

Supplementary Materials: Bridging Visual Affective Gap: Borrowing Textual Knowledge by Learning from Noisy Image-Text Pairs

Anonymous Authors

1 TEXT PREPROCESSING

We enhance the pre-trained visual model with PACL by training on TumEemo [12], a social media post dataset. However, the texts of samples from TumEemo [12] contain many characters that are uninformative or cannot be encoded, including web URLs, usernames, emojis, and tokens from rare languages. To effectively leverage the factual and emotional connections between images and texts, we preprocess texts by replacing or removing these characters. (1). For web URLs, we replace them with [URL] characters. (2). For usernames, we replace them with [USER] characters. (3). For emojis and rare languages, we remove them. (4). For hashtags starting with “#”, we remove “#” and retain the contents of them.

2 DATASET PARTITION TOOLS

In dataset partition, the first stage of PACL, we employ CLIP [9] as the grounding evaluator of the factual connection, and DeepSentiBank [2], BERTweet [8] as the grounding evaluator of the emotional connection. We treat these models as off-the-shelf tools to approximate the accurate partition at low computational costs. Therefore, these models are replaceable by other tools. To prove this, we adopt other tools as the evaluators to enhance pre-trained ResNet-50 [5] with PACL and report the experimental results on FI [13] under linear evaluation in Table 1.

Table 1: PACL’s emotional enhancement of ResNet-50 on FI [13] under linear evaluation by adopting different tools as factual and emotional evaluators.

Factual		Emotional		FI	
Image	Text	Image	Text	Acc-8	F1-8
	CLIP [9]	DeepSentiBank [2]	BERTweet [8]	59.50	59.03
	CLIP [9]	Affection [1]	BERTweet [8]	60.12	59.60
	CLIP [9]	DeepSentiBank [2]	SKEP [11]	59.37	58.96
	CLIP [9]	Affection [1]	SKEP [11]	59.91	59.65
M ² [3]	BERT [4]	DeepSentiBank [2]	BERTweet [8]	59.24	58.86
M ² [3]	RoBerta [6]	DeepSentiBank [2]	BERTweet [8]	59.43	59.07
GRIT [7]	BERT [4]	DeepSentiBank [2]	BERTweet [8]	59.57	59.22
GRIT [7]	RoBerta [6]	DeepSentiBank [2]	BERTweet [8]	59.70	59.13

We use a pair of image and text tools as the evaluators except for CLIP. The image tools convert images into texts, while the text tools calculate the cosine similarities to assess the factual or emotional connections. As shown in Table 1, the emotional perception of ResNet-50 is consistently enhanced by PACL with different evaluators. Additionally, the enhancements across these evaluators are quite stable. Compared to our adopted tools, other tools bring at most 0.62% accuracy improvements or 0.26% accuracy degradations. These results confirm the robustness of PACL in the selection of evaluators.

3 DATASET SELECTION

To transfer knowledge from texts to images, we train PACL on TumEemo [12], a dataset rarely adopted in vision-language pre-training. In this section, we explain why we choose it. Specifically, we comprehensively compare it with CC3M [10] to demonstrate its advantages. CC3M is popular in vision language pre-training and contains approximately 3.3M image-text pairs. It is also collected from the web, yet undergoes heavy pre-processing that removes noun modifiers and weakens the emotional information linked to them. This results in its samples having strong factual connections but relatively weak emotional connections.

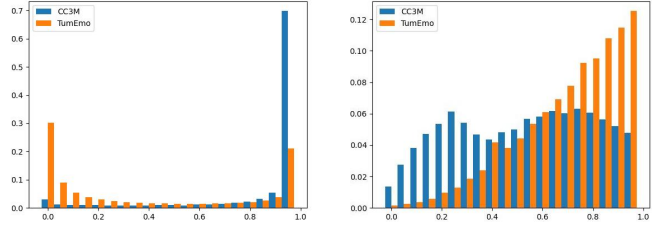


Figure 1: Distributions of the quantitatively assessed factual connections (left) and the emotional connections (right) of samples from TumEemo and CC3M.

We visualize the distributions of factual and emotional connections of samples from two datasets in Fig. 1. Under our partition, the percentage of factual-matched to factual-mismatched samples is 39%/61% in TumEemo and 91%/9% in CC3M, the percentage of emotional-matched to emotional-mismatched samples is 56%/44% in TumEemo and 29%/71% in CC3M. Samples in TumEemo possess relatively strong factual connections, though not as strong as those in CC3M, and simultaneously relatively stronger emotional connections, which makes them more suitable for our goal of bridging the visual ‘affective gap’.

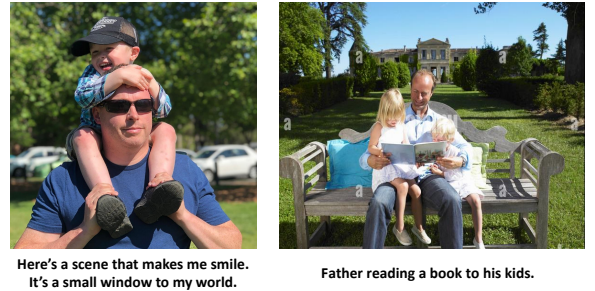


Figure 2: Examples from TumEemo (left) and CC3M (right).

Table 2: Emotional enhancement of ResNet-50 on various emotion-related downstream tasks. PACL (CC3M) and PACL (TumEmo) represent training PACL on CC3M and TumEmo, respectively. The best results are marked in bold.

Model	Single-label						Hierarchical-label						Multi-label		VAD			
	FI		Emotion-6		UnbiasedEmo		WebEmo						Emotic		Emotic		IAPS	
	Acc-8	F1-8	Acc-6	F1-6	Acc-6	F1-6	Acc-2	F1-2	Acc-6	F1-6	Acc-25	F1-25	mAP	AUC	MSE	R^2	MSE	R^2
<i>Linear Evaluation:</i>																		
ResNet-50	57.72	56.72	46.47	45.18	60.20	60.08	66.25	66.21	41.08	39.03	24.94	22.80	25.85	66.36	2.825	0.1267	2.130	0.2790
ResNet-50 + PACL (CC3M)	58.03	57.33	46.98	46.20	64.22	64.16	66.54	66.46	41.27	38.68	24.10	21.94	26.32	66.40	2.843	0.1255	2.107	0.2836
ResNet-50 + PACL (TumEmo)	59.50	59.03	51.14	50.59	66.78	66.39	67.53	67.42	43.60	40.95	25.85	24.32	27.02	67.60	2.646	0.1404	1.982	0.3104
<i>Fine-tuning:</i>																		
ResNet-50	65.14	64.84	53.46	53.25	73.68	73.15	73.81	73.80	48.88	48.34	30.85	29.89	28.36	68.43	2.757	0.1337	2.114	0.2842
ResNet-50 + PACL (CC3M)	66.35	65.42	56.94	56.72	75.81	75.03	74.26	74.17	49.35	48.60	31.24	30.67	29.14	69.22	2.599	0.1382	2.043	0.2916
ResNet-50 + PACL (TumEmo)	67.11	66.79	58.60	58.16	77.30	76.65	74.76	74.75	50.18	49.21	32.72	31.42	30.39	70.12	2.531	0.1460	1.927	0.3130

We present examples from two datasets in Fig. 2 to provide a more intuitive explanation. In images depicting similar scenes of a father interacting with their children, the text in TumEmo tends to focus on emotional interpretation. In contrast, the text in CC3M is more inclined toward factual descriptions. These differences emphasize the potential of TumEmo in conducting emotional knowledge transfer, which is also validated by the zero-shot performance achieved by the enhanced ResNet-50.

In addition to qualitative analysis, we provide quantitative results by directly training PACL on CC3M to enhance ResNet-50 [5]. As shown in Table 2, although training PACL on CC3M enhances ResNet-50 in most cases, especially under fine-tuning, training PACL on TumEmo can always achieve more effective enhancements. It proves that compared to solely strong factual connections, relatively strong factual and emotional connections are more beneficial to emotional knowledge transfer. Given all these reasons, we adopt TumEmo as the training dataset of PACL.

REFERENCES

- [1] Panos Achlioptas, Maks Ovsjanikov, Leonidas J. Guibas, and Sergey Tulyakov. 2023. Affection: Learning Affective Explanations for Real-World Visual Data. In *IEEE/CVF CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 6641–6651. <https://doi.org/10.1109/CVPR52729.2023.00642>
- [2] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. DeepSentBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. *CoRR* abs/1410.8586 (2014). arXiv:1410.8586 <http://arxiv.org/abs/1410.8586>
- [3] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *IEEE/CVF CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE/CVF, 10575–10584. <https://doi.org/10.1109/CVPR42600.2020.01059>
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/V1/N19-1423>
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Computer Vision Foundation / IEEE, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 <http://arxiv.org/abs/1907.11692>
- [7] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2022. GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features. In *ECCV 2022, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI (Lecture Notes in Computer Science, Vol. 13696)*. Springer, 167–184. https://doi.org/10.1007/978-3-031-20059-5_10
- [8] Juan Manuel Pérez, Juan Carlos Giudici, and Franco M. Luque. 2021. Pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. *CoRR* abs/2106.09462 (2021). arXiv:2106.09462 <https://arxiv.org/abs/2106.09462>
- [9] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [10] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Association for Computational Linguistics, 2556–2565. <https://doi.org/10.18653/V1/P18-1238>
- [11] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. In *ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 4067–4076. <https://doi.org/10.18653/V1/2020.ACL-MAIN.374>
- [12] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2021. Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. *IEEE Trans. Multim.* 23 (2021), 4014–4026. <https://doi.org/10.1109/TMM.2020.3035277>
- [13] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. In *AAAI 2016, February 12-17, 2016, Phoenix, Arizona, USA*. AAAI Press, 308–314. <https://doi.org/10.1609/AAAI.V30I1.9987>