FLYPE🌀: Multitask Prompt Tuning for Multimodal Human Understanding of Social Media

Bohua Peng[1,*], Chengfu Wu[1], Wei He[2], William Thorne[2], Aline Villavicencio[2], Yujin Wang[3] and Aline Paes[4]

[1]Northwestern Polytechnical University, Xian, 710072, China
[2]The University of Sheffield, Sheffield, S3 7HB, UK
[3]Tsinghua University, Shuangqing Road, 100084, China
[4]Universidade Federal Fluminense, Niterói, RJ, 24210-346, Brazil

Abstract

Large-scale pretraining and instruction tuning has facilitated visual language understanding on general purposes with broad competence. Social media processing will be highly benefit from large visual language models because messages are conveyed through joint reasoning over texts and images. Although vision-language pretraining has been widely studied, meta vision-language tuning remains under-explored. Given the ubiquity of visual content in social media, adapting pretrained Visual Language Models (VLMs) to meta social science is essential avoid extra computational expense on hyper-parameter search. This paper takes inspiration from cognitive studies to intrinsically and efficiently integrate a cross-modal reasoning into a method named FLYPE, as the runner up winner in CheckThat! 2023 task 1A. FLYPE integrates visual and text components of multiple tasks with cross-task shared prompts to guide a frozen VLM to perform as a meta classifier for unseen tasks. We evaluate our model across six social visual language understanding tasks and perform an ablation study on several modifications to the architecture. Our empirical study shows the competitive performance and training efficiency of the method. Using soft prompts can curate biased pretrained attention to focus on more task-related visual content. We release improved benchmarks with our model at https://anonymous.4open.science/r/Flype-2096.

Keywords

Large visual language model, parameter efficient training, matrix decomposition, bias analysis.

1. Introduction

Designing machine learning with multimodal perception abilities has recently gained significant attention due to its ubiquity in various computational social science (CSS) tasks such as emotion recognition, hate speech identification, and fact checking [1, 2]. Reasonably, many linguistic phenomena practiced in these tasks have visual representations. Furthermore, the environments where those tasks are more useful, such as social media, favor message conveying in textual or visual formats. This way, identifying the correlation and contrasts between textual and visual content to retrieve and classify relevant information is essential. For CSS tasks it is even

MUWS'23: 2nd International Workshop on Multimodal Human Understanding for the Web and Social Media, October 22, 2023, Birmingham, UK
*Corresponding author.
bohua\protect\TU_peng@mail.nwpu.edu.cn (B. Peng); alinepaes@ic.uff.br (A. Paes)

more challenging as it is common to disguise messages by including implicit visual elements or exploiting linguistic phenomena such as irony and sarcasm.

Although the interplay of how the brain processes images and text is an active area of research, evidence suggests that different channels analyse them separately toward the final goal and fusion happens at a later stage [3, 4, 5]. Computational methods for visual language understanding (VLU) take them as inspiration in the following pipeline [6]. First, the text associated with the image is analysed to check for expressions related to the final task. Next or parallel to that, the image is inspected to find task-related concepts. Finally, they are inspected together to fuse their correlations and predict the final labels. While such a pipeline can be instantiated with many different components, state-of-the-art VLU relies on transformers and very large language models [7], which in turn require expensive hardware and a considerable training time. This urges the development of methods that run on affordable hardware while still achieving high predictive performance. However, while spreading fake news and hateful messages are problematic universal behaviors, state-of-the-art multimodal transformers [8] are hardware-hungry, making them inaccessible to a large population. This urges the development of methods that can run on affordable hardware while still achieving high predictive performance.

This paper presents FLYPE, an efficient vision-language understanding method targeting CSS tasks. It adopts two general assumptions for efficiency: (i.) data addressed by different CSS tasks hold correlated underlying patterns as spreaders of harmful content often rely on similar visual and linguistic cues to manipulate people and deceive automatic methods, and (ii.) prompt and prefix-based tuning has demonstrated state-of-the-art results while still requiring less computational resources than fine-tuning entire models. FLYPE leverages both assumptions with a novel training regime of multiple-task knowledge distillation by fusing prompts of the two modalities and sharing prompts among different tasks. Moreover, as an AI-facilitated expert system, optical character recognition (OCR) [9] has been widely used for aiding understanding of visual language content [10, 11]. However, a question arises whether OCR is still necessary given that robust pretrained visual language models [12] can transliterate between images and texts. We have refurbished six visual language understanding benchmarks with an up-to-date OCR toolkit using Vision Transformer [13] to investigate that. Experimental results on several tasks, including emotion recognition, offensive comments, and automatic fact-checking, show that FLYPE improves parameter efficiency and generalization, particularly in low-resource settings.

2. Related work

Prompt & Instruction Tuning. Parameter-efficient fine-tuning methods have gained traction to address ever-increasing training times and computational requirements. One line of methods, Prompt-Tuning [14, 15], explore continuous prompts in the input space beyond the original vocabulary. This method aims to improve both parameter and data efficiency of fine-tuning by automatically searching for effective prompts. Instruction tuning involves fine-tuning large language models by formulating prompts in natural language [16]. This technique has demonstrated effective enhancement in zero-shot performance of general natural language understanding. Notably, Instruction Tuning methods often build on robust PLMs, enabling

faster convergence and yielding substantial improvements in downstream tasks. For example, Vicuna [17] is an instruction fine-tuned version of a foundation language model, LLaMA [18]. Inspired by these works, we opt to investigate parameter efficient Prompt-Tuning in cross-modal multitask learning settings.

Vision-Language Models CLIP proposes a simple yet powerful foundation model for general visual language representation learning. The model achieves this goal with large scale contrastive image-text alignment. This model significantly enhances the performance of multimodal downstream tasks because the generalizable representations aligned during pretraining can substantially improve visual content retrieval and visual grounding with its surprisingly pleasant zero-shot predictions. Frozen [19] extends Prompt-Tuning to the cross-modal setting by grounding a frozen large language model. Instead of learning soft text embedding, this approach efficiently tunes the visual tokens, corresponding to the convolutional layers of the model, presenting strong few-shot performance. Flamingo [6] proposes gated cross attention to fuse interleaved visual and language data, allowing for superb few-shot visual language generation which can be more aligned with real-world applications. BLIP-2 [7] introduces a Q-Former, which uses a bidirectional encoder [20] for cross-modal query tensor encoding. The approach essentially learns soft visual tokens to bootstrap visual features from the frozen visual encoder for visual question answering tasks. These visual tokens incorporate task information that regularize text generation search space, mitigating the image-to-text generation loss observed in Frozen or Flamingo. We extend the above ideas to the multitask visual language learning by fusing task-specific prompts with high order matrix decomposition methods.

Multimodal Social Computing Multimodal social computing aims to learn the patterns of human understanding about social media. However, some social bias, e.g., hatefulness, can be less perceptible hidden in multiple modes of information, including texts and images. These biased data can pose ethical threats to certain population and therefore should be curated before training any AI models.

3. Method

In Figure 1, we provide an overview of our cross-modal prompt tuning method, FLYPE, for social media understanding. The method consists of two stages: a task-targeting prompt tuning stage and a prompt fusion stage.

Task-targeted Prompt Tuning. Inspired by [21, 1], task-targeted prompt tuning aims to adapt a pretrained VLM to maximizing cross-modal semantic consistency of a specific CSS task. To this end, we pass the prompts through the first few layers of the decoder to understand entities and their relations through LLM reasoning. Then the prompt goes through a modality connector and enter Qformer to bootstrap the most conducive visual representations from the frozen visual backbone. Then the average visual query is fused with the last token of the decoder using cross-attention.

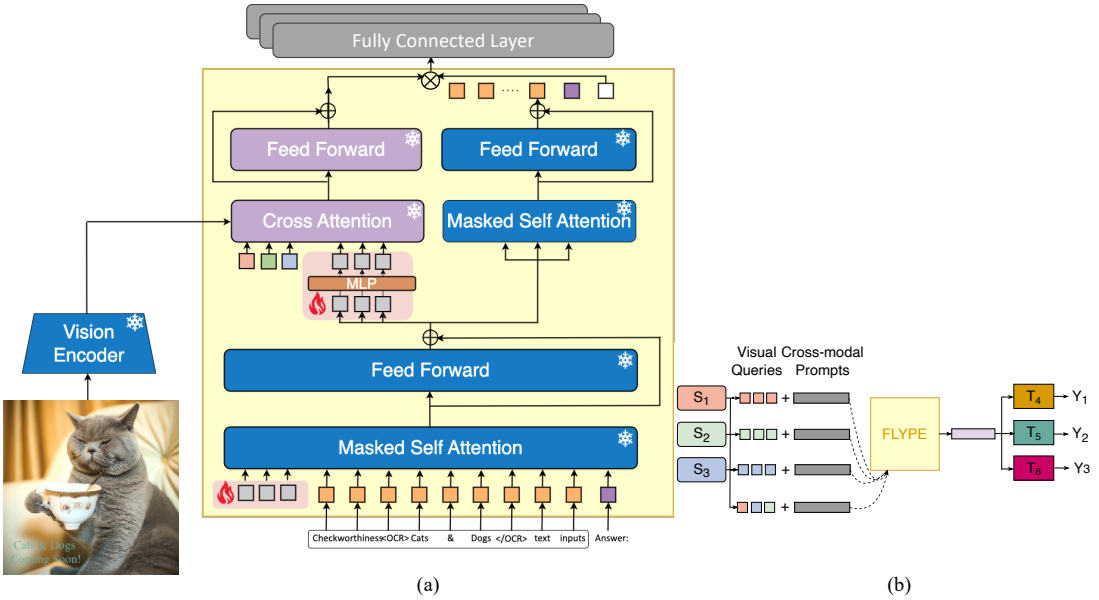

Figure 1: An overview of our multitask learning method for computational social science. (a) FLYPE utilizes a Query Transformer, or Q-former, and an LLM decoder. The Q-former is coloured in purple in Figure 1. The soft prompts (grey) guide both the Q-former the decoder to collaborate on the social media understanding tasks. In particular, these soft prompts hinge texts and images, modeling the consistency between two modalities. FLYPE predicts final labels based on aligned features, visual query tokens from the Q-former and the last token from the LLM decoder. (b) Prompt fusion: Based on HOSVD, we fuse a joint prompt by maximizing the correlation of prompts trained task-targeted datasets (S_1, S_2, S_3), hoping unseen datasets of held-out tasks (T_4, T_5, T_6).

In the text channel, we organize the texts as "(instruction, OCR and text)" to emulate the cross-modal consistency reasoning, a cognitive process for data collection, described in [22, 23]. In general, the reasoning process requires human annotators to first analyses the textual description of the image and checks for task semantic clues. Next, they inspect the image looking for task-oriented concepts. Finally, they connect both scrutinized components to predict the final label.

To emulate the process, we sequentially process the soft prompts, optical characters and text description with the causal self-attention mechanism of a frozen text decoder. The causal mask can enforce the self-attention to proceed from left to right, and the output probability indicates surprisal of entailment, which will be fused into the final decision. In the image channel, we use the same set of soft prompts but change their dimensions to bootstrap visual tokens. These soft prompts bring task information to visual tokens, and the latter becomes input features for classification. The process maximize a visual entailment probability, with the log-likelihood written as follows,

$$p_t' = \underset{p_t \in \mathbb{R}^{c \times d}}{\arg\max} \mathbb{E}[\log P(Y_t \mid I_t, p_t, O_t, X_t)] \tag{1}$$

where p_t' is the task-oriented prompt, θ is the frozen parameters of backbone encoders, and p_t is the prompt before augmentation. I_t, O_t, and X_t denote input images, OCR results, and text description, respectively. Y_t are the labels.

Linear Prompt Fusion. Conducting independent prompt tuning for each task can be a chal-

lenging and time-consuming process. However, [24] introduced an approach called Multitask Prompt Tuning based on knowledge distillation for text categorization tasks. This method incorporates task-independent components and shared prompts within its soft embedding. By utilizing shared prompts, zero-shot learning becomes possible even for unseen tasks. Nevertheless, when dealing with heterogeneous data in multimodal CSS tasks, the complexity of the MPT task weights can make hyperparameter searching more difficult.

To address this issue, we propose a method that leverages Higher-Order Singular Value Decomposition (HOSVD) [25] to maximize the mutual information of task instruction through prompt fusion. This approach allows us to reuse the computation results obtained from task-target prompts, thereby minimizing additional computations while obtaining a shared prompt and a meta classifier. As shown in Figure 1, the algorithm first calculates the covariance

Algorithm 1 Prompt fusion algorithm for FLYPE

Input: X - List of unique combinations of 3 flatten prompts;
Input: ϵ - Error tolerance for High Order Singular Value Decomposition (HOSVD);
Output: *core* - The share prompt fused from task-targeted prompts;
Output: *projection* - The codebook for task-specific prompt decoding;
$core \leftarrow [0]^d$
$projection \leftarrow [0]^{t \times d}$
for $(id_s, p_s), (id_t, p_t), (id_d, p_d) \in X$ **do**
  $p_s, p_t, p_d \leftarrow centerize((p'_s, p'_t, p'_d))$
  $covariance\_matrix \leftarrow covariance(p'_s, p'_t, p'_d)$
  $G, u, v, w \leftarrow HOSVD(covariance\_matrix, \epsilon)$
  for $G_i \in G$ **do**
    if $abs(G_i) > core[i]$ **then**
      $core[i] \leftarrow abs(G_i)$
      $projection[id_s, i] \leftarrow u[i]$, $projection[id_t, i] \leftarrow v[i]$, $projection[id_d, i] \leftarrow w[i]$
    end if
  end for
end for

matrix of *every three-task-combo* because the maximum dimensionality is three as we write this paper. Covariance matrix is computed on flatten prompts. For each covariance matrix, we apply higher-order singular value decomposition. For each dimension, we take the largest singular value (mode) and singular vectors to preserve and discard the rest. The proposed linear fusion method is shown in Algorithm 1. The algorithm can also be considered as learning a *codebook* to encode task-specific instructions as a *joint instruction* in a high dimensional space.

Training parameter efficiency. We analyze the theoretical training efficiency of Model Souping [26], task targeted prompt tuning and prompt fusion with regard to self-attention mechanism. Model Souping is a meta fientuning method that averages kT fully finetuned model weights with k-fold cross-validation on T tasks, the complexity is $O(T \times k \times n^2 \times d \times m)$, where n is the input sequence length, d is the number of dimensions of transformer embeddings, and m is the average number of validation samples in each held-out dataset. In the task-targeted prompt tuning, the complexity is reduced to $O(T \times k \times p \times (n + p) \times d \times m)$, where only p prompt tokens need to be tuned. In FLYPE, the HOSVD prompt fusion has a complexity of

$O((T \times k)^3 \times d)$. But since $T < k << d$, it can be considered as $O(\varepsilon \times d)$. There is no need for redundant cross-validation as the meta classifier shares hyper across all tasks. The complexity is therefore $O((k \times p \times (n + p) \times m + \varepsilon) \times d)$. HOSVD has explicitly enforced mutual information maximization across different social tasks. The hyper-parameter search space of the meta classifier is also reduced because the prompt length of the meta classifier is fixed. The parameter efficiency of prompt fusion is superior to Model Souping and task targeted prompt tuning.

4. Experimental Setup

Table 1
Dataset size of image and text proportions across train, validation and test sets combined.

Dataset	Text	Image	Classes
EmoRecCom	6,064	6,064	8
CheckThat!	3,911	3,911	2
SER30K	5,886	30,739	7
MAMI	11,000	11,000	2
HatefulMemes	16,428	16, 428	2
COSMOS	454,185	204,458	2

Tasks and Datasets. The statistics of the datasets used in our experiments can be found in Table 1. **Emotion Recognition on Comic Scenes (EmoRecCom)** is a competition dataset designed to tackle the task of recognising emotions based on comic panels, text in speech balloons or captions, and onomatopoeia [22]. **CheckThat! 2023 subtask 1A** explores whether tweets are worth fact-checking. Each instance is an image and a tweet, used to determine *check-worthiness* [27]. **SER30K** classifies the emotion conveyed by stickers and accompanying text, if present, used in online conversations [28]. **Multimedia Automatic Misogyny Identification (MAMI)** classifies whether text-image examples are misogynistic and then further classifies into a misogyny type [29]. Here we explore the binary setting only. **HatefulMemes** tackles the spread of toxic content in memes [30, 31]. **Catching Out-of-Context Misinformation using Self-Supervised Learning (COSMOS)** tries to classify images that have been used out of context to mislead a reader [32]. All the datasets we are using are benchmarks from shared task competitions and ethical issues were already tackled there, to the best of our knowledge.

Implementation details. We select the VQA foundation model, BLIP-2 [7], with ViT [33] as the visual encoder and OPT [34] as the text encoder for CSS classfication tasks. We employ the largest model that can fit into one RTX3090 to leverage the power of scale in prompt tuning. To avoid overfitting, we randomly mask 20% of patches to avoid overfitting. Instead of using the provided validation sets, we apply 10-fold cross-validation for hyperparameter search of the current task and reuse the resulting prompts for fusion. Ideally, more folds will result in better results because the estimated covariance matrix will also be more precise. To compute HOSVD [35], we follows the implementation of TensorToolbox [36]. Additional experimental details and hyper-parameters are in Appendix B.

Results. Table 1 shows the performance of FLYPE and zero-shot FLYPE compared with baselines

Table 2
Results of FLYPE and baselines across multimodal computational science datasets.

Model	EmoRecCom AUROC	CheckThat! F1	SER30K Accuracy
Zero-shot BLIP-2	0.228	0.449	0.124
CLIP	-	0.628	-
Zero-shot FLYPE	0.601	0.611	0.549
Task Best Performer	0.630	0.712	0.710
FLYPE (ours)	**0.779**	**0.717**	**0.744**
Model	**MAMI**	**HatefulMemes**	**COSMOS**
	Avg. F1-Macro	**AUROC**	**Accuracy**
Zero-shot BLIP-2	0.486	0.516	0.587
CLIP	0.704	0.610	-
Zero-shot FLYPE	0.703	0.612	0.729
Task Best Performer	0.731	0.765	0.850
FLYPE(ours)	**0.745**	**0.804**	**0.892**

including finetuned CLIP and frozen BLIP-2. Task Best Performers are winning models of each challenge, with concrete architecture detailed in A. FLYPE tunes task-targeted prompts. Zero-shot FLYPE fuses task-targeted prompts into a shared prompt and plug it into the frozen VLM, testing the model with unseen data from *held-out tasks*. Compared with CLIP, FLYPE uses cross-modal prompts to bootstrap image content into vision tokens with cross attention, and then fuses them with self-attention, which can be considered as a multi-stage fusion technique.

Table 3
Training efficiency comparison on the Check Worthiness dataset. $|p_t|$ denotes the prompt length.

Methods	#parameters (MB)	#memory (GB)	#training time (mins)	F1
FLYPE ($|p_t| = 1$)	0.003	11.966	9.771	0.562
FLYPE ($|p_t| = 5$)	0.013	15.830	10.210	0.717
FLYPE ($|p_t| = 10$)	0.026	19.682	12.166	0.704
FLYPE ($|p_t| = 20$)	0.051	22.055	13.312	0.683
Tuning CLIP	151.281	64.015	15.044	0.628
Tuning Q-former	188.144	33.812	22.937	0.682
Fully Finetuning	3744.710	-	-	-

Table 3 shows the improvement of FLYPE on training parameter efficiency. The table also shows that the number of training parameters increases linearly with the prompt length. Compared to fine-tuning a Q-former, FLYPE reduces the number of training parameters from 188MB to 0.01MB, significantly reducing the memory usage without loss of precision. Lengthy soft prompts improve the expressiveness of the model, but can lead to unexpected overfitting or noise injection to the visual tokens. The length of soft prompt is essential for CSS tasks, where each class sometimes only has a few hundred labeled samples, and overfitting can easily happen.

To understand other components of FLYPE, we perform an ablation of a number of techniques on the MAMI and CheckThat! datasets, as shown in Tables 4- A. The expert system, OCR, is still helpful for solving CSS tasks with visual language foundation models. As the pretrained fusion

Table 4
Ablation study on CheckThat! subtask 1A. The dataset has more negative samples than positive samples. High F1 and recalls are encouraged in the check worthiness task. Texts and soft prompts are more important components compared to pretrained visual tokens, which can be replaced by random noise during tuning.

Combination	Evaluation				Test			
	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall
soft prompts + ocr + visual tokens	**0.75**	0.80	0.62	**0.93**	**0.72**	**0.78**	**0.68**	0.77
ocr + visual tokens	0.65	0.69	0.51	0.90	0.65	0.67	0.56	0.82
soft prompts + visual tokens	0.71	**0.81**	**0.63**	0.87	0.70	0.75	0.65	0.76
replace image with noise	0.68	0.75	0.58	0.82	0.69	0.76	0.66	0.72
replace text with noise	0.56	0.57	0.41	0.86	0.62	0.58	0.47	**0.91**
replace visual tokens with noise	0.74	0.80	**0.63**	0.90	0.70	0.76	0.67	0.73

layers, the Q-former's pretrained self-attentions are biased towards visual question answering, which are curated by soft prompts tuned for CSS tasks. This is demonstrated by unnoticeable performance drop of replacing visual tokens with noise.

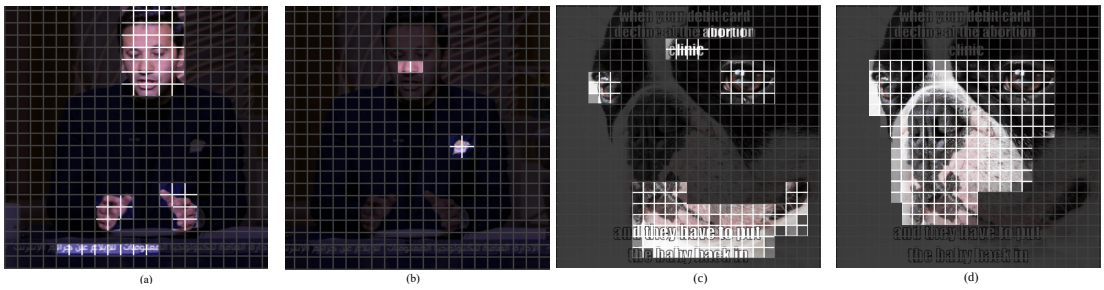

Figure 2: Comparison of layer-wise relevance propagation (LRP) saliency maps of prompts with protective attributes. (a) and (b) analyze bias against race. (a) "Being an Asian, Arabic fact checking texts."; (b) "Being a White, the same texts.". (c) and (d) analyze bias against species. (c) "Being a human, hateful texts."; (d) "Being a dog, the same hateful texts.". The correct prompts surprisingly activate relevant salient facial features and optical character features. By contrast, the biased prompts have much less salient activation, or rather sparse activation.

Cross-modal Bias Figure 3 shows the saliency map for a check worthiness sample and a hatefulness detection sample. We implement the salient feature detection with LRP[37] to the image pixels. If certain claims can be inferred from the image, the relevant image regions will be activated and considered for classification. The saliency maps validate our hypothesis. BLIP is converting CSS tasks into visual entailment tasks. Figure 3 compares the false negative equality difference of groups with protective attributes for the emotion recognition task. The adult group shows consistently lower false precision rate, which means their emotions are less likely to be wrongly classified as positive. Words with protective attributes may intrigue cross-modal bias when using hard prompts. By contrast, soft prompts can curate the potential bias from pretraining tasks. The use of soft prompts can contribute to making the deployment of LLMs safer and more ethical.

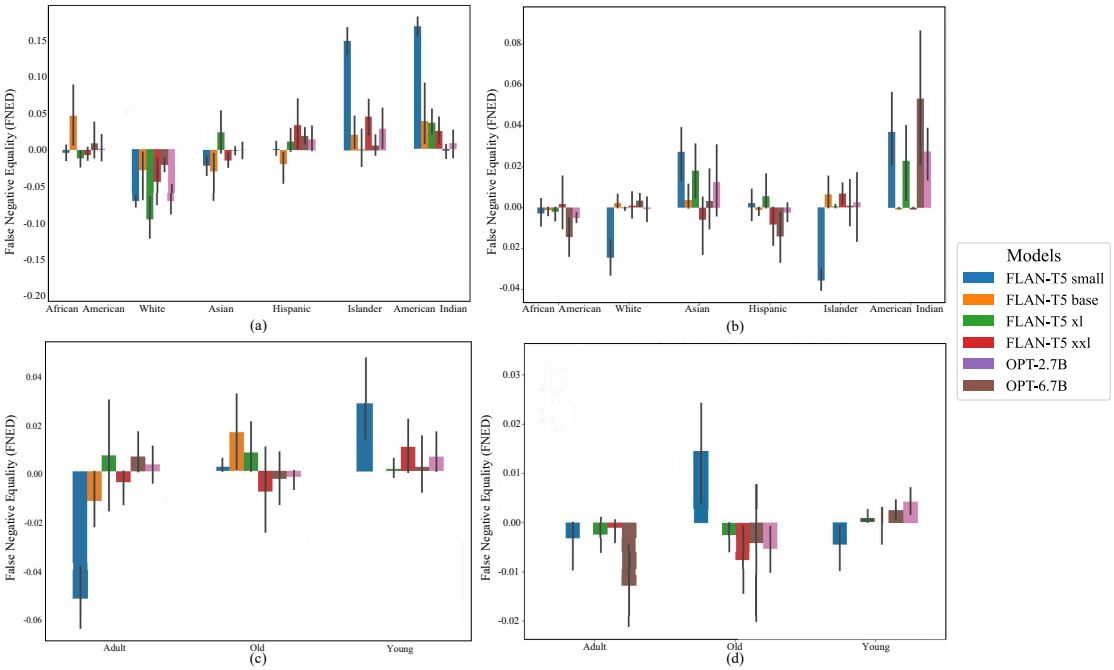

Figure 3: False negative equality difference for the protective attributes of age and race on the EmoRecCom dataset. (a) and (c) are results for samples with hard prompts, "Being a [protective attribute], multimodal emotion recognition context.". (b) and (d) are results for the same samples with soft prompts. Dark lines over the bars denotes the standard deviation.

5. Conclusion

In this paper, we investigate visual language reasoning problem and introduces FLYPE, a cross-modal meta prompt tuning method, which efficiently enhances visual language understanding. Our extensive experiments on six benchmark datasets demonstrate the effectiveness of this method for unseen dataset. The efficiency-wise comparison shows the training efficiency of this method. Through bias analysis, our work finds hard prompts can activate cross-modal bias which can raise ethical concerns. In our future work, we aim to categorize these cross-modal bias and address them with efficient approaches for safer deployment of large visual language models.

Limitations

Constrained by the existing pretrained foundation model, FLYPE faces a limitation in its ability to achieve long-distance attention for long input sequences. This constraint arises from the image-text alignment pretraining task of BLIP-2, which only considers 224×224 image patches and sentences with a maximum sequence length of 77. However, our proposed method has the potential to be extended to visual language models processing long contexts, thereby overcoming this bottleneck in the near furture.

References

[1] E. Blaier, I. Malkiel, L. Wolf, Caption enriched samples for improving hateful memes detection, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, ACL, 2021, pp. 9350–9358. URL: https://aclanthology.org/2021.emnlp-main.738. doi:10.18653/v1/2021.emnlp-main.738.

[2] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6625–6643. URL: https://aclanthology.org/2022.coling-1.576.

[3] T. C. Biggs, H. H. Marmurek, Picture and word naming: Is facilitation due to processing overlap?, The American Journal of Psychology (1990) 81–100.

[4] S. K. Reed, Cognitive architectures for multimedia learning, Educational psychologist 41 (2006) 87–98.

[5] S. Li, S. Chen, H. Zhang, Q. Zhao, Z. Zhou, F. Huang, D. Sui, F. Wang, J. Hong, Dynamic cognitive processes of text-picture integration revealed by event-related potentials, Brain Research 1726 (2020) 146513.

[6] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning, in: NeurIPS, 2022.

[7] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. arXiv:2301.12597.

[8] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.

[9] C. Amrhein, S. Clematide, Supervised ocr error detection and correction using statistical and neural machine translation methods, J. Lang. Technol. Comput. Linguistics 33 (2018) 49–76.

[10] H. Afli, Z. Qui, A. Way, P. Sheridan, Using smt for ocr error correction of historical texts (2016).

[11] A. Azadbakht, S. R. Kheradpisheh, H. Farahani, Multipath vit ocr: A lightweight visual transformer-based license plate optical character recognition, 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE) (2022) 092–095.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021. URL: https://api.semanticscholar.org/CorpusID:231591445.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Trans-

formers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[14] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, J. Tang, Gpt understands, too, arXiv:2103.10385 (2021).
[15] W. L. Tam, X. Liu, K. Ji, L. Xue, X. Zhang, Y. Dong, J. Liu, M. Hu, J. Tang, Parameter-efficient prompt tuning makes generalized and calibrated neural text retrievers, 2022. arXiv:2207.07087.
[16] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS (2022).
[17] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. arXiv:2306.05685.
[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, 2023. arXiv:2302.13971.
[19] M. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, F. Hill, Multimodal few-shot learning with frozen language models, in: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, 2021, pp. 200–212.
[20] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, ACL, 2019, pp. 4171–4186.
[21] E. Müller-Budack, J. Theiner, S. Diering, M. Idahl, R. Ewerth, Multimodal analytics for real-world news using measures of cross-modal entity consistency, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 16–25.
[22] N.-V. Nguyen, X.-S. Vu, C. Rigaud, L. Jiang, J.-C. Burie, Icdar 2021 competition on multimodal emotion recognition on comics scenes, in: J. Lladós, D. Lopresti, S. Uchida (Eds.), Document Analysis and Recognition – ICDAR 2021, Springer International Publishing, Cham, 2021, pp. 767–782.
[23] G. S. Cheema, S. Hakimov, A. Sittar, E. Müller-Budack, C. Otto, R. Ewerth, MM-claims: A dataset for multimodal claim detection in social media, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 962–979. URL: https://aclanthology.org/2022.findings-naacl.72. doi:10.18653/v1/2022.findings-naacl.72.
[24] Z. Wang, R. Panda, L. Karlinsky, R. Feris, H. Sun, Y. Kim, Multitask prompt tuning enables parameter-efficient transfer learning, arXiv preprint arXiv:2303.02861 (2023).
[25] L. R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.
[26] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al., Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, in: International Conference on Machine Learning, PMLR, 2022, pp. 23965–23998.
[27] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struss, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The clef-2023

checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 506–517. URL: https://checkthat.gitlab.io/clef2023/task1/.

[28] S. Liu, X. Zhang, J. Yang, Ser30k: A large-scale dataset for sticker emotion recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, MM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 33–41. URL: https://doi.org/10.1145/3503161.3548407. doi:10.1145/3503161.3548407.

[29] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://competitions.codalab.org/competitions/34175#learn_the_details. doi:10.18653/v1/2022.semeval-1.74.

[30] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 2020.

[31] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, C. A. Fitzpatrick, P. Bull, G. Lipstein, T. Nelli, R. Zhu, N. Muennighoff, R. Velioglu, J. Rose, P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, H. Yannakoudakis, V. Sandulescu, U. Ozertem, P. Pantel, L. Specia, D. Parikh, The hateful memes challenge: Competition report, in: H. J. Escalante, K. Hofmann (Eds.), Proceedings of the NeurIPS 2020 Competition and Demonstration Track, volume 133 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 344–360. URL: https://proceedings.mlr.press/v133/kiela21a.html.

[32] S. Aneja, C. Bregler, M. Nießner, COSMOS: Catching Out-of-Context Misinformation with Self-Supervised Learning, in: ArXiv preprint arXiv:2101.06278, 2021.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, 2021.

[34] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., Opt: Open pre-trained transformer language models, arXiv preprint arXiv:2205.01068 (2022).

[35] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM review 51 (2009) 455–500.

[36] B. W. Bader, T. G. Kolda, et al., Tensor toolbox for matlab, 2023. URL: http://www.tensortoolbox.org, general software, latest release.

[37] A. Binder, G. Montavon, S. Bach, K.-R. Müller, W. Samek, Layer-wise relevance propagation for neural networks with local renormalization layers, 2016. arXiv:1604.00825.

[38] R. Frick, I. Vogel, Fraunhofer sit at checkthat! 2023: Mixing single-modal classifiers to estimate the check-worthiness of multi-modal tweets, arXiv preprint arXiv:2307.00610 (2023).

[39] T.-V. La, Q.-T. Tran, T.-P. Tran, A.-D. Tran, D.-T. Dang-Nguyen, M.-S. Dao, Multimodal

cheapfakes detection by utilizing image captioning for global context, in: Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, 2022, pp. 9–16.
[40] G. K. Kumar, K. Nanadakumar, Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features, arXiv preprint arXiv:2210.05916 (2022).
[41] J. Zhang, Y. Wang, SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 585–596. URL: https://aclanthology.org/2022.semeval-1.81. doi:10.18653/v1/2022.semeval-1.81.
[42] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019.
[43] A. Gotmare, N. S. Keskar, C. Xiong, R. Socher, A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation, arXiv preprint arXiv:1810.13243 (2018).

Appendix

A. Explanation of Task Best Performers

To provide further clarification on the model comparison, we present the Task Best Performer for each task as follows:

- Check-worthiness (CheckThat!): [38] submitted an ensemble BERT classifier based on Model Souping. Model Souping is a meta-learning technique that aims to provide a better zero-shot classifier by averaging the weights of multiple independently fine-tuned models. The averaging process requires finetuning with k-fold cross-validation on each task, and the averaged models are further merged across tasks, leading to a meta classifier similar to FLYPE.
- Emotion Recognition (EmoRecCom): S-NLP used ResNet as image encoder and RoBERTa as text encoder. Both early and late level fusion are performed and combined in the final submission. In the early fusion stage, RoBERTa merges static image embedding and learnable text embeddings with finetuning. This model, however, requires much more training memory than prompt tuning an early stage fusion encoder, e.g., Q-former, as used in our work.
- Cheap Fake Detection (COSMOS): Boosting Image Captioning for Global Context [39] employs pretrained image caption models to describe images and predicts labels based on text description and context. The method essentially leverages the semantic understanding ability acquired through the next sentence prediction task of BERT's pretraining. The model further measures if two sentences are logically connected from the perspective of fake news detection.
- Hatefulness Detection (HatefulMemes): Hate-CLIPper [40] extracted image and text features independently with a frozen CLIP encoder. Then it computed cross-modal correlation scores with a bilinear pooling layer, and the the outer products are used for classification.
- Sticker Emotion Recognition (SER30K): The method used multi-stage fusion with Pyramid Vision Transformer (PVT), which is similar to visual feature extraction of BLIP-2. The method applied a spatial-reduction attention mechanism, where tokens are selected based on their attention scores with the CLS token. This actually models a recursive probability conditioned on a global latent variable.
- Gender Bias Misogyny Identification (MAMI): [41] finetuned late stage fusion layers on static image and text features extracted by a CLIP encoder with domain-adversarial loss, and used robust features to finetune a task specific classifier.

B. Additional Implementation Details

For image preprocessing, ViT randomly resized and cropped inputs to 224×224, and further cropped them to 16×16 non-overlapping patches. We search the optimal prompt token length from $\{5, 10, 20\}$. We use a batch size of 16, a learning rate of 1×10^{-3} for AdamW [42] with

learning rate warm-up [43]. We train FLYPE with the weighted cross-entropy loss for 5 epochs, which empirically guarantees convergence.

C. Visualization of Prompt Fusion

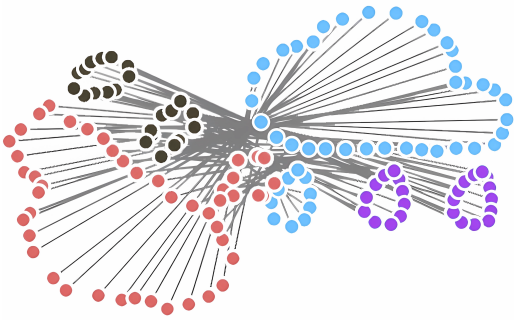

Figure A: A visualization of multi-task prompt fusion. The colourful scatter points represent task targeted prompts of four CSS tasks before fusion. The dark gray lines indicate the projections. The shared prompt is used for the other held-out tasks in this case.

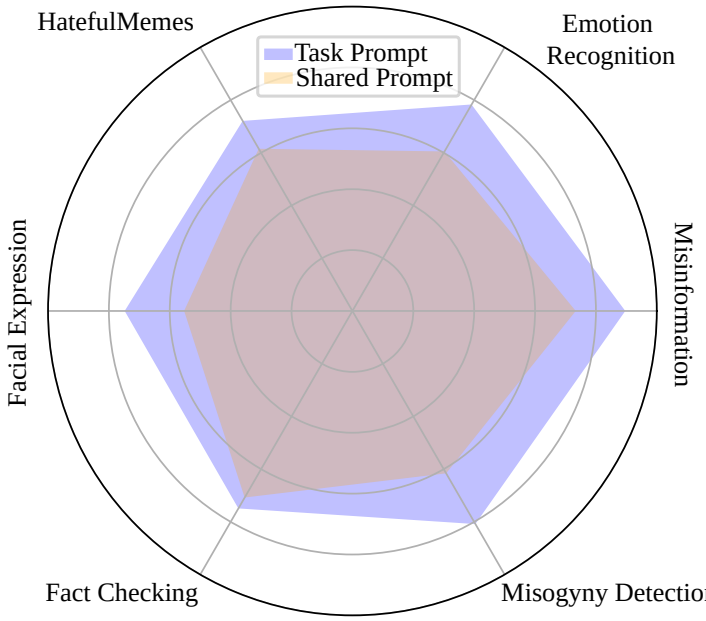

Figure B: The radius of the circle denotes 100% model performance and the orange is the performance of multitask shared prompt. As a linear machine learning algorithm, HOSVD maximize the mutual instruction information between the top three similar tasks.

Table A
Ablation study on MAMI.

Combination	Evaluation				Test			
	F1	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**
soft prompts + ocr + visual tokens	**0.84**	**0.85**	**0.82**	0.84	**0.75**	**0.71**	**0.66**	0.91
ocr + visual tokens	0.79	0.81	0.77	0.82	0.74	0.69	0.64	0.88
soft prompts + visual tokens	0.81	0.83	0.79	0.84	0.74	0.69	0.63	0.89
replace image with noise	0.77	0.78	0.71	0.84	0.64	0.59	0.57	0.74
replace text with noise	0.79	0.78	0.68	**0.93**	0.72	0.63	0.58	**0.94**
replace visual tokens with noise	0.82	0.84	**0.82**	0.82	0.73	0.69	0.65	0.86