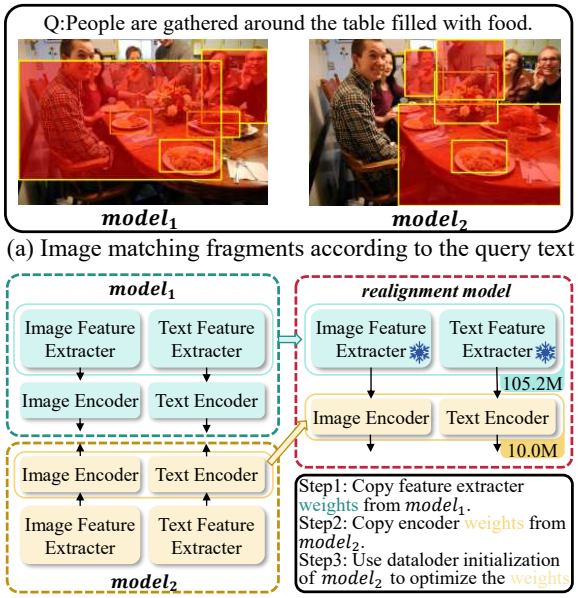
Achieving Ensemble-Like Performance in a Single Model: A Feature Diversification Framework for Image-Text Matching

Zhao Zhou[1,2], **Yiqun Wang**[1], **Weizhong Zhang**[1,3*], **Yingbin Zheng**[4], **Xiangcheng Du**[1], **Cheng Jin**[1,2*]
[1]Fudan University, Shanghai, China
[2]Innovation Center of Calligraphy and Painting Creation Technology, MCT, China
[3]Shanghai Collaborative Innovation Center of Intelligent Visual Computing, China
[4]Videt Lab, Shanghai, China
{zzhou21,yiqunwang23,xcdu22}@m.fudan.edu.cn, zyb@videt.cn, {weizhongzhang,jc}@fudan.edu.cn

Abstract

Model ensembling is a widely used technique that enhances performance in image-text matching tasks by combining multiple models, each trained with different initializations. However, the inefficiencies associated with training several models and generating outputs from them constrain their practical applicability. In this paper, we argue that while the parameters of two randomly initialized models can differ significantly, their feature distributions can be similar at certain stages. By employing a proposed technique called cross-modal realignment, we demonstrate that features derived from differently initialized models maintain similarity at the feature extraction stage and can be effectively transformed by fine-tuning a small number of parameters. These findings provide an efficient way to achieve ensemble-like performance within a single model. Specifically, we propose a Feature Diversification Framework (FDF) that emulates the outputs of multiple model initializations to generate diverse features from a common shared feature. Firstly, we introduce feature conversion methods to transform shared features into a set of distinct features. Next, a realignment training strategy is presented to optimize negative pairs for realigning these transformed features, thereby enhancing their diversification to resemble the outputs of different models. Additionally, we propose a reweighting module that assigns weights to these features, enabling a weighted fusion approach for robust feature representation. Extensive experiments on the Flickr30K and MS-COCO datasets demonstrate the effectiveness and generalizability of our framework.



(a) Image matching fragments according to the query text

(b) Pipeline for cross-model realignment

Figure 1: (a) Models trained with diverse initializations demonstrate varying degrees of cross-model alignment. (b) A cross-model realignment technique to demonstrate that features from models with distinct initializations are similar at feature extractor stage and can be effectively transformed.

Introduction

In recent years, the task of image-text matching has attracted considerable attention due to its extensive applications in multimedia retrieval (Li et al. 2022b), image captioning (Luo et al. 2021), and visual question answering (Anderson et al. 2018). Image-text matching methods encompass a wide array of techniques aimed at establishing meaningful correspondences between images and textual descriptions, which is learned through a large number of positive and negative sample pairs during training. The crucial challenge of image-text matching resides in accurately learning the semantic correspondence between images and texts for measuring their similarity.

Due to the intrinsic complexity of vision and language tasks, models trained with distinct initializations can generate diverse alignment results between images and text and biased understandings of data, as depicted in Figure 1(a). To alleviate the impacts of initialization variance and explore the peak of performance, some image-text matching methods (Chen et al. 2021; Fu et al. 2023; Li et al. 2023a; Lee et al. 2018; Zhang et al. 2022; Kim, Kim, and Kwak 2023) adopt model ensembling techniques. These methods average the similarity outputs of multiple models and capitalize on the complementary strengths of each, thereby integrating their predictions to enhance overall performance.

*Corresponding author.

However, model ensembling is inefficient as it requires the repeated training of multiple models and perform numerous forward passes as well as similarity calculations during the prediction phase. This issue raises a natural question: *Are all model parameters required, i.e., can a single model achieve ensemble-like performance?*

To answer the question, we begin by exploring the relationship between output features of models trained with different initializations. The technique of neural network grafting (Gu et al. 2020) suggests that the feature distributions of standard CNNs with random initialization are similar. However, the image-text matching task, which involves multimodal data and a more complex network design (typically divided into feature extractor and encoder stages), differs from classification tasks. To determine whether the features from different initializations in image-text matching networks remain similar at certain stages and to understand the relationship between features from different models, we propose a cross-model realignment technique to compares feature distributions across different initialization models at each stage (see Section Method for details). The central idea is to determine whether the representation learned by one model can be effectively employed by another without considerable decreasing in accuracy.

To be specific, considering that the feature extractor focuses on understanding the content of images and text, and that this stage involves a large number of parameters, which accounts for the majority of the computational load. To assess the similarity of features output by the feature extractor, as shown in Figure 1(b), we freeze the parameters of the feature extractor and optimize the encoder module (which contains only a few parameters) to realign cross-modal features according to the given data loader initialization. Through cross-model realignment, we observe that the feature distributions learned by two models with the same architecture but different initializations are almost the same during the feature extractor stage. Additionally, we find that negative pairs are crucial for model optimization and can effectively transform features between different models by leveraging the differences in negative samples.

Our surprising finding suggests that ensemble results can be more effectively achieved by focusing on feature transformation rather than trainable parameters. This insight leads us to propose transforming shared features into a set of distinct and diversified features, akin to those produced by different initializations, in order to achieve ensemble-like performance within a single model. Specifically, we introduce a novel image-text matching training framework called the Feature Diversification Framework (FDF). This framework first generates base image and text features from a shared feature extractor, then employs multiple feature conversions to derive a set of transformed features from these base features. To further diversify the transformed features, we employ a realignment training strategy that simulates the optimization process of different initializations. Finally, we introduce a reweighting module that assigns adaptability weights to each transformed feature and performs a weighted sum to create a robust representation. Experimental results demonstrate that our approach not only achieves significant improvements over existing image-text matching baselines but also enhances model stability. Notably, our framework exhibits strong generalizability, enabling seamless integration with various image-text matching methodologies without extensive modifications.

Our main contributions can be summarized as follows:

- We propose a cross-model realignment technique that demonstrates features from different initialization models are similar at the feature extractor stage. Furthermore, the output features can be effectively transformed between different initializations by fine-tuning a small subset of parameters.
- We introduce a novel feature diversification framework that achieves ensemble-like effects efficiently within a single model. A realignment training strategy is proposed to enhance the diversity of the transformed features, and an adaptive weight generated by the reweighting module is used to fuse the transformed features, resulting in a robust representation.
- We apply FDF on three image-text matching baselines, conducting experiments on Flickr30K and MS-COCO. The results show its robust generality and effectiveness.

Related Work

Global-Level Matching Methods. Among image-text matching methods, global-level matching methods appear earlier than the local-level ones and stay active all along the time. Generally speaking, these methods focus on understanding the holistic semantic content of an image or a text and measuring the degree of semantic similarity between them. To achieve a better result, some methods aim to better contextualize and aggregate multi-modal features into holistic embeddings (Chen et al. 2021; Fu et al. 2023). For example, VSE∞ (Chen et al. 2021) uses a Generalized Pooling Operator (GPO) to automatically adapt to the best pooling strategy for different features; HREM (Fu et al. 2023) explicitly captures both fragment-level and instance-level relations to learn discriminative and robust cross-modal embeddings. Besides, other methods aim at designing a better strategy to select positive and negative sample pairs during training (Faghri et al. 2018; Li et al. 2023a; Zhang et al. 2024). For example, VSE++ (Faghri et al. 2018) only uses the hardest negative sample pair within each mini-batch to train the model; FNE (Li et al. 2023a) makes the model concentrate on hard negatives by decreasing the sampling probability of both false negatives and easy negatives.

Local-Level Matching Methods. Different from global-level ones, local-level matching methods focus on the more fine-grained relationships between specific visual and textual elements within the images and texts, and thus always achieve more precise matching results. Early methods like SCAN (Lee et al. 2018) first introduce the calculation process of the similarity between regions within an image and words within a text, which motivates various subsequent methods. Most recent works aim at better modeling the region-word relationships and further aligning the semantics (Zhang et al. 2022; Kim, Kim, and Kwak 2023; Pan, Wu, and Zhang 2023). For example, NAAF (Zhang et al.

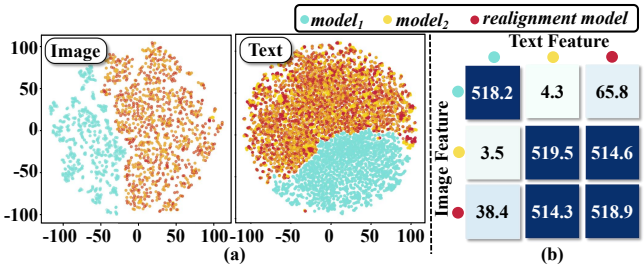


Figure 2: The experimental result of the proposed cross-model realignment. (a) The t-SNE (Van der Maaten and Hinton 2008) results from different models for image and text. (b) The similarity matrix for different models.

2022) explicitly exploits both the positive effect of matched fragments and the negative effect of mismatched fragments to jointly infer image-text similarity; DivE (Kim, Kim, and Kwak 2023) uses slot attention mechanism to capture diverse semantics of input, and uses a new similarity function called smooth-Chamfer similarity to avoid sparse supervision and set collapsing; CHAN (Pan, Wu, and Zhang 2023) only exploits the most relevant region-word pairs and eliminates all other alignments as redundant or irrelevant ones.

Model Ensembling for Image-Text Matching. Generally speaking, model ensembling combines the predictions from multiple models together. Those models can have different architectures, and ensembling methods include voting, averaging, boosting, etc. In the field of image-text matching, the simplest model ensemble technique is always used, namely averaging the output of multiple models with the same architecture (Li et al. 2023a; Zhang et al. 2022). Experimental results show that this method can always bring an obvious improvement of about 5% on the RSUM metric (Chen et al. 2021; Kim, Kim, and Kwak 2023). In this paper, our aim is to achieve a competitive improvement efficiently within a single model, without training multiple models repetitively.

Method

In this section, we first outline the standard training procedure utilized in previous works. Next, we detail the cross-modal realignment technique. Finally, inspiring from the cross-modal realignment technique, we present the Feature Diversification Framework for image-text matching.

Problem Statement

Most contemporary methods for image-text matching (Pan, Wu, and Zhang 2023; Chen et al. 2021) utilize bottom-up and top-down attention (BUTD) (Anderson et al. 2018) for extracting image features and employ either a bi-directional gated recurrent unit (BiGRU) or a pre-trained BERT (Devlin et al. 2019) for extracting text features. Formally, the visual features $\{f_k^v\}_{k=1}^K$ are extracted by a visual feature extractor, with each feature f_k^v capturing the semantic of the k-th salient region in an image. The variable K signifies the total number of salient regions. Similarly, the text features $\{f_l^t\}_{l=1}^L$ are generated by a text feature extractor, with each feature f_l^t representing the l-th word in a text. The length of the text is denoted by L. Previous works (Pan, Wu, and Zhang 2023; Chen et al. 2021; Kim, Kim, and Kwak 2023; Fu et al. 2023) have designed an image and text encoder architecture to enhance the features and embed them into d-dimensional vectors. This encoder process can be summarized as:

$$V = \phi^v(\{f_k^v\}_{k=1}^K), \;\; T = \phi^t(\{f_l^t\}_{l=1}^L). \tag{1}$$

Here, ϕ^t and ϕ^v symbolize the encoder functions for text and visual data, respectively.

To training the network, existing methodologies (Pan, Wu, and Zhang 2023; Chen et al. 2021; Lee et al. 2018) adopt the hinge-based bi-directional triplet ranking loss, integrated with online hard negative mining as proposed by VSE++ (Chen et al. 2021). Triplet loss seeks to minimize the distance between anchor and positive samples while maximizing the distance between anchor and negative samples. And online hard negative mining selects the negative sample with the highest matching score within a minibatch for optimization. The objective function is formated as:

$$\begin{aligned} L_{\text{online}} = \sum_{(i,j)\in P} \Big(& \Big[\alpha + S(\hat{i}, j) - S(i,j)\Big]_+ \\ & + \Big[\alpha + S(i, \hat{j}) - S(i,j)\Big]_+ \Big), \end{aligned} \tag{2}$$

where α denotes the margin parameter, (i, j) represents a positive image-text pairs within minibatch, and $[x]_+$ is defined as $\max(x, 0)$. Additionally, $\hat{i}$ and $\hat{j}$ refer to the most challenging negative image and text within a training minibatch, respectively. The similarity S is employed to measure the distance between the cross-model features. Randomly initialization directly influence the sampling of positive and negative pairs, which is crucial for triplet loss optimization and can lead to a biased understanding of the data.

Cross-Modal Realignment

To explore the relationship between output features of models trained with different initializations, we propose a cross-modal realignment technique. This technique examines whether two models with different initializations learn similar feature representations at the feature extraction stage and whether the output features can be easily transformed across different initializations. The intuition is that if two models, $model_1$ and $model_2$, learn similar feature representations, then features extracted by $model_1$ can be used in $model_2$ with minimal increase in error. The pipeline for cross-modal realignment is illustrated in Figure 1(b). In this process, the parameters of the image and text feature extractors in $model_1$ are copied and frozen during the realignment model training. This implies that the features from the feature extractor stage are fixed and denoted as $F^v_{model_1}$ and $F^t_{model_1}$. Meanwhile, the parameters of the image and text encoders are initialized from $model_2$. The output of realigment model can be format as:

$$\begin{aligned} V_{\text{realign}} &= \phi^v(W^v_{model_2}; F^v_{model_1}), \\ T_{\text{realign}} &= \phi^t(W^t_{model_2}; F^t_{model_1}). \end{aligned} \tag{3}$$

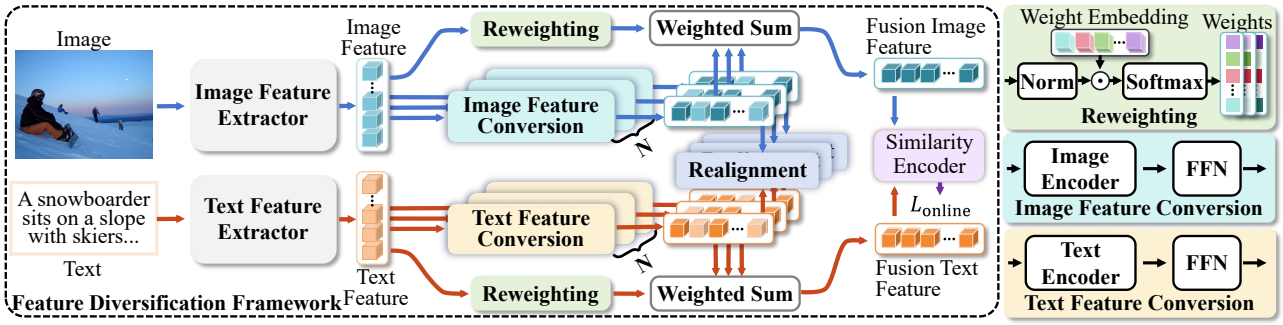


Figure 3: The Feature Diversification Framework (FDF). Here, we exemplify the local-level matching method. N represents the number of image or text feature conversion involved in the FDF.

Here $W^v_{model_2}$ and $W^t_{model_2}$ denote the image and text encoder parameters, which are fine-tuned using L_{online} with the data loader initialization from $model_2$.

We visual the feature outputs from three models, as shown in Figure 2(a), although features from different initialization models exhibit substantial differences (represented by the blue and yellow circles), cross-model realignment technique can get similar feature distribution (represented by the red and yellow circles). We also calculate the RSUM for image and text features from both the realignment model and different initialization models, as shown in Figure 2(b). The realignment model not only exhibits close performance but also shows high similarity with the target $model_2$.

Using this technique, we first observe that different initialization models produce similar features at the feature extractor stage, indicating that the feature extractor can be shared across models. Secondly, while cross-model features from different initializations are somewhat unrelated, they can be effectively transformed into one another by fine-tuning only a small portion of the parameters. Additionally, modifying the sampling of positive and negative pairs can help guide the feature transformation toward a specified distribution. This finding suggests a new approach: using a single feature extraction model to generate base features, from which diversified features can be derived to align with the outputs of other models with different initializations. This approach addresses the limitations of model ensembling.

Feature Diversification Framework

Based on cross-modal realignment, we propose a novel image-text training framework called feature diversification framework (FDF). As illustrated in Figure 3, the pipeline of FDF involves generating cross-modal features from image and text feature extractors. Feature conversion is then applied to transform the base features into multiple variants. The realignment training strategy further diversifies these transformed features by adjusting the negative pairs used for optimization. Finally, the reweighting module estimates the weights of the transformed features and generates the fused features through a weighted sum. Details of the proposed modules are provided below.

Feature Conversion. For a given image and text feature set, $(\{f^v_k\}^K_{k=1}, \{f^t_l\}^L_{l=1})$, generated by the feature extractor, we design N image and text feature conversions $\{(\theta^v_n, \theta^t_n)\}^N_{n=1}$ to obtain the transformed features $\{(\hat{V}_n, \hat{T}_n)\}^N_{n=1}$. To enhance the feature transformation capability of these conversions, we incorporate a Feedforward Neural Network (FFN) after the encoder architecture. The structure of the encoder remains consistent with the individual baselines. The process is formatted as follows:

$$\begin{aligned}\hat{V}_n &= \theta^v_n(\{f^v_k\}^K_{k=1}) = \sigma(\text{FFN}^v_n(\phi^v_n(\{f^v_k\}^K_{k=1}))),\\ \hat{T}_n &= \theta^t_n(\{f^t_l\}^L_{l=1}) = \sigma(\text{FFN}^t_n(\phi^t_n(\{f^t_l\}^L_{l=1}))).\end{aligned} \tag{4}$$

Here, n denotes the n-th feature conversion, and σ denotes L2 normalization.

Realignment Training Strategy. Feature diversification is crucial for achieving better ensemble results. To enhance the diversification of transformed features, we propose a realignment training strategy based on the cross-model realignment technique. Our approach aims to imitate the training process of differently initialized models that exhibit significant differences in positive and negative sampling. As detailed in Algorithm 1, previous minibatch data is utilized as negative samples for training different feature conversions. After optimization with diverse negatives, the transformed features are realigned to resemble the outputs of differently

Algorithm 1: Realignment Training Strategy

Input: N minibatch data B_1, B_2, ..., B_N.
Output: Realignment Loss L_{realign}.

Utilize θ^v_1 and θ^t_1 to calculate $\{\hat{V}_1\}$ and $\{\hat{T}_1\}$ from B_1.
Compute the similarity between $\{\hat{V}_1\}$ and $\{\hat{T}_1\}$.
Identify the positive pairs P and the corresponding image hardest negatives N^v_1 and text hardest negatives N^t_1 from B_1.
Calculate L_{online} using P, N^v_1, and N^t_1 with Eq. (2).
$L_{\text{realign}} \leftarrow L_{\text{online}}$.
for $n = 2, \ldots, N$ **do**
    Calculate $\{\hat{V}_n\}$ and $\{\hat{T}_n\}$ from B_n with Eq. (4).
    Update $\{\hat{V}_1\}$ and $\{\hat{T}_1\}$ from B_1 with Eq. (4).
    Identify N^t_n from B_n with similarity of $\{\hat{T}_n\}$ and $\{\hat{V}_1\}$.
    Identify N^v_n from B_n with similarity of $\{\hat{V}_n\}$ and $\{\hat{T}_1\}$.
    Calculate L_{online} using P, N^v_n, and N^t_n with Eq. (2).
    $L_{\text{realign}} \leftarrow L_{\text{realign}} + L_{\text{online}}$.
end for
return L_{realign}.

Data Split	MS-COCO 5-fold 1K Test (Chen et al. 2015)							MS-COCO 5-fold 5K Test (Chen et al. 2015)						
Eval Task	IMG → TEXT			TEXT → IMG				IMG → TEXT			TEXT → IMG			
Method	R@1	R@5	R@10	R@1	R@5	R@10	RSUM	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
BUTD + BiGRU														
SGRAF† (Diao et al. 2021)	79.3	96.7	98.3	64.5	90.0	95.8	524.6	55.8	83.0	91.0	42.0	72.4	82.1	426.3
CGMN (Cheng et al. 2022)	76.8	95.4	98.3	63.8	90.7	95.7	520.7	53.4	81.3	89.6	41.2	71.9	82.4	419.8
NAAF† (Zhang et al. 2022)	78.1	96.1	98.6	63.5	89.6	95.3	521.2	58.9	85.2	92.0	42.5	70.9	81.4	430.9
CHAN (Pan, Wu, and Zhang 2023)	79.7	96.7	**98.7**	63.8	90.4	95.8	525.0	60.2	85.9	92.4	41.7	71.5	81.7	433.4
DivE (Kim, Kim, and Kwak 2023)	79.8	96.2	98.6	63.6	90.7	95.7	524.6	58.8	84.9	91.5	41.1	72.0	82.4	430.7
HREM (Fu et al. 2023)	80.0	96.0	**98.7**	62.7	90.1	95.4	522.8	58.9	85.3	92.1	40.0	70.6	81.2	428.1
SCAN⋆	67.6	93.3	97.6	54.6	86.3	93.3	492.9	41.8	75.1	85.5	31.7	62.7	75.0	371.8
SCAN+FDF	67.9	94.1	98.1	56.5	87.2	93.6	497.4(+4.5)	40.9	75.4	86.3	33.2	65.1	76.9	377.9(+6.1)
CHAN⋆	79.7	96.3	**98.7**	64.1	90.6	95.9	525.3	59.6	85.3	91.9	42.1	71.8	82.0	432.7
CHAN+FDF	**80.9**	**96.9**	**98.7**	**65.5**	**91.1**	**96.2**	**529.3(+4.0)**	**61.2**	**87.0**	**93.2**	**43.9**	**73.1**	**82.9**	**441.4(+8.7)**
BUTD + BERT														
VSE∞ (Chen et al. 2021)	79.7	96.4	98.9	64.8	91.4	96.3	527.5	58.3	85.3	92.3	42.4	72.7	83.2	434.3
TERAN† (Messina et al. 2021)	80.2	96.6	99.0	67.0	92.2	**96.9**	531.9	59.3	85.8	92.4	45.1	**76.4**	84.4	443.4
CHAN (Pan, Wu, and Zhang 2023)	81.4	**96.9**	98.9	66.5	92.1	96.7	532.6	59.8	87.2	93.3	44.9	74.5	84.2	443.9
HREM (Fu et al. 2023)	81.1	96.6	98.9	66.1	91.6	96.5	530.7	62.3	**87.6**	93.4	43.9	73.6	83.3	444.1
USER (Zhang et al. 2024)	**82.8**	96.8	98.8	66.1	90.6	95.6	530.5	**63.7**	87.4	**93.5**	44.8	73.4	82.7	445.5
DCIN (Li et al. 2023b)	80.9	96.5	98.8	65.1	91.5	96.3	529.1	59.8	85.8	92.4	42.9	73.5	83.6	438.0
VSE∞⋆	79.2	96.6	98.9	64.7	91.2	96.2	526.8	58.4	85.2	92.0	42.5	72.7	83.0	433.9
VSE∞+FDF	81.1	96.4	98.9	65.4	91.6	96.3	529.7(+2.9)	59.4	85.9	92.7	42.9	73.4	83.6	437.9(+4.0)
CHAN⋆	80.7	96.8	**99.1**	66.2	91.9	96.6	531.3	59.7	86.8	93.2	44.4	74.1	83.9	441.9
CHAN+FDF	81.1	96.8	99.0	**67.1**	**92.5**	96.8	**533.3(+2.0)**	61.2	87.4	93.2	**45.8**	75.1	**84.8**	**447.6(+5.7)**

†Ensemble models of two hypotheses. ⋆The results are obtained from model re-trained with code provided in paper.

Table 1: Image-text retrieval results on MS-COCO. Bold and underlined texts denote the top and the runner-up, respectively.

initialized trained models. This enhances the diversification of different transformed features and provides a solid foundation for generating robust feature representations in the fusion stage.

Reweighting Module. Previous approaches (Zhang et al. 2022; Li et al. 2019; Diao et al. 2021) employed an ensemble strategy that computed the mean score of two output similarities, treating the two models as having equal weights. However, we argue that distinct transformed features may have varying degrees of influence on the final output for each visual or text feature. Therefore, we propose the utilization of a reweighting module to dynamically learn individualized weights for each feature from different feature conversions. As shown in Figure 3, cross-model trainable weight embedding, denoted as $E_v \in \mathbb{R}^{d\times N}$ and $E_t \in \mathbb{R}^{d\times N}$, are designed to discern the contribution of each transformed features based on the base feature. A set of cross-model transformed feature weights, γ^v and γ^t, are generated as follow:

$$\begin{aligned}\gamma^v &= \text{softmax}(\sigma(\{f_k^v\}_{k=1}^K)\cdot E_v),\\ \gamma^t &= \text{softmax}(\sigma(\{f_l^t\}_{l=1}^L)\cdot E_t).\end{aligned} \tag{5}$$

Finally, the image and text fusion features are obtained by a weighted sum of the transformed features using γ^v and γ^t, and are optimized through L_{online}.

In summary, our final loss formulation is as follows:

$$L = L_{\text{realign}} + \lambda_1 L_{\text{online}}, \tag{6}$$

where λ_1 is scaling coefficients.

Experiments

Experimental Setup

Datasets. We selected two widely-used datasets for our experimentation: Flickr30K (Young et al. 2014) and MS-COCO (Chen et al. 2015). The MS-COCO dataset comprises 123,287 images, each accompanied by 5 annotated captions. Our data partitioning adheres to established practices (Faghri et al. 2018; Lee et al. 2018), allocating 113,287 images for training, 5,000 for validation, and 5,000 for testing. We ensure robustness by reporting results averaged over 5 folds of 1,000 test images and validated on entire 5,000-image test set. The Flickr30K dataset comprises 31,783 images obtained from Flickr platform, with each image meticulously paired with five corresponding captions. Within the Flickr30K dataset, 29,000 images are allocated for training, 1,000 for testing, and 1,014 for validation purposes.

Evaluation Metrics. In accordance with conventional information retrieval protocols, we evaluate performance using $R@K$, denoting the percentage of accurately matched queries within the top K retrieved instances. Elevated $R@K$ values signify superior performance. For a thorough evaluation and in line with previous methods, we aggregate all recall values into RSUM, covering both image-to-text and text-to-image matching.

Implementation Details. To ensure a comprehensive and equitable evaluation, we maintain the network architectures and configurations of all baseline methods exactly as detailed in their paper. The default setting of λ_1 is set to 0.5. The number of feature conversions N is fixed at 8 across all baseline experiments.

Main Results

We assess the effectiveness and generalization capability of FDF across three image-text matching methods (Pan, Wu, and Zhang 2023; Chen et al. 2021; Lee et al. 2018), as depicted in Table 1 and Table 2. For the global-level matching method, we apply FDF to VSE∞. Remarkably, both

Eval Task	IMG → TEXT			TEXT → IMG			
Method	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
BUTD + BiGRU							
SGRAF†	78.4	94.6	97.5	58.2	83.0	89.1	500.8
CGMN	77.9	93.8	96.8	59.9	85.1	90.6	504.1
NAAF†	79.6	**96.3**	**98.3**	59.3	83.9	90.2	507.6
CHAN	79.7	94.5	97.3	60.2	85.3	90.7	507.8
DivE	77.8	94.0	97.5	57.5	84.0	90.0	500.8
HREM	79.5	94.3	97.4	59.3	85.1	91.2	506.8
SCAN⋆	67.0	89.9	94.8	43.6	73.7	82.6	451.6
SCAN+FDF	67.9	89.6	94.1	44.9	74.1	83.2	454.8(+3.2)
CHAN⋆	78.1	94.8	97.7	59.6	84.9	90.5	505.6
CHAN+FDF	**82.1**	95.0	97.9	**62.6**	**86.0**	**91.5**	**515.1(+9.5)**
BUTD + BERT							
VSE∞	81.7	95.4	97.6	61.4	85.9	91.5	513.5
TERAN†	79.2	94.4	96.8	63.1	87.3	92.6	513.4
CHAN	80.6	96.1	97.8	63.9	87.5	92.6	518.5
HREM	83.3	96.0	98.1	63.5	87.1	92.4	520.4
USER	82.7	**97.0**	98.3	63.1	86.7	92.1	519.9
DCIN	83.0	96.4	**98.6**	63.3	87.8	92.4	521.5
VSE∞⋆	80.1	95.5	97.8	62.1	86.4	91.9	513.8
VSE∞+FDF	82.3	95.9	98.2	64.4	87.8	92.6	521.1(+7.3)
CHAN⋆	81.7	95.6	97.5	64.1	87.4	92.3	518.6
CHAN+FDF	**84.1**	96.9	98.4	**67.9**	**89.6**	**93.9**	**530.8(+12.2)**

Table 2: Image-text retrieval results on Flickr30K.

	IMG → TEXT			TEXT → IMG			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
Baseline	81.7	95.6	97.5	64.1	87.4	92.3	518.6
+ FC	81.5	**96.7**	98.1	64.6	87.8	92.8	521.5
+ RTS	81.8	96.6	**98.4**	**68.1**	88.8	93.3	527.0
+ RW	**83.5**	96.3	97.8	68.0	**89.2**	**93.8**	**528.5**

Table 3: Results with different model components.

Flickr30K and MS-COCO 5-fold 1K datasets exhibit gains of 7.3% and 2.9% RSUM over the respective baselines. The MS-COCO 5-fold 5K dataset experiences a notable performance boost from 433.9 to 437.9 RSUM. Additionally, experiments were conducted on the famous local-level matching method SCAN, resulting in observed gains of +3.2% RSUM on the Flickr30K dataset, and +4.5% and +6.1% on the MS-COCO dataset. For the state-of-the-art local-level method CHAN, integration with our framework yields notable improvements. Specifically, on the Flickr30K dataset employing text feature extractor BERT, CHAN achieves a RSUM of 515.1, surpassing the baseline by +10.1% RSUM. On the MS-COCO datasets with text feature extractor BERT and BiGRU, significant performance enhancements are observed. With the BiGRU feature extractor, the performance of CHAN on the MS-COCO dataset 5-fold 5K setting sees a substantial boost from 432.7 to 441.4. Overall, our results demonstrate significant improvements across both basic and state-of-the-art baseline methods upon integrating our framework. Notably, on the smaller Flickr30K dataset, the improvements are particularly pronounced, underscoring the effectiveness of our approach.

Comparisons with the State-of-the-arts

We integrate our FDF with the CHAN model to compare with recent state-of-the-art methods across two benchmark datasets. Unlike methodologies (Chen et al. 2020; Zhang et al. 2022; Li et al. 2022a) that enhance performance through ensemble techniques by averaging similarities from two models, our approach focuses on single-model retrieval results. To ensure a fair comparison, we categorize the methods based on their feature extraction backbones.

Table 2 presents the quantitative outcomes of our approach on the Flickr30K test set. Remarkably, CHAN with our FDF outperforms all other methods, achieving RSUM scores of 515.1 for BiGRU-based and 530.8 for BERT-based approaches. Compared to the baseline CHAN model, our BiGRU-based and BERT-based FDF yield notable improvements exceeding 5.1% and 7.6% across sum of $R@1$, $R@5$, and $R@10$ for text-image retrieval. Moreover, our FDF surpasses other state-of-the-art methods by significant margins, achieving improvements of 7.3% and 9.3% in RSUM with BiGRU and BERT feature extraction, respectively.

The quantitative comparison results on the larger and more intricate MS-COCO dataset are depicted in Table 1. Notably, our BiGRU-based FDF exhibits superior performance compared to recent counterparts such as HREM and Dive across both MS-COCO 5-fold 1K and MS-COCO 5K test sets. Furthermore, for BERT-based models, our FDF demonstrates slightly better results than the Moco-based method USER, as evidenced towards the bottom of Table 1. The experimental results illustrate that the CHAN model trained with our framework outperforms other models and achieves superior results.

Ablation Study

We perform an ablation study on the Flickr30K dataset using the BERT backbone. The evaluations are conducted with the CHAN model and 2 feature conversions by default.

Effects of Model Components. We systematically examined the impacts of the feature conversion (FC), realigment training strategy (RTS), and reweighting module (RW) components on the overall model performance. For experiments without a reweighting module, we directly averaged the transformed features to obtain final features. Firstly, the incorporation of the realigment training strategy yielded a significant performance enhancement, achieving an RSUM of 527.0, marking a notable improvement of 8.4% over the baseline. This underscores the importance of feature diversification for achieving optimal performance and highlights the effectiveness of the realignment training strategy. Lastly, the reweighting module contributed an additional 1.5% improvement to RSUM on the superior model results by learning distinct weights for each transformed features. This highlights the capacity of the reweighting module to refine model performance by effectively adjusting feature weights from different feature conversions.

Feature Conversion Structure. As shown in Table 4, we explore the effectiveness of various feature conversion structures within our FDF. Initially, without additional structure, we achieved an RSUM of 523.6, representing a 5.0%

	IMG → TEXT			TEXT → IMG			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
None	82.6	95.9	98.0	66.1	88.1	92.9	523.6
Adapter	81.3	**96.4**	**98.4**	67.6	88.4	93.5	525.6
FFN	**83.5**	96.3	97.8	**68.0**	**89.2**	**93.8**	**528.5**

Table 4: Results with different feature conversion structure

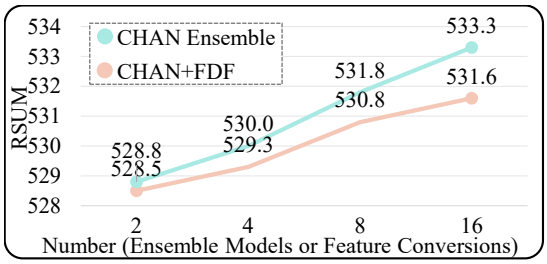


Figure 4: Comparison of model ensembling and FDF.

improvement over the baseline. Introducing the adapter structure yields a notable performance boost of 7.0% for the entire model, underscoring the significant enhancement achievable by augmenting feature conversion feature extraction capabilities. This improvement highlights the importance of designing and refining the architecture to leverage the strengths of individual feature conversions. Furthermore, we evaluated the performance of the Feed-Forward Networks (FFN) structure, getting a 1.9% RSUM improvement compared to the adapter structure. This highlights the potential for continuous performance enhancement through the expansion of feature conversion parameters. To preserve computational efficiency and reasoning effectiveness, we only consider the integration of smaller structures within the feature conversion architecture.

Compare with Model Ensembling. In Figure 4, We investigate the effectiveness of employing varying numbers of feature conversions within our method and compare it with ensemble methods. Compared to the baseline CHAN, our FDF with 2 feature conversions achieves a notable 9.9% improvement in RSUM, delivering performance comparable to an ensemble of 2 models. As the number of feature conversions increases, performance improves, and our method yields competitive results compared to model ensembling with the same number of models. Notably, the increase in parameters for our method is only about 10% of the cost of preparing a new model. With a similar number of parameters, our approach using 8 feature conversions outperforms an ensemble of 2 models.

Discussion

Efficiency and Stability Analysis. We first analyzed the inference time of our proposed FDF. Since the output feature dimension of our framework matches that of the baseline CHAN, we only computed the time taken for feature generation, excluding the time required for similarity comparison. As illustrated in Figure 5 (a), the inference time was computed by summing the inference times for 5,000 image-

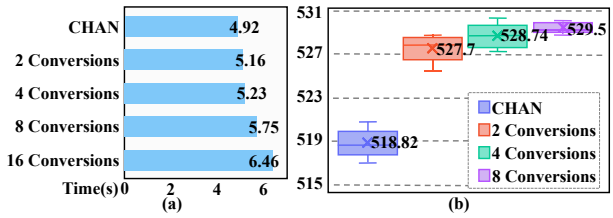


Figure 5: (a) The inference time relative to the baseline across varying numbers of feature conversions; (b) The RSUM distribution is analyzed for both baseline and varying numbers of feature conversions on the FDF , employing different initialization.

text pairs from the Flickr30K dataset with a batch size of 128. With only 2 feature conversions, our method showed a marginal increase of 0.24 seconds in total time and 0.048 milliseconds per pair. This increase accounts for only 4.9% of the total time compared to the baseline, owing to the lightweight design of our modules. Regarding model stability, we evaluated 5 sets of model weights trained with random seeds for each configuration. As depicted in Figure 5 (b), we observed that employing 2 feature conversions led to an 8.88% improvement in mean RSUM and a reduction in RSUM variance. Moreover, increasing the number of feature conversions resulted in further improvements in mean RSUM and increased stability. This trend suggests that FDF fosters a robust understanding of image-text alignment and mitigates the biases arising from different initializations.

Limitation. A limitation of our approach is the extended training duration, which results from the need to concurrently optimize multiple feature conversions. Although we only added lightweight parameters for these feature conversions compared to the baseline, training still requires optimization of these components. Although our framework increases training time by 14% compared to baseline models, our inference time remains close to that of a single model while achieving ensemble-like performance.

Conclusion

In this paper, we presented a novel framework named Feature Diversification Framework (FDF) for image-text matching, achieving ensemble-like performance in a single model. In our framework, a novel realignment training strategy was proposed to optimize the model across a range of diverse sample configurations simultaneously, instead of training multiple different models separately and repeatedly. After a weighted ensemble, our method enhanced the understanding of image-text relationships, leading to improved model performance. Extensive experimentation on both the Flickr30K and MS-COCO datasets demonstrated the efficacy of our framework, leading to state-of-the-art results in image-text matching tasks. Ablation studies further underscored the robustness and versatility of our approach across various configurations and settings. Looking ahead, we intend to extend our framework to encompass a wider range of vision and language tasks driven by metric learning.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62472097) and National Archives Administration of China Research Program (2024-X-013). The computations in this research were performed using the CFFF platform of Fudan University.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; and Han, J. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12655–12663.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15789–15798.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Cheng, Y.; Zhu, X.; Qian, J.; Wen, F.; and Liu, P. 2022. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4): 1–23.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *AAAI Conference on Artificial Intelligence*, volume 35, 1218–1226.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*.

Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning semantic relationship among instances for image-text matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15159–15168.

Gu, Y.; Zhang, W.; Fang, C.; Lee, J. D.; and Zhang, T. 2020. How to characterize the landscape of overparameterized convolutional neural networks. *Advances in Neural Information Processing Systems*, 33: 3797–3807.

Kim, D.; Kim, N.; and Kwak, S. 2023. Improving cross-modal retrieval with set of diverse embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23422–23431.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *European Conference on Computer Vision*, 201–216.

Li, H.; Bin, Y.; Liao, J.; Yang, Y.; and Shen, H. T. 2023a. Your negative may not be true negative: Boosting image-text matching with false negative elimination. In *ACM International Conference on Multimedia*, 924–934.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual semantic reasoning for image-text matching. In *IEEE/CVF International Conference on Computer Vision*, 4654–4662.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2022a. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 641–656.

Li, P.; Xie, H.; Ge, J.; Zhang, L.; Min, S.; and Zhang, Y. 2022b. Dual-stream knowledge-preserving hashing for unsupervised video retrieval. In *European Conference on Computer Vision*, 181–197. Springer.

Li, W.; Su, X.; Song, D.; Wang, L.; Zhang, K.; and Liu, A.-A. 2023b. Towards Deconfounded Image-Text Matching with Causal Inference. In *ACM International Conference on Multimedia*, 6264–6273.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.-W.; and Ji, R. 2021. Dual-level collaborative transformer for image captioning. In *AAAI Conference on Artificial Intelligence*, volume 35, 2286–2293.

Messina, N.; Amato, G.; Esuli, A.; Falchi, F.; Gennaro, C.; and Marchand-Maillet, S. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(4): 1–23.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19275–19284.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022. Negative-aware attention framework for image-text matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15661–15670.

Zhang, Y.; Ji, Z.; Wang, D.; Pang, Y.; and Li, X. 2024. USER: Unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE Transactions on Image Processing*.