Emphasizing Complementary Samples for Non-literal Cross-modal Retrieval (Supplementary Material)

Christopher Thomas*
Columbia University
New York, NY
christopher.thomas@columbia.edu

Adriana Kovashka
University of Pittsburgh
Pittsburgh, PA
kovashka@cs.pitt.edu

1. Introduction

We provide supplemental results to those in our main text. We first provide results further comparing our method to [5] using a different retrieval loss than triplet (i.e. **angular loss**) for direct comparison with that method. We show that **our method continues to outperform [5] under its strongest configuration**, even without finetuining our hyperparameters in any way. We also compare our method against this method using top-3 (rather than just top-1) and demonstrate that our conclusions do not change (i.e. our method continues to outperform significantly).

Next, we show an additional way of comparing our method to the strongest baseline using **mean rank** of the correct sample as judged by our method. We show that the trends among the methods remain and that our method continues to outperform all other methods. We then provide a visualization of the distribution of $|\alpha_X - \alpha_Y|$ (as computed by our method) across the different datasets tested, showing that our different weighting measures emphasize different aspects of each dataset. We finally provide a discussion of the motivation of our measures and how to properly evaluate cross-modal retrieval for abstract samples.

2. Comparison to Thomas and Kovashka [5]

The experiments in our main text are all shown using triplet loss, the most common loss used in cross-modal retrieval [5]. However, [5] also test their method with angular loss and show stronger results than with triplet. We chose to omit angular in our main text for three reasons: 1) most baselines (e.g. PVSE [4], Mithun [2], HAL [1]) use some variation of triplet; 2) [5]'s contribution is not angular loss, but new loss constraints that impose structure on the learned space; and 3) our weighting measures can be applied to any retrieval loss.

For fairness and closer comparison to [5] however, we retrained our model using angular loss. To make the shift to this setting particularly challenging for our method, we performed no hyperparameter tuning of our method. We show the result of training our method using angular loss in Table 1, using both Recall@1 and Recall@3. **We outperform [5] with angular loss**, without tuning the diversity / discrepancy combination weights or any other hyperparameters in our method. On Politics we achieve 3% gain over Thomas; 2-4% gain on ConcCap. Interestingly, we observe that as we move from Recall@1 to Recall@3 the benefit of our method becomes even more pronounced over the baseline.

3. Results using Mean Rank

We wanted to verify that our conclusions held if we used the rank metric rather than recall. The rank metric calculates the average rank of the correct retrieval result within each multiple choice task. Lower is better and a correct retrieval would have rank 0, since the retrieved result would be first (i.e. counting from 0). We compared our best non-combination method (either Ours-Discrepancy or Ours-Diversity) to the strongest baseline per dataset in Table 2. We observe that our non-combination methods are weaker than their combined counterparts (see Table 2 in the main text). Specifically, [2] outperforms our individual weighting methods for COCO, but when they are combined either with Ours-Combined-Val or Ours-Combined-Stats our method outperforms [2] even for COCO. We note that COCO features highly aligned captions and images, where the caption literally describes the content of the image. This is not our focus. Instead, we are interested in more abstract, real-world data like Politics, GoodNews, or Conceptual Captions. For these datasets, we observe that our best-performing individual weighting measures significantly outperform the best baseline. This result shows that the trends shown in the main text continue to hold for alternative metrics like mean rank.

*Work done while at University of Pittsburgh

	COCO (100-Way)		**GoodNews** (5-way)		**Politics** (5-way)		**ConcCap** (20-way)	
Method	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**
	Recall@1							
[5]	0.6932	0.6925	0.8867	0.8864	0.6264	0.6314	0.7865	0.7924
Ours	**0.6988**	**0.7087**	**0.8881**	**0.8877**	**0.6554**	**0.6584**	**0.8095**	**0.8357**
	Recall@3							
[5]	0.9051	0.9055	0.9691	0.9691	0.8923	0.8919	0.9246	0.9249
Ours	**0.9178**	**0.9173**	**0.9833**	**0.9843**	**0.9057**	**0.9053**	**0.9429**	**0.9428**

Table 1. Retrieval Recall@1 and Recall@3 with Ours-Combined-Stats vs the complete version of [5] using PVSE [4]; both use angular loss. We significantly outperform [5] **even without tuning hyperparameters** of our method (diversity / discrepancy combination weights). We observe that the benefit of our method over [5] becomes even more pronounced for Recall@3.

	COCO (100-Way)		**GoodNews** (5-way)		**Politics** (5-way)		**ConcCap** (20-way)	
Method	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**
	Mean Rank							
Best baseline	**15.1556**	**15.2450**	0.3411	0.3326	0.9861	0.9779	2.3503	2.3811
Best non-combination Ours-Discrepancy or Ours-Diversity	15.6370	15.7489	**0.3167**	**0.3055**	**0.9465**	**0.9376**	**2.2607**	**2.3152**

Table 2. We compare the best of our non-combination methods with the best baseline using mean rank (lower is better, starting at zero). We observe that our method outperforms the baselines significantly for the more abstract datasets that we target. We note that our non-combination methods are outperformed by [2] for COCO for mean rank (as well as top-1 accuracy / recall@1) in the main text. However, **our combination methods outperform it for COCO** (see Table 2 in the main text).

4. $|\alpha_X - \alpha_Y|$ per Dataset

We calculated $|\alpha_X - \alpha_Y|$ using both of our proposed weighting measures (discrepancy and diversity) for each sample in each dataset. We show the distribution of these weights per dataset in Figure 1. We observe a correlation with Table 2 in the main text and our individual (non-combination) methods (Ours-Discrepancy and Ours-Diversity). A "thin" (small variance) plot for Diversity correlates with strong performance on that dataset. For Discrepancy, large variance correlates with strong performance. We observe higher X-axis values (not visible) at the peak for Politics (0.6 vs 0.2 for COCO) which suggests our sample weighting is especially important for this dataset.

5. Motivation of Our Measures

We hypothesize real-world image-text data falls along a spectrum from literal (text literally describes image) to abstract (image/text usage figurative). We propose two weighting strategies that measure *how* abstract an image-text sample is and focus the model on abstract samples during training. **Diversity** measures how visually/textually diverse the first-degree neighbors of a text/image sample are (i.e. diversity in the other modality). That is to say that samples with high *diversity* suggest the same semantic concept shares many visual expressions (e.g. "justice" has visually distinct portrayals) compared to a concrete concept (e.g. car). On the other hand, **Discrepancy** uses second-degree neighbors to check how much the meaning of a sample changes two degrees removed from the sample (i.e. discrepancy in the same modality). *Discrepancy* quantifies the number of senses of a sample using second-degree neighbors and their relation to the sample (far means concept can be interpreted multiple ways). Our measures explicitly attempt to measure the abstractness of a sample. While prior hard/soft negative mining methods also attempt to emphasize certain samples (i.e. samples whose embeddings remain far from their cross-modal counterpart), these methods may prioritize noisy rather than abstract samples (e.g. image-text wrongly paired). We compare against a number of such methods in our main text (i.e. soft negative mining / weighting) and significantly outperform them.

6. Evaluation Metrics for Non-literal Cross-modal Retrieval

We believe that finding a good eval metric for abstract samples remains an open issue, because of the subjectivity of the task. For example, in traditional "literal" cross-modal retrieval, leveraging the fact that the image and caption were paired is sufficient, since the caption literally de-

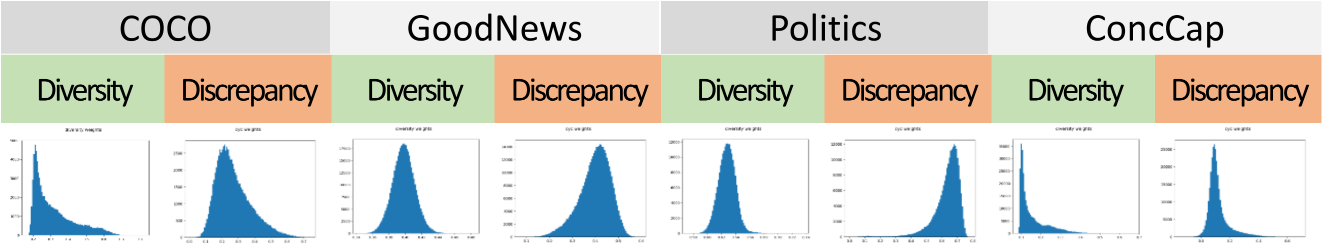

Figure 1. We show the distribution of $|\alpha_X - \alpha_Y|$ for both diversity and discrepancy per dataset. We observe a correlation between the distribution of weights with the method performance in Table 2 in the main text.

scribes the image contents and it is less likely another image (or text) would exactly describe the same image. In contrast, for abstract "non-literal" retrieval the problem is more challenging and subjective. For example, a passage about Isaac Newton could be paired with an image of an apple, an apple tree, or Newton himself. Thus, the challenge with traditional Recall@k over the entire test set is that abstract retrieval is subjective, i.e. should a method be penalized since it ranks an image with a scale of justice higher than Lady Justice for text about justice? The SemanticMap metric [3] has recently been proposed to give credit to retrieval methods based on the semantics of the retrieved image, rather than only looking at whether the text query was originally paired with the image. Unfortunately, SemanticMap requires labeled image/text class data while our data (ConcCap, Politics, and GoodNews) lacks such labels.

Both PVSE [4] and [5] target ambiguous or abstract retrieval and both exclusively use Recall. However, [5] attempt to address the subjectivity problem by formulating mini-retrieval tasks and perform Recall@1. The smaller retrieval tasks reduce the likelihood subjective samples are in each task. We adopt this formulation and report on the same mini-retrieval tasks as [5].

References

[1] Fangyu Liu, Rongtian Ye, Xun Wang, and Shuaipeng Li. Hal: Improved text-image matching by mitigating visual semantic hubs. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020. 1

[2] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 1, 2

[3] Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, Alessandro Calefati, and Faisal Shafait. Do cross modal systems leverage semantic relationships? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[4] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. 1, 2, 3

[5] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3