modal retrieval.

Table 1: Ablation to evaluate effect of number of proxies on matching performance on Flickr8k dataset

Min Freq.	**No. Of Proxies**	**Text-to-Image**		**Image-to-Text**	
		R@1	Rsum	R@1	Rsum
1	3874	39.2	189.7	54.5	226.6
2	2444	**39.2**	188.0	**56.2**	**229.7**
3	1950	39.1	**189.8**	53.9	227.1
5	1438	38.5	187.9	55.0	227.7
8	1102	38.5	188.7	53.4	224.2
No NAPReg	0000	32.3	168.9	51.2	216.0

2. Ablation on backbone

In this section, we present the performance of our proposed regularization method for cross modal retrieval with different textual backbones. In particular, we show augmentation of our method with a **transformer** based text encoder. As can be observed in table 2, with a **BERT** backbone, NAPReg provides a 4.0% improvement on Text to Image retrieval and 4.4% improvement on R@1 for Image to Text retrieval with respect to the baseline (SGRAF with BiGRU and triplet loss). In comparison to NAAF'21 [7], we show 1.5% improvement on R@1 on Text to Image retrieval and 0.3 % improvement on R@1 on Image to Text. Further, we also present the ensembled backbone (BERT + Bi-GRU) results, which shows 2.5% improvement over NAAF on Text to Image retrieval R@1. We also show performance improvement, when using GloVe textual embeddings on SCAN architecture. These experiments demonstrate the robustness of our proposed regularization towards different textual encoder architectures.

3. Comparison with Proxy Anchor Loss

Proxy based metric learning was introduced first in the domain of uni-modal (image-to-image) metric learning by Proxy-NCA [3]. This method, in its standard setting, assigns a unique proxy per class. Consider sample $X = (x_1, y_1)$ in which x_1 is the data point and y_i is corresponding label associated with the data point. In Proxy-NCA, the data point is attracted towards the class proxy corresponding to y_i and separated from all the other proxies (corresponding to other classes). Recently, inspired by [5], [1] proposed Proxy Anchor Loss, which introduced an improvement over the Proxy-NCA by considering the relative hardness of samples. The formulation of proxy anchor loss is given by

$$\mathcal{L}_{AP} = \sum_{\mathcal{P}} \left\{ \frac{1}{P^+} \log\left(1 + \sum_{x \in X_p^+} e^{-\alpha_1 (S_{np} - \lambda_1)}\right) + \frac{1}{P} \log\left(1 + \sum_{x \in X_p^-} e^{\alpha_1 (S_{np} + \lambda_1)}\right) \right\} \tag{1}$$

Where λ_1 is the margin, α_1 is the scaling factor and P^+ refers to the positive proxies. From the above formulation one can note that, the proxies corresponding to the individual classes in the dataset act as anchor points. The relationship between a data point to the corresponding positive anchor point (proxy) is mutually exclusive, as a data point cannot belong to two different classes. Adapting existing proxy based uni-modal metric learning to cross-modal retrieval is non-trivial due to the class dependent nature of the way the proxies are defined. The proposed regularization NAPReg overcomes this challenge by defining a proxy cor-

Table 2: Ablation demonstrating performance using BERT backbone

Method	Backbone	Loss	Text to Image			Image To Text		
			R@1	R@5	R@10	R@1	R@5	R@10
SCAN$_{i2t}$	BiGRU	Triplet	43.9	74.2	82.8	67.9	89.0	94.4
SCAN$_{i2t}$	BiGRU	**Ours**	51.4	77.6	85.7	70.8	90.9	95.3
SCAN$_{i2t}$	GloVe	**Ours**	**54.6**	**81.7**	**88.4**	**74.9**	**92.6**	**96.6**
SGRAF	BiGRU	Triplet	58.5	83.0	88.8	77.8	94.1	97.4
NAAF	BiGRU	Triplet	58.9	83.3	89.0	78.3	94.1	97.7
SGRAF	BiGRU	**Ours**	**60.0**	**84.1**	**90.2**	**79.6**	**95.6**	**98.0**
NAAF	GloVe	Triplet	61.0	85.3	90.6	81.9	**96.1**	**98.3**
SGRAF	BERT	**Ours**	**62.5**	**87.1**	**92.1**	**82.2**	95.5	98.0
SGRAF	Ensemble	**Ours**	**63.5**	**87.8**	**92.6**	**82.2**	**96.4**	**98.3**

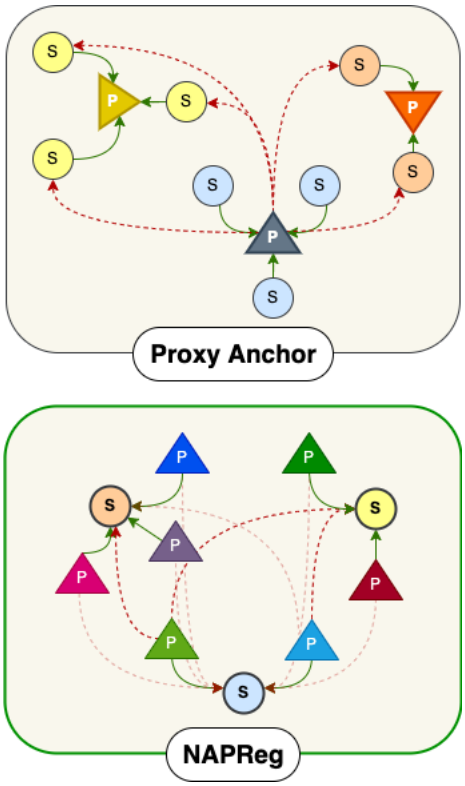

Figure 1: Comparison with proxy anchor loss [1] formulation. Here, P denotes a proxy and S denotes a data point/sample. Proxy anchor assigns each proxy as an anchor, whereas NAPReg (ours) formulation assigns samples as anchors. Refer to section 3 for related discussion.

responding to each noun occurring more than K times in the dataset. When defining proxies based on nouns, adopting a Proxy Anchor based formulation leads to sub optimal performance. This is due to the fact that in cross-modal scenarios, single data point (image) may have multiple positive proxies (nouns) associated with them, unlike the mutually exclusive data point - proxy relationship that is observed in the uni-modal metric learning problem. So, in our proposed regularization term, the formulation is designed to make each individual data point act as an anchor point. The objective of the proposed term tries to reduce the distance between the positive proxies associated with each data point and increase the separation to the other proxies (as shown in figure 1. The final formulation of proposed regularization is given by

$$\mathcal{L}_{nap} = \sum_{\mathcal{X}} \left\{ \frac{1}{\alpha_1} \log\left(1 + \sum_{p \in P^+} e^{-\alpha_1 (S_{np} - \lambda_1)}\right) + \frac{1}{\beta_1} \log\left(1 + \sum_{p \notin P^+} e^{\beta_1 (S_{np} - \lambda_1)}\right) \right\} \tag{2}$$

4. Qualitative Analysis on Flickr30K

Here, we provide a qualitative comparison of NAPLoss against Polyloss [6] . The first five rows in figure 2 show examples where NAPLoss correctly retrieved the image for a given caption at top-1. Rows 6-9 show examples where NAPLoss correctly retrieved in the top 5 whereas Polyloss[6] did not retrieve the image in the top 5. Rows 10-12 show examples where NAPLoss correctly retrieved images in top-1, whereas Polyloss[6] retrieved in top-5.

5. Other Hyper-parameters

As discussed in the main paper, we use the same hyper-parameter values as described in [5]. For completeness and reproducibility we mention these hyper-parameter values here: $\lambda_1, \lambda_2 = 0.5$, positive scale $\alpha_1, \alpha_2 = 2.0$, negative scale $\beta_1, \beta_2 = 40.0$ and margin 0.1.

6. Discussion on Large Scale Vision-Language Models

Large scale vision-language models[4] have shown superior performance on most vision-language tasks such as

Figure 2: Qualitative results of top 5 retrieval on flickr30k - (green represents correct retrieval, red represents incorrect retrieval)

image-to-text retrieval, visual grounding, ref-retrieval, visual question answering etc. Majority of these methods use large scale pre-training where the goal is to optimize a constrastive objective. Since our proposed regularization creates shared semantic proxies, these could be easily adapted to enhance the optimization objective for these methods. For instance, one potential way to incorporate NAPreg into CLIP like models is to only use the individual modality embedding to generate the noun context vector. This is due to the fact that these contrastively trained models do not use cross attention during evaluation. In other methods such as [2] which use cross attention modules, an alignment-based approach for creating noun context vector similar to that of ours can incorporate the proposed regularization.

References

[1] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020.

[2] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021.

[3] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017.

[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. *CoRR*, abs/2103.00020, 2021.

[5] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

[6] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[7] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15661–15670, June 2022.