Prompt Learning with Cross-Modal Feature Alignment for Visual Domain Adaptation*

Jinxing Liu[1], Junjin Xiao[1], Haokai Ma[1], Xiangxian Li[1], Zhuang Qi[1], Xiangxu Meng[1], and Lei Meng[1*]

Shandong University, Jinan, Shandong, China
liujinxing@mail.sdu.edu.cn junjxiao@163.com mahaokai@mail.sdu.edu.cn
xiangxian_lee@mail.sdu.edu.cn 97qizhuang@gmail.com mxx@sdu.edu.cn
lmeng@sdu.edu.cns

Abstract. Exploring the capacity of pre-trained large-scale models to learn common features of multimodal data and the effect of knowledge transfer on downstream tasks are two major trends in the multimedia field. However, existing studies usually use pre-trained models as feature extractors, or as the teacher model to achieve knowledge distillation of downstream tasks.Therefore, the cross-modal knowledge transfer mechanism and the knowledge forgetting problem of pre-trained large models have not been fully investigated.To address the above issues, this paper explores the fine-tuning strategy, feature selection strategy and semantic guidance approach in the migration process of pre-trained large models.Aiming at the problem of knowledge forgetting during "fine-tuning", an image classification algorithm (PMHANet) integrating a pre-trained large-scale model and heterogeneous feature alignment is proposed.More importantly, this provides a cross-modal knowledge transfer paradigm for multimodal pre-training of large models.We conducted experiments on VireoFood-172 and NUS-WIDE and found that large models trained on datasets such as COCO performed better on the similar domain dataset NUS-WIDE than the small domain dataset VireoFood-172; PMHANet effectively implements multimodal representation enhancement in downstream tasks based on a partially fine-tuned pre-trained large model to achieve SOTA performance on both datasets.

Keywords: Pre-trained multimodal models · Cross-modal knowledge transfer · Heterogeneous feature alignment · Image classification.

1 Introduction

In recent years, knowledge transfer based on multimodal pre-trained large models has received wide attention due to its excellent performance[15][10], which can bridge the semantic gap between different modal data and improve the performance of downstream tasks. However, current large models rely on large-scale multimodal data, while the data distribution and data modality are often limited in downstream tasks, leading to unsatisfactory performance after knowledge

* Corresponding author

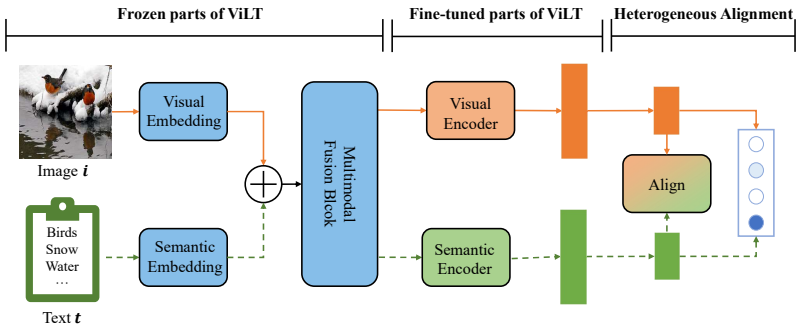

Fig. 1. Illustration of the PMHANet. Combining cross-modal transfer and heterogeneous feature alignment mechanisms with existing pre-trained large models, image representation enhancement is achieved by maximising the consistency between multimodal representations.

transfer. Therefore, how to combine the knowledge of multimodal pre-trained large models with small domain data to improve the performance of downstream tasks and explain their mechanisms is an urgent problem.

There are two main approaches to knowledge transfer for multimodal pre-trained models, one is to use the model as a feature extractor and fine-tune[21][7] , and the other is to use the model as a teacher to achieve knowledge distillation[20]. However, existing methods may lead to knowledge forgetting when transfering, limiting the performance of knowledge transfer. To address this issue, we use heterogeneous feature alignment to supplement the interaction information in the transfer phase. Heterogeneous feature alignment is an important technique for feature enhancement with multimodal data[14][16][3][11], but the effectiveness of alignment is currently limited due to the different distribution and value range between heterogeneous features.

To address the above issues, this paper explores how to use multimodal pre-trained large models for cross-modal knowledge transfer and proposes a model transfer method based on partial heterogeneous modal feature alignment (PMHANet). Specifically, this paper explores the transfer method for pre-trained large models by comparing the feature capabilities of different stages and types. Based on the above exploration, we use semantic information to enhance image representation and propose an image classification algorithm (PMHANet) that incorporates a pre-trained large model and partially heterogeneous modal feature alignment. As shown in Figure 2, this paper divides the pre-trained large model into four modules, which Shallow Feature Extraction(SFE), Feature Fusion Network(FFN), Fine-tune Network(FTN), and Heterogeneous Feature Alignment(HFA). Specifically, in the SFE Module, PMHANet forms shallow features for multimodal interactions by mapping visual modality and semantic modality information, respectively. To achieve heterogeneous modal feature interactions, PMHANet fuses the underlying features of images and text

and implements self-focused implicit feature alignment in the FFN; to achieve fine-grained feature extraction, PMHANet retrains the transfer network module on a segmented domain task dataset to enhance the image representation capability of the model in the FTN. Finally, to mitigate the loss of interaction information during multimodal large-scale model transfer, PMHANet learns text-enhanced visual representations for image classification by maximizing the distributional consistency between visual and semantic representations in the HFA. The PMHANet proposed in this paper solves the problem of interaction information loss during multimodal large-scale model transfer, achieves the effective transfer of multimodal pre-trained large-scale models in image classification tasks, and culminates in a fine-tuning paradigm for pre-trained models.

We conducted performance comparisons, an ablation study of the alignment method, and several in-depth analyses of the transfer paradigm of the pre-trained large model on two real-world datasets, VireoFood-172 and NUS-WIDE. The results show that PMHANet can further improve the image classification performance by partial fine-tuning and heterogeneous feature alignment. To summarize, this paper includes two main contributions:

(1) Based on ViLT, we explore the knowledge transfer capability of features at different stages and types of features, and propose a knowledge transfer paradigm for multimodal pre-trained large models.
(2) An image classification method (PMHANet) is proposed based on heterogeneous feature alignment, which can achieve semantic-guided image feature enhancement and effectively improve image classification performance.

2 Related Works

2.1 Multi-modal Pre-trained Large-scale Models

Multimodal pre-trained large-scale models are capable of automatically learning different modal features without supervision and quickly migrating to different down- stream tasks. Most of the current multimodal pre-trained large models are based on BERT[4] which can be structurally classified into single-stream and dual-stream models. Single-stream models such as VisualBERT[10], UNICODER[6], VL-BERT[12], and ImageBERT[15] use the collocation of visual and semantic features of image-text pairs as input to the interaction network, and multimodal information is fused without constraints at an early stage. In contrast, in dual-stream models such as ViLBERT[12], LXMERT[17], and ERNIE-ViL[22], different modal data are encoded respectively and the resulting embeddings interact through a co-attentive-based encoding layer. However, less research has been done to explore the transfer mechanism of pre-trained large models, and how to combine the knowledge of pre-trained large models to achieve feature enhancement for downstream tasks is still a problem waiting to be solved.

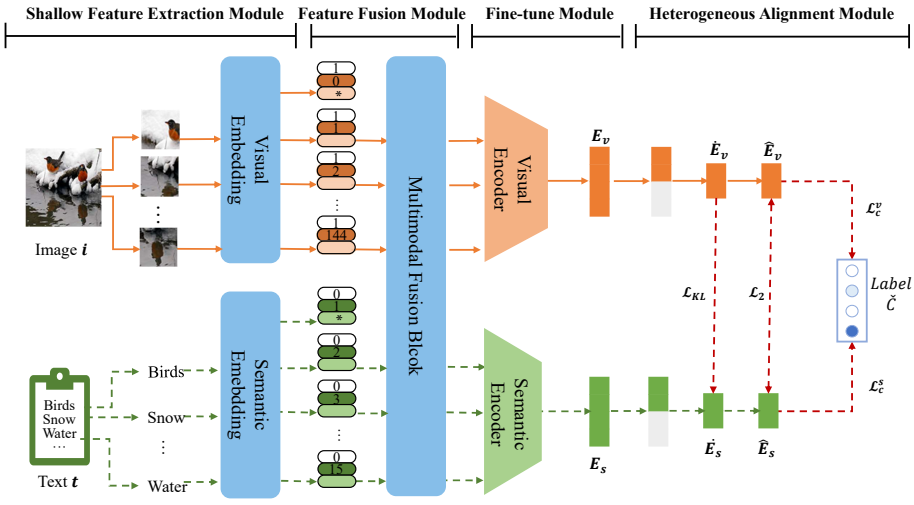

Fig. 2. Illustration of the PMHANet. PMHANet combines heterogeneous feature alignment mechanisms with existing pre-trained large-scale models to enhance image characterization and enable cross-modal knowledge transfer.

2.2 Heterogeneous Feature Alignment

Heterogeneous feature alignment techniques are able to learn linear or nonlinear mappings to align features from different domains in the latent space.Current research typically uses shared neural networks to learn high-level features of heterogeneous data and constrain their similarity, such as the KL-divergence[14], the loss of generative adversarial networks[3] and the covariance matrix of feature distributions[16].However, the existing methods do not address the inherent heterogeneity of data in different domains and cannot resolve the differences between text features and image features heterogeneous modalities in terms of feature distribution and value domain range, resulting in limited mapping effect of heterogeneous feature alignment techniques between modalities.

3 Technique

3.1 Framework Overview

In view of the powerful characterization ability of pre-trained large models and the feature enhancement ability of heterogeneous feature alignment, this paper proposes an image classification algorithm PMHANet.As shown in the Figure 2, PMHANet contains four main submodules.

3.2 Shallow Feature Extraction Module (SFE)

PMHANet generates low-level and shallow features for information interaction in Feature Fusion Module by linearly mapping image and text in the Shallow

Feature Extration Module respectively.Specifically, PMHANet accepts multimodal original image text pairs $\{(I_j, T_j) \mid j = 1, \ldots, N\}$ as input, where N represents the number of samples contained in a training batch.The input image $I \in R^{C \times H \times W}$ is sliced into equal-sized image patches, and embedded to $v \in R^{L \times D}$ through visual feature linear mapping layer $\mathcal{M}_v(\cdot)$. Then add the absolute position encoding $E_v^{pos} \in R^{L \times D}$ and the type encoding $E_v^{type} \in R^{L \times D}$ to generate the visual shallow feature $F_v \in R^{L \times D}$, where L is the content tokens and D is the hidden dimension of features. The processing of text data is similar. After the text data t is mapped into a word embedding matrix by the semantic feature embedding layer $\mathcal{M}_s(\cdot)$, the position encoding E_s^{pos} and type encoding E_s^{type} are added to form semantic shallow features F_s.

$$F_v = \mathcal{M}_v\left(I\right) + E_v^{pos} + E_v^{type} \tag{1}$$

$$F_s = \mathcal{M}_s\left(T\right) + E_s^{pos} + E_s^{pos} \tag{2}$$

3.3 Feature Fusion Network Module (FFN)

In order to leverage the generic knowledge in the multimodal pre-trained model and to enhance the representational power of the model when migrating to downstream tasks of image classification, we extract interaction features based on the multimodal feature fusion layer of the pre-trained large model.In the Feature Fusion module, PMHANet fuses heterogeneous modal features F_v and F_s and facilitates multimodal information interaction based on pre-trained knowledge to reduce the variability of distribution of heterogeneous features in the feature space and generate multimodal interaction features:

$$F_{sv} = \left(\theta\left(LN\left(MSA\left(LN\left[F_s : F_v\right]\right)\right)\cdots\right)\right. \tag{3}$$

where $MSA(\cdot)$ denotes multi-headed self-attention, $LN(\cdot)$ denotes LayerNorm normalization method, $\theta(\cdot)$ denotes two-layer fully connected network.

3.4 Fine-tune Network Module (FTN)

The Transfer Network Module fine-tunes the layers to generate fine-grained features in the segment and alleviates data limitations between the pre-training and downstream datasets such as data distribution.In detail, the fine-grained heterogeneous feature extraction was achieved by mapping interaction features to different feature spaces via visual mapping $\mathcal{T}_v(\cdot)$ and semantic mapping $\mathcal{T}_s(\cdot)$.After the multimodal interaction features are subjected to multi-headed attention operations, the content representation information is aggregated and the visual representation feature E_v is generated through the visual perceptual mapping module.Multimodal interaction features aggregate content representation information through semantic information mapping and generate semantic representation features E_s. The specific formulas are as follows:

$$E_v = \mathcal{T}_v\left(LN\left(MSA\left(LN\left(F_{sv}\right)\right)\right)\right) \tag{4}$$

$$E_s = \mathcal{T}_s\left(F_{sv}\right) \tag{5}$$

Table 1. Datasets used in the experiment.Tags means the number of text words

Dataset	Class	Tags	Train Split	Test Split
Vireo-Food172	172	353	68,175	25,250
NUS-WIDE	81	1,000	121,962	81,636

3.5 Heterogeneous Feature Alignment Module (HFA)

Visual and semantic encoders are able to learn visual and semantic features E_v and E_s independently, and given that semantic embeddings have better discriminative power on classification tasks, aligning visual and semantic features directly allows their distribution to converge and enables semantically guided visual feature enhancement.However, the inherent variability of heterogeneous modal characteristics leads to limited effects in alignment.

To solve this problem, we align the shared information of visual and semantic features in the HFA module, i.e. the partial heterogeneous alignment method. Based on the assumption that the embedding E_k consists of E_k^1, which represents label information, and E_k^2, which represents stylistic information.Therefore separating E_v^1 and E_s^1 from E_v and E_s and aligning them can alleviate the heterogeneity between heterogeneous modes, and endow visual features with semantic information, ultimately enhancing visual representation.To preserve the semantic modal feature distribution, we used a one-way KL-divergence loss function for cross-modal feature alignment to compensate for heterogeneous modal distribution differences. In addition, we use the ℓ_2 norm alignment method to align the heterogeneous modal features in the potential space.

$$\mathcal{L}_{KL} = KL\left(\dot{\mathbf{E}}^s \| \dot{\mathbf{E}}^v\right) \tag{6}$$

$$\mathcal{L}_{align} = \left\|\hat{\mathbf{E}}^v - \hat{\mathbf{E}}^s\right\|_2 \tag{7}$$

The representations output the category prediction information through a nonlinear mapping and use cross-entropy (CE) loss and binary cross-entropy (BCE) loss to calculate the classification loss in single-label classification and multi-label classification tasks, respectively.

$$\mathcal{L}_c^v = CE\left(softmax\left(\phi\left(\hat{\mathbf{E}}^v\right)\right), y\right) \tag{8}$$

$$\mathcal{L}_c^s = CE\left(softmax\left(\phi\left(\hat{\mathbf{E}}^s\right)\right), y\right) \tag{9}$$

4 Experiments

4.1 Datasets

We conducted experiments on two datasets, Vireo-Food172 and NUS- WIDE, and the statistical information corresponding to the datasets is given in Table 1:

- **VireoFood-172[1]**: A single-label classification dataset containing a total of 110,241 images of dishes, corresponding to 172 categories and 353 semantic elements with an average of three semantic elements annotated per image.We sliced all images according to the design in paper[1].
- **NUS-WIDE[2]**: A multi-label classification dataset containing a total of 269,648 image samples corresponding to 81 categories and 1000 semantic elements.We refer to the related paper[2][18][19], preprocessed the original dataset and divided the remaining 203,598 samples.

Table 2. Compare the experimental results, where the VireoFood-172 dataset uses the accuracy (Acc) of top1 and top5 as a measure of single classification task, and the NUS-WIDE dataset uses the precision (Prec) and recall (Recall) of top1 and top5 as the performance performance of multi-classification task

	Algorithm	Vireo-Food172		NUS-WIDE			
		Acc@1	Acc@5	Prec@1	Prec@5	Recall@1	Recall@5
Visual Classification	ResNet50	81.6	95.0	73.2	36.1	39.8	86.4
	WRN	82.3	95.5	73.4	36.5	39.8	78.7
	WISeR	82.8	96.5	73.7	36.7	40.1	79.0
	ViLT	69.4	90.2	80.7	40.5	45.8	88.0
Heterogeneous Feature Alignment	IG-CMAN	82.9	96.4	73.3	37.0	42.0	80.0
	ATNet	82.9	93.1	75.7	35.8	41.3	77.1
	MSMVFA	83.1	96.6	78.2	39.3	43.7	85.6
	PMHANet	**83.7**	**96.7**	**81.6**	**41.1**	**46.3**	**88.8**

4.2 Model details

In our experiments, we followed the feature dimension setting of the pre-trained model VILT[9], and the input image resolution was resized to 384. Model was optimised using the Adam optimiser during training, with the learning rate picked from 5e-5 to 5e-3 and with every four periods completed, the optimiser's learning rate decayed to a factor of 0.1 of the original.For NUS-WIDE, the positive sample weights for setting BCE losses were selected from 20 to 150.

4.3 Performance Comparison

In this section, we show the effect of PMHANet on two datasets, comparing the performance of the base vision models ResNet50[5] and the improved models WRN[23] and WISeR[13] based on ResNet50, and we also compare ViLT[9] transfered directly to the classification task. To compare with multimodal inference methods, we experimented with ATNet[14] and MSMVFA[8] methods based on feature alignment, and ARCH-D[1] methods based on multimodal information constraints based on ResNet50. As shown in Table 2.

- **For pre-trained basic vision models**: models with deeper layers such as ResNet50 and VGG19 achieved better results than ResNet-18 in both datasets.Improvements in model structure by the WRN and WISeR models yielded better predictive performance than ResNet50. This shows that structural improvements and parametric enhancement enhance the task migration capability of the model
- **For the large pre-trained model ViLT**: it produced inconsistent results when migrating to the classification task on both datasets. Compared to ResNet50, ViLT showed a 14.9% decrease in Top-1 on the VireoFood172 dataset, but a 10.2% increase on the NUS-WIDE dataset, indicating that the pre-trained large model is susceptible to data distribution in downstream task migration, which leads to inconsistent quality of its learned representations.
- **For multimodal inference methods**, the visual representation is enhanced by feature alignment-based or multimodal information constraint methods, thus improving the classification performance, where the MSMVFA method fuses multi-scale information and the CMRR method adds constraints on the local information thus further enhancing the inference capability.

Table 3. Performance of the PMHANet with different semantic guidance approaches on the Vireo172 and NUS-WIDE datasets.FT: partial fine-tuned ViLT.

Algorithm	VIreo172		NUS-WIDE			
	Acc@1	Acc@5	Prec@1	Prec@5	Recall@1	Recall@5
baseline	69.4	90.2	80.5	40.5	45.4	87.6
+FT	82.9	96.3	80.7	40.6	45.8	88.0
+FT+l2norm	82.9	96.5	80.7	40.6	45.8	88.0
+FT+KL	83.5	96.6	81.2	40.8	46.1	88.4
+FT+l2norm+KL	**83.7**	**96.7**	**81.6**	**41.1**	**46.3**	**88.8**

– **PMHANet achieves the best results on both datasets**, and obtains a 20.6% Top-1 accuracy improvement relative to the original ViLT on the Vireo-Food172 dataset. This demonstrates the effectiveness of the design of partial fine-tuning and heterogeneous feature alignment, allowing the model to learn a better visual representation , which also demonstrates the potential of PMHANet to facilitate model learning for out-of-distribution data.

4.4 Ablation Study

In addition to the overall performance comparison, we conducted ablation experiments in order to explore the effectiveness of the heterogeneous feature alignment method in PMHANet. The results are shown in Table 3, and we have the following findings:

– **Partial fine-tuning facilitates cross-modal knowledge transfer:** Due to the significant heterogeneity between the task targeted by the pre-trained large model and the downstream task, it is difficult to directly migrate it to the image classification task, so only 69.4% top1 accuracy is obtained on vireo172. In contrast, fine-tuning the model according to the downstream task to achieve cross-modal knowledge transfer for segmented datasets can achieve a performance improvement of 19.4%. Meanwhile, the performance improvement of "baseline+FT" on the NUS-WIDE dataset is not obvious, which may be due to the similar data distribution between the training dataset of the pre-trained large model and NUS-WIDE.
– **Heterogeneous feature alignment facilitates image feature enhancement:** L2norm can achieve similarity of heterogeneous features by aligning heterogeneous modal features in a uniform feature space, but bi-directional similarity cannot achieve specific feature enhancement.KL Divergence can asymmetrically achieve one-way similarity from image features to text features, which eventually yields 83.5% top1 accuracy on vireo172 and a simultaneous boost on nuswide. Combining l2norm and KL-divergence, the model can mitigate the semantic "gap" when migrating from the pre-trained large model to the downstream task by fine-grained heterogeneous alignment of visual and semantic modal representations through partial feature alignment methods, and compensate for the information loss during migration, thus achieving the best classification improvement on both datasets.

4.5 In-depth Analysis of PMHANet

In this section, we investigate the impact of using different fine-tuning strategies, feature selection approaches, and semantic guidance approaches when pre-trained large models

Table 4. Classification performance of ViLT combined with different fine-tuning strategies and token features.Ft-3 fine-tunes the last three layers of the ViLT

Token Type	Accuracy	Frozen	FT-3	FT-6	FT-9	FT-all
Class token	Acc@1	35.3	74.7	79.5	82.7	83.1
	Acc@5	66.5	93.4	95.1	96.5	96.5
Content token	Acc@1	69.4	78.6	82.3	82.7	82.9
	Acc@5	90.3	95.1	96.1	96.3	96.3
Fusion token	Acc@1	68.9	78.6	82.3	82.7	83.2
	Acc@5	90.1	95.0	96.0	96.4	96.5

Table 5. Comparison of transfer ability of pre-trained models with features at different stages in classification tasks. Shallow: features generated by Embedding; Frozen: features generated by frozen ViLT; Fine-tune: features generated by fine-tuned ViLT

Feature		Vireo172		NUS-WIDE			
		Acc@1	Acc@5	Prec@1	Prec@5	Recall@1	Recall@5
Image	Shallow	5.2	17.0	42.3	25.6	19.5	57.0
	Frozen	69.4	90.2	80.5	40.5	45.4	87.6
	Fine-tune	82.9	96.3	80.7	40.6	45.8	88.0
Text	Shallow	97.1	99.6	71.9	36.7	39.7	80.3
	Frozen	76.9	93.8	42.5	26.1	20.0	57.0
	Fine-tune	97.9	99.9	75.0	37.7	41.7	82.2
Multi-modal	Shallow	82.4	95.7	67.3	34.9	37.2	77.6
	Frozen	90.9	98.3	80.9	40.5	46.0	88.0
	Fine-tune	98.6	99.9	84.3	41.8	48.1	90.5

are used in the knowledge migration process, thus providing insight into the design of PMHANet on model migration.

Comparison of feature selection methods We explore the impact of fine-tuning strategy and the choice of token feature type on migration performance on the VireoFood-172 dataset, and the results are shown in Table 4.

- **At the level of the fine-tuning strategy:** We compare the effects of tuning different fine-tuning layers on the model. The fine-tuned network layers (from frozen to FT-3, FT-6) enable fine-grained extraction of features, which leads to a significant improvement in its classification performance. As the number of fine-tuned layers rises (FT-9 and FT-all), the generalization ability of shallow networks may be weakened, limiting their performance improvement. More importantly, the fine-tuning of more layers brings an increase in computational overhead. Therefore, FT-6 is chosen for our algorithm to balance the model performance and computational efficiency.
- **At the level of the feature selection strategy:** In the case of partial fine-tuning, the information obtained by class token in the shallow pre-trained knowledge deviates more from the information in the downstream task, while content token and fusion token are able to better combine the pre-trained knowledge with the multi-modal information in the downstream task and therefore obtain better classification results. Since the content token already has sufficient representational power, it is used in this model.

Table 6. Experimental comparison of semantic guidance modalities on the ability to represent visual features.

Approaches	Vireo172		NUS-WIDE			
	Acc@1	Acc@5	Prec@1	Prec@5	Recall@1	Recall@5
Deep PI	81.4	96.1	80.7	40.5	45.6	88.0
Shallow PI	81.5	96.2	80.7	40.5	45.7	88.0
Concat	83.7	96.8	81.6	41.1	46.3	88.8

Analysis of cross-modal migration ability In this section, we analyze the experimental results of image, text and multimodal data for the features extracted from different layers to analyze the cross-modal knowledge transfer capability of ViLT, as shown in Table 5.Due to the diversity of image data the shallow features lack category-related information and thus have poor performance. The interaction space information provided by the pre-trained interaction network and further extraction of image features can improve the prediction of the model. However, due to the problem of data distribution differences in knowledge migration, the model still needs to be fine-tuned to achieve better classification performance. Compared to images, text features mapped by semantic embedding can obtain sufficient classification information. However, the hybrid knowledge of the pre-trained interaction network makes the representation of text features less capable due to the different value range and distributions of heterogeneous modal features. The use of multimodal fusion features consistent with pre-training can fully utilize the pre-training just to achieve the best performance.

Comparison of semantic guidance approaches In this section, we conduct experiments to compare the effects of different semantic guidance approaches on image features, as shown in Table 6. The Deep PI method passes different modal data asynchronously through the feature extraction network to map to the same feature space, FTN Module is affected by the optimization of text classification and degrades the image classification performance. The Shallow PI method achieves semantic enhancement of visual features by directly aligning shallow text features with deep image features and avoids the influence of text on FTN Module. The best results can be achieved by stitching image and text features together according to the ViLT settings, making full use of pre-training knowledge, and achieving feature fusion and interaction through self-attention in a shallow network.

5 Conclusion

In this paper, we propose an image classification algorithm PMHANet that combines multimodal pre-trained large model and heterogeneous feature alignment, and propose a "fine-tuning" paradigm for the pre-trained large model. The experimental results show that partial heterogeneous feature alignment can further improve the image representation capability of the model. Future work in this research is focused on two directions.First, the learning capability of the model for multimodal interaction representations is improved by introducing higher-order graph structure information between image and text information. Second, causal inference techniques can help the model learn the relationship between context and target, combined with the mapping alignment of heterogeneous modalities to further alleviate the semantic gap problem between modalities.

Acknowledgments

This work is supported in part by the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048) and the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073)

References

1. Chen, J., Ngo, C.W.: Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 32–41 (2016)
2. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. pp. 1–9 (2009)
3. Chung, Y.A., Weng, W.H., Tong, S., Glass, J.: Unsupervised cross-modal alignment of speech and text embedding spaces. Advances in neural information processing systems **31** (2018)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Huang, H., Liang, Y., Duan, N., Gong, M., Shou, L., Jiang, D., Zhou, M.: Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. arXiv preprint arXiv:1909.00964 (2019)
7. Iki, T., Aizawa, A.: Effect of visual extensions on natural language understanding in vision-and-language models. arXiv preprint arXiv:2104.08066 (2021)
8. Jiang, S., Min, W., Liu, L., Luo, Z.: Multi-scale multi-view deep feature aggregation for food recognition. IEEE Transactions on Image Processing **29**, 265–276 (2019)
9. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
10. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
12. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
13. Martinel, N., Foresti, G.L., Micheloni, C.: Wide-slice residual networks for food recognition. In: 2018 IEEE Winter Conference on applications of computer vision (WACV). pp. 567–576. IEEE (2018)
14. Meng, L., Chen, L., Yang, X., Tao, D., Zhang, H., Miao, C., Chua, T.S.: Learning using privileged information for food recognition. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 557–565 (2019)

15. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 (2020)
16. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: European conference on computer vision. pp. 443–450. Springer (2016)
17. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
18. Tang, J., Shu, X., Li, Z., Qi, G.J., Wang, J.: Generalized deep transfer networks for knowledge propagation in heterogeneous domains (2016)
19. Tang, J., Shu, X., Qi, G.J., Li, Z., Wang, M., Yan, S., Jain, R.: Tri-clustered tensor completion for social-aware image tag refinement. IEEE transactions on pattern analysis and machine intelligence **39**(8), 1662–1674 (2016)
20. Tang, Z., Cho, J., Tan, H., Bansal, M.: Vidlankd: Improving language understanding via video-distilled knowledge transfer. Advances in Neural Information Processing Systems **34** (2021)
21. Wang, J., Wang, H., Deng, J., Wu, W., Zhang, D.: Efficientclip: Efficient cross-modal pre-training by ensemble confident learning and language modeling. arXiv preprint arXiv:2109.04699 (2021)
22. Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H.: Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3208–3216 (2021)
23. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)