Cross-Modal and Uni-Modal Soft-Label Alignment for Image-Text Retrieval

Hailang Huang[1], Zhijie Nie[1,2], Ziqiao Wang[3], Ziyu Shang[4]

[1]SKLSDE, School of Computer Science and Engineering, Beihang University, Beijing, China

[2]Shen Yuan Honors College, Beihang University, Beijing, China

[3]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada

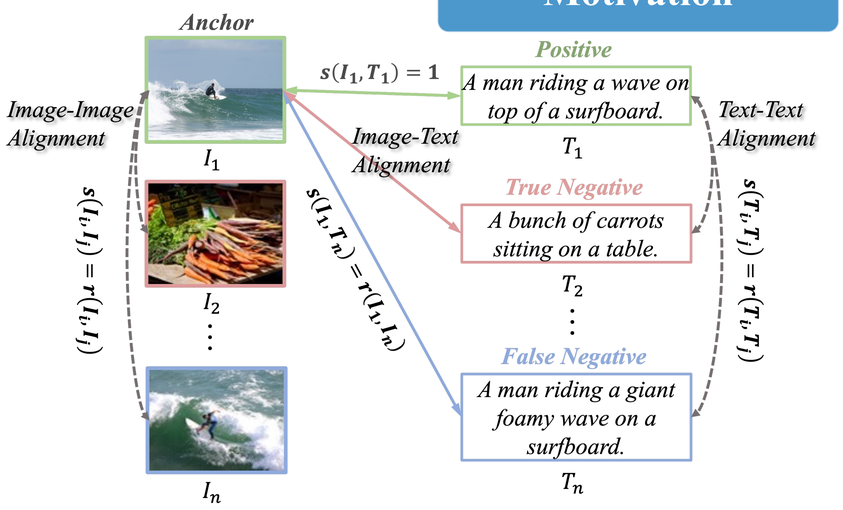
[4]School of Computer Science and Engineering, Southeast University, Nanjing, China

{huanghl, niezj}@act.buaa.edu.cn, zwang286@uottawa.ca, ziyus1999@seu.edu.cn

AAAI 2024

GitHub Link

Motivation



Most existing Image-Text Retrieval methods have two limitations: ***the inter-modal matching missing problem*** *and* ***the intra-modal semantic loss problem****.*

- ***The inter-modal matching missing problem*** *refers to the situation where, during model training, samples that should be matched are mistakenly treated as unmatched due to contrastive learning techniques and random sampling, resulting in a decrease in model performance.*
- ***The intra-modal semantic loss problem*** *refers to the insufficient capability of current ITR models to recognize similar input samples.*

Method

Cross-modal Soft-label Alignment

$$Q_{ij}^{\mathrm{i2t}} = \frac{\exp\left(s_{ij}^{\mathrm{i2t}}/\tau\right)}{\sum_{k=1}^{N}\exp\left(s_{ik}^{\mathrm{i2t}}/\tau\right)}, Q_{ij}^{\mathrm{t2i}} = \frac{\exp\left(s_{ij}^{\mathrm{t2i}}/\tau\right)}{\sum_{k=1}^{N}\exp\left(s_{ik}^{\mathrm{t2i}}/\tau\right)}.$$

$$P_{ij}^{\mathrm{i2t}} = \frac{\exp\left(r_{ij}^{\mathrm{i2t}}\right)}{\sum_{j=1}^{N}\exp\left(r_{ij}^{\mathrm{i2t}}\right)}, \quad P_{ij}^{\mathrm{t2i}} = \frac{\exp\left(r_{ij}^{\mathrm{t2i}}\right)}{\sum_{j=1}^{N}\exp\left(r_{ij}^{\mathrm{t2i}}\right)}.$$

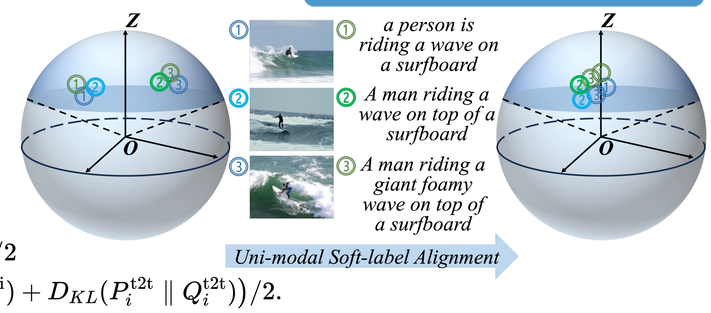
$$\begin{aligned}\mathcal{L}_{\mathrm{CSA}} &= \left(\mathcal{L}_{\mathrm{CSA}}^{\mathrm{i2t}} + \mathcal{L}_{\mathrm{CSA}}^{\mathrm{t2i}}\right)/2 \\ &= \left(D_{KL}(P_i^{\mathrm{i2t}} \parallel Q_i^{\mathrm{i2t}}) + D_{KL}(P_i^{\mathrm{t2i}} \parallel Q_i^{\mathrm{t2i}})\right)/2.\end{aligned}$$

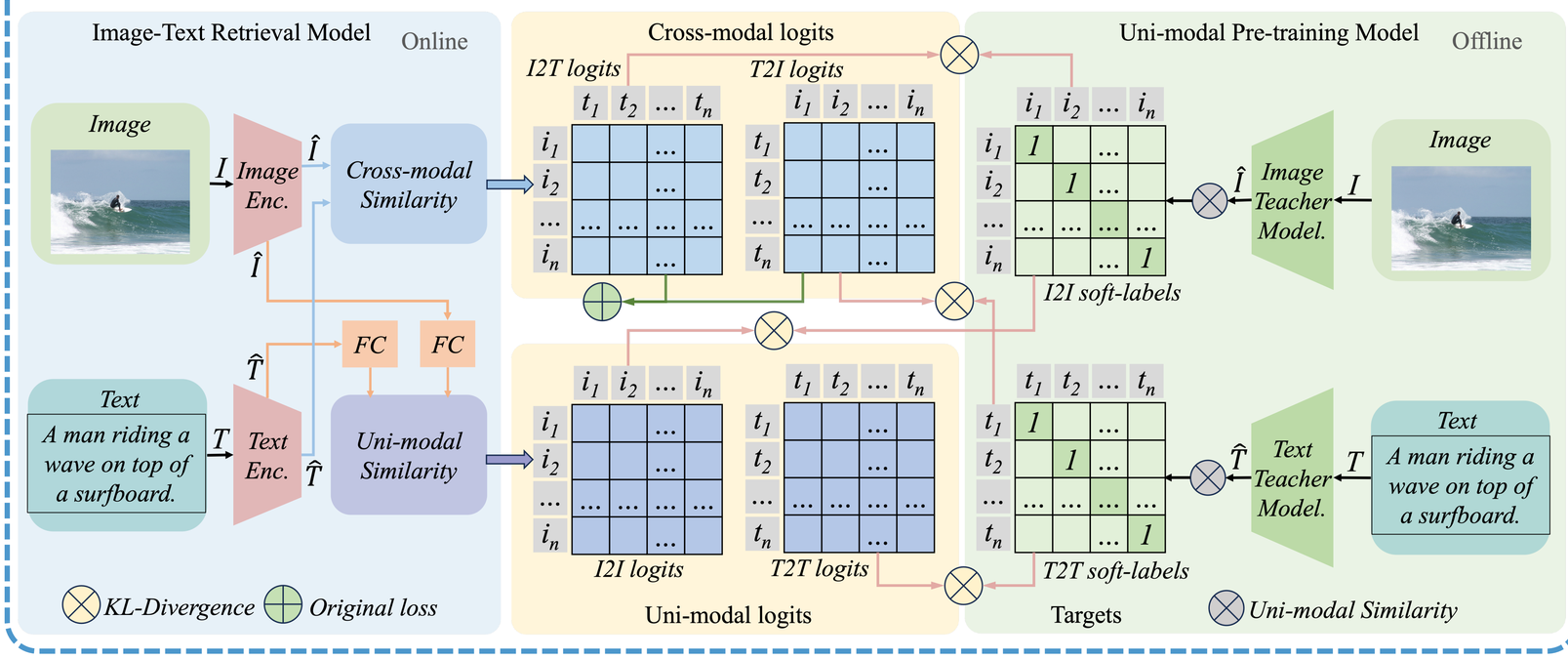
Uni-modal Soft-label Alignment

$$Q_{ij}^{\mathrm{i2i}} = \frac{\exp\left(s_{ij}^{\mathrm{i2i}}/\tau\right)}{\sum_{j=1}^{N}\exp\left(s_{ij}^{\mathrm{i2i}}/\tau\right)},$$

$$Q_{ij}^{\mathrm{t2t}} = \frac{\exp\left(s_{ij}^{\mathrm{t2t}}/\tau\right)}{\sum_{j=1}^{N}\exp\left(s_{ij}^{\mathrm{t2t}}/\tau\right)}.$$

$$\begin{aligned}\mathcal{L}_{\mathrm{USA}} &= \left(\mathcal{L}_{\mathrm{USA}}^{\mathrm{i2i}} + \mathcal{L}_{\mathrm{USA}}^{\mathrm{t2t}}\right)/2 \\ &= \left(D_{KL}(P_i^{\mathrm{i2i}} \parallel Q_i^{\mathrm{i2i}}) + D_{KL}(P_i^{\mathrm{t2t}} \parallel Q_i^{\mathrm{t2t}})\right)/2.\end{aligned}$$





Experiment

Model	MSCOCO (5K Test Set) Image-to-Text R@1	R@5	R@10	Text-to-Image R@1	R@5	R@10	RSUM	Flickr30K (1K Test Set) Image-to-Text R@1	R@5	R@10	Text-to-Image R@1	R@5	R@10	RSUM
Faster-RCNN, ResNet-101, without pre-training														
SCAN	50.4	82.2	90.0	38.6	69.3	80.4	410.9	67.4	90.3	95.8	48.6	77.7	85.2	465.0
VSE∞	56.6	83.6	91.4	39.3	69.9	81.1	421.9	76.5	94.2	97.7	56.4	83.4	89.9	498.1
VSRN++	54.7	82.9	90.9	42.0	72.2	82.7	425.4	79.2	94.6	97.5	60.6	85.6	91.4	508.9
NAAF	58.9	85.2	92.0	42.5	70.9	81.4	430.9	**81.9**	**96.1**	98.3	61.0	85.3	90.6	513.2
SGR†	57.3	83.2	90.6	40.5	69.6	80.3	421.5	76.6	93.7	96.6	56.1	80.9	87.0	490.9
+ **CUSA**	57.4	84.5	92.0	40.9	71.2	81.9	427.9	79.3	94.9	97.5	58.4	84.2	89.5	503.7
SAF†	55.5	83.8	91.8	40.1	69.7	80.4	421.3	75.6	92.7	96.9	56.5	82.0	88.4	492.1
+ **CUSA**	55.6	84.7	92.3	40.8	71.7	82.4	427.5	77.8	95.0	98.0	58.5	83.9	90.3	503.5
SGRAF†	58.8	84.8	92.1	41.6	70.9	81.5	429.7	78.4	94.6	97.5	58.2	83.0	89.1	500.8
+ **CUSA**	**59.8**	**86.1**	**93.3**	**43.3**	**73.2**	**83.6**	**439.2**	81.4	95.6	**98.5**	**61.0**	**86.1**	**91.5**	**514.1**
Dual-Encoder, pre-training														
$\mathrm{CLIP}_{\mathrm{ViT\text{-}B/32}}$	56.3	81.7	89.4	42.8	71.2	81.1	422.6	78.7	**95.4**	**98.0**	66.3	88.6	93.1	520.0
+ **CUSA**	**57.3**	**83.1**	**90.3**	**44.2**	**72.7**	**82.1**	**429.7**	**82.1**	95.3	97.9	**67.5**	**89.6**	**93.9**	**526.3**
$\mathrm{CLIP}_{\mathrm{ViT\text{-}L/14}}$‡	67.1	89.4	94.7	51.6	79.1	87.7	469.6	87.3	99.0	99.5	76.4	94.8	97.4	554.5
+ **CUSA**	**67.9**	**90.3**	**94.7**	**52.4**	**79.8**	**88.1**	**473.1**	**90.8**	**99.1**	**99.7**	**77.4**	**95.5**	**97.7**	**560.2**
Dual Encoder + Fusion encoder reranking, pre-training														
$\mathrm{BLIP}_{\mathrm{base}}$	81.9	95.4	97.8	64.3	85.7	91.5	516.6	97.3	99.9	100.0	87.3	97.6	98.9	581.0
OmniVL	82.1	95.9	98.1	64.8	86.1	91.6	518.6	97.3	99.9	100.0	87.9	97.8	99.1	582.0
$\mathrm{X2VLM}_{\mathrm{base}}$	**83.5**	96.3	98.5	66.2	87.1	92.2	523.8	98.5	100.0	100.0	90.4	98.2	99.3	586.4
+ **CUSA**	83.3	**96.6**	**98.5**	**67.1**	**87.6**	**92.7**	**525.8**	**98.5**	**100.0**	**100.0**	**91.3**	**98.8**	**99.5**	**588.1**
$\mathrm{X2VLM}_{\mathrm{large}}$	84.4	96.5	98.5	67.7	87.5	92.5	527.1	98.8	100.0	100.0	91.8	98.6	99.5	588.7

Model	CUB	SOP	In-Shop	INaturalist	Avg.
Faster-RCNN, ResNet-101, without pre-training					
SGR†[1]	31.1	51.9	19.5	33.7	34.1
+ **CUSA**	**34.6**	**60.7**	**31.6**	**41.9**	**42.2**
SAF†[1]	34.1	52.8	20.3	37.0	36.0
+ **CUSA**	**39.9**	**59.6**	**32.2**	**44.6**	**44.1**
Dual-Encoder, pre-training					
$\mathrm{CLIP}_{\mathrm{ViT\text{-}B/32}}$	41.5	51.8	28.1	41.3	40.7
+ **CUSA**	**49.6**	**56.5**	**34.1**	**45.6**	**46.5**
$\mathrm{CLIP}_{\mathrm{ViT\text{-}L/14@336px}}$	58.3	61.1	46.9	63.5	57.4
+ **CUSA**	**67.2**	**63.0**	**48.2**	**68.7**	**61.8**
Dual Encoder + Fusion encoder reranking, pre-training					
$\mathrm{X2VLM}_{\mathrm{base}}$ †[2]	53.6	64.2	52.6	59.3	57.4
+ **CUSA**	**58.9**	**67.0**	**54.2**	**62.2**	**60.6**

Model	$\mathrm{STS12\text{-}16}_{\mathrm{Avg.}}$	STS-B	SICK-R	Avg.‡
Faster-RCNN, ResNet-101, without pre-training				
SGR†[1]	51.8	58.1	62.7	54.3
+ **CUSA**	**55.9**	**65.2**	**64.9**	**58.5**
SAF†[1]	53.9	64.5	63.5	56.8
+ **CUSA**	**54.8**	**66.3**	**64.5**	**57.8**
Dual-Encoder, pre-training				
$\mathrm{CLIP}_{\mathrm{ViT\text{-}B/32}}$	67.4	76.2	72.9	69.4
+ **CUSA**	**71.6**	**78.3**	**75.8**	**73.2**
$\mathrm{CLIP}_{\mathrm{ViT\text{-}L/14@336px}}$	69.8	78.6	**75.5**	71.9
+ **CUSA**	**73.4**	**79.9**	74.9	**74.5**
Dual Encoder + Fusion encoder reranking, pre-training				
$\mathrm{X2VLM}_{\mathrm{base}}$ †[2]	26.6	22.3	50.4	29.4
+ **CUSA**	**46.8**	**47.9**	**76.2**	**51.2**

Model	ECCV Caption Image-to-Text mAP@R	R-P	R@1	Text-to-Image mAP@R	R-P	R@1
Faster-RCNN, ResNet-101, without pre-training						
SGR†[1]	26.8	38.7	70.3	42.2	51.2	83.6
+ **CUSA**	28.0	40.0	72.4	44.0	53.0	83.4
SAF†[1]	26.6	38.5	69.6	43.1	52.0	83.8
+ **CUSA**	27.4	39.8	71.4	44.4	53.6	84.6
SGRAF†[1]	28.1	39.8	72.3	43.7	52.5	84.4
+ **CUSA**	**29.5**	**41.4**	**74.5**	**46.4**	**55.1**	**85.7**
Dual-Encoder, pre-training						
$\mathrm{CLIP}_{\mathrm{ViT\text{-}B/32}}$	28.5	39.4	**72.5**	41.7	50.8	83.0
+ **CUSA**	**29.6**	**40.7**	72.0	**45.2**	**53.6**	**85.7**
$\mathrm{CLIP}_{\mathrm{ViT\text{-}L/14@336px}}$	32.8	43.4	79.7	45.5	54.2	87.2
+ **CUSA**	**33.6**	**44.1**	**80.9**	**47.6**	**55.8**	**88.2**
Dual Encoder + Fusion encoder reranking, pre-training						
$\mathrm{X2VLM}_{\mathrm{base}}$ †[2]	36.6	45.2	89.7	43.8	51.2	93.5
+ **CUSA**	**37.6**	**46.5**	**89.8**	**48.4**	**55.9**	**94.1**

- *Our method achieves* ***SOTA*** *results on three image-text retrieval datasets (MSCOCO, Flickr30K, ECCV Caption).*
- *Our method can also boost the uni-modal performance (Image Retrieval, STS Benchmark) of the image-text retrieval model, enabling it to* ***achieve universal retrieval****.*
- *The case study results show that our method not only* ***recalls more false negative cases****, but also allows the model to* ***better recognize similar input samples****.*

Conclusion

In this paper, we have proposed a novel method for image-text retrieval, called ***Cross-modal and Uni-modal Soft-label Alignment****. Our method leverages a uni-modal pre-training model to provide soft-label supervision signals for the ITR model, and uses two alignment techniques, CSA and USA, to overcome false negatives and enhance similarity recognition between uni-modal samples. Our method is* ***plug-and-play*** *and can be easily applied to existing ITR models without changing their original architectures. We have conducted extensive experiments on various ITR models and datasets and demonstrated that our method can consistently improve the performance of image-text retrieval and achieve new* ***state-of-the-art*** *results. Moreover, our method can also boost the uni-modal retrieval performance of the ITR model, enabling it to* ***achieve universal retrieval****.*