Dynamic Soft Labeling for Visual Semantic Embedding

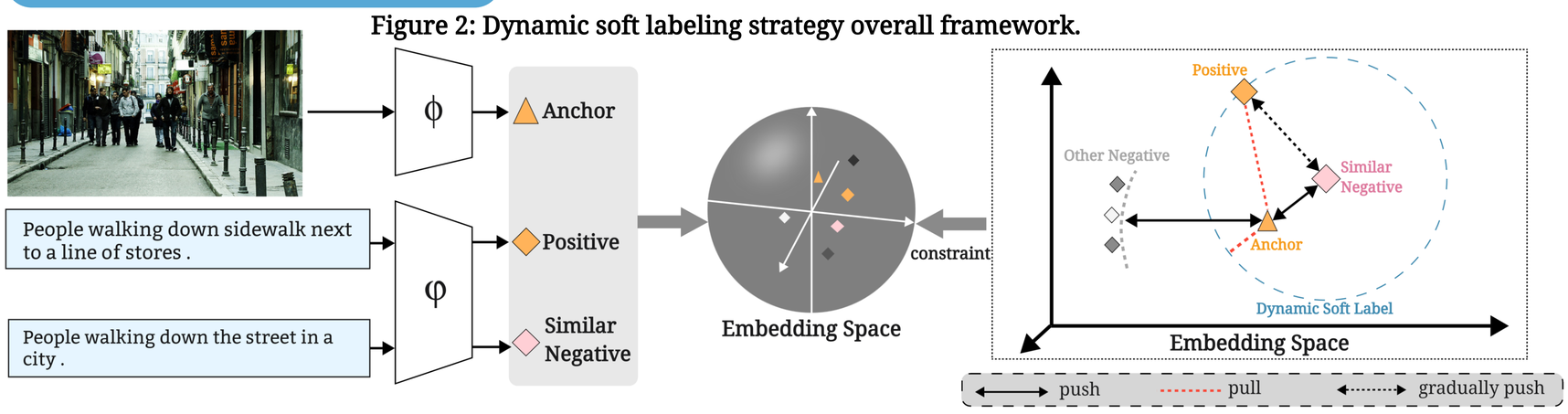
Jiaao Yu, Yunlai Ding *, Yuezun Li, Junyu Dong
College of Computer Science and Technology, Ocean University of China, Qingdao, Shandong, China
Sanya Oceanographic Institution, Ocean University of China, Qingdao, Shandong, China

INTRODUCTION

- We propose a dynamic soft labeling strategy that is simple, efficient, plug-and-play, and assigns optimization objectives to all sample pairs to fully learn semantic feature representations and distinguish similar negative samples.
- We propose Soft Negative Alignment Loss (SNAL) and Step-wise Negative Pair Decoupling Loss (SNDL). During training, the dynamic correlation of intra-modal samples is used to optimize the inter-modal correlation, and the distance between intra-modal samples in common embedding space is gradually distanced by adaptive constraint weights.
- We conduct extensive experiments on two public dataset benchmarks Flickr30K and MS-COCO to demonstrate that our strategy has performance-enhancing effects in visual semantic embedding methods.

METHODOLOGY

Figure 2: Dynamic soft labeling strategy overall framework.



- This strategy constrains the distribution of samples in the common feature space to train the encoders.
- The strategy constrains the anchors and negative samples to have a reasonable degree of correlation with each other as well.
- The strategy relatively distances the positive and negative samples from each other.

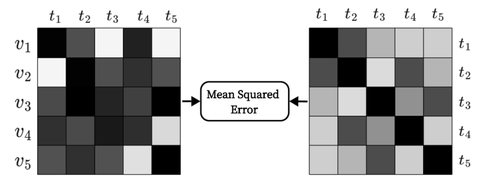
How To Constraint ?

$$Loss = \text{Triplet ranking loss with Hard Negative mining.} + \lambda \text{Soft Negative Alignment Loss (SNAL).} + \gamma \text{Stepwise Negative Decoupling Loss (SNDL).}$$

Triplet ranking loss with Hard Negative mining.

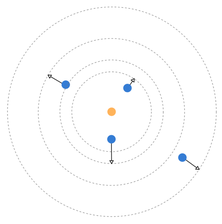
$$\mathcal{L}_{tri-HN} = [\beta + s(v_p, \hat{t_n}) - s(v_p, t_p)]_+ + [\beta + s(\hat{v_n}, t_p) - s(v_p, t_p)]_+$$

Soft Negative Alignment Loss (SNAL).



$$\mathcal{L}_{snal} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[s(v_i, t_j) - s(t_i, t_j)]^2$$

Stepwise Negative Decoupling Loss (SNDL).



$$\mathcal{L}_{sndl} = ||(1 - E^\intercal E) \cdot E^\intercal E||_F^2$$

EXPERIMENTS

Data Split	MS-COCO 5-fold 1K Test							MS-COCO 5K Test						
Eval Task	Image→Text			Text→Image				Image→Text			Text→Image			
Method	R@1	R@5	R@10	R@1	R@5	R@10	RSUM	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
ResNet-152														
VSE++ [8]	64.6	90.0	95.7	52.0	84.3	92.0	478.6	41.3	-	81.2	30.3	-	72.4	-
SCO [13]	69.9	92.9	97.5	56.7	87.5	94.8	499.3	42.8	72.3	83.0	33.1	62.9	75.5	369.6
ResNet-101,Faster-RCNN														
SCAN† [17]	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
CVSE [32]	74.8	95.8	98.3	59.9	89.4	95.2	512.7	-	-	-	-	-	-	-
LIWE [34]	73.2	95.5	98.2	57.9	88.3	94.5	507.6	-	-	-	-	-	-	-
VSRN† [18]	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
IMRAM [2]	76.7	95.6	98.5	61.7	89.1	95.0	516.6	53.7	83.2	91.0	39.7	69.1	79.8	416.5
CAMERA† [27]	77.5	96.3	98.8	63.4	90.9	95.8	522.7	55.1	82.9	91.2	40.5	71.7	82.5	423.9
DCPA [30]	78.9	95.5	98.6	64.7	90.8	96.3	524.8	53.5	82.4	90.2	40.4	71.0	82.0	419.5
VSRN++† [19]	77.9	96.0	98.5	64.1	91.0	96.1	523.6	54.7	82.9	90.9	42.0	72.2	82.7	425.4
MV-VSE[21]	80.4	96.6	99.0	64.9	91.2	96.0	528.1	59.1	86.3	92.5	42.5	72.8	83.1	436.3
CHAN[26]	81.4	96.9	98.9	66.5	92.1	96.7	532.6	59.8	87.2	93.3	44.9	74.5	84.2	443.9
GPO [3]	79.7	96.4	98.9	64.8	91.4	96.3	527.5	58.3	85.3	92.3	42.4	72.7	83.2	434.3
GPO+DSL(ours)	**84.6**	**97.6**	**99.2**	**66.9**	**92.4**	**96.5**	**537.2**	**62.3**	**87.0**	**93.3**	**44.6**	**74.3**	**83.9**	**445.4**
LLP	81.7	96.9	98.8	65.9	91.2	95.6	530.1	60.8	86.2	92.8	43.2	72.8	82.7	438.6
LLP+DSL(ours)	**84.3**	**97.5**	**98.8**	**67.6**	**92.3**	**96.2**	**536.6**	**62.8**	**87.5**	**93.6**	**44.6**	**74.6**	**84.1**	**447.1**

This strategy achieved good results and was validated by ablation experiments

SNAL	SNDL	i2t(R@1)	t2i(R@1)	RSUM
×	×	81.7	61.4	513.5
✓	×	82.5	63.0	520.9
×	✓	82.0	62.6	517.7
✓	✓	**83.1**	**64.0**	**523.9**

Figure 5: The above figure shows the Top-15 search results for an image retrieval task on the Flickr30K 1K test set.

RESULTS

Figure 6: The above figure shows the Top-6 search results of the text retrieval task on the Flickr30K 1K test set, where green represents ground-truth text, red represents mismatched text.

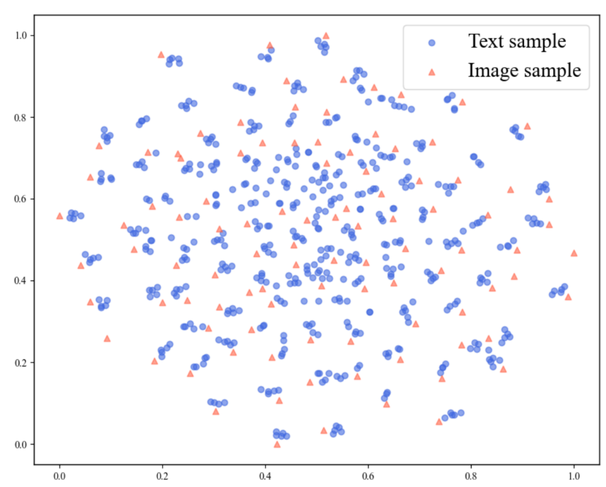


Figure 7: Feature embedding of 100 images and 500 text de□scriptions in the Flickr30K 1K test set in a feature space distribution map.