Cross-Modal Feature Distribution Calibration for Few-Shot Visual Question Answering

Jing Zhang*†, Xiaoqiang Liu*, Mingzhe Chen, Zhe Wang†

Department of Computer Science and Engineering, East China University of Science and Technology, China
jingzhang@ecust.edu.cn, 2681332916lxq@gmail.com, ecustcmz@gmail.com, wangzhe@ecust.edu.cn

Abstract

Few-shot Visual Question Answering (VQA) realizes few-shot cross-modal learning, which is an emerging and challenging task in computer vision. Currently, most of the few-shot VQA methods are confined to simply extending few-shot classification methods to cross-modal tasks while ignoring the spatial distribution properties of multimodal features and cross-modal information interaction. To address this problem, we propose a novel **C**ross-modal feature **D**istribution **C**alibration **I**nference **N**etwork (CDCIN) in this paper, where a new concept named visual information entropy is proposed to realize multimodal features distribution calibration by cross-modal information interaction for more effective few-shot VQA. Visual information entropy is a statistical variable that represents the spatial distribution of visual features guided by the question, which is aligned before and after the reasoning process to mitigate redundant information and improve multi-modal features by our proposed visual information entropy calibration module. To further enhance the inference ability of cross-modal features, we additionally propose a novel pre-training method, where the reasoning sub-network of CDCIN is pretrained on the base class in a VQA classification paradigm and fine-tuned on the few-shot VQA datasets. Extensive experiments demonstrate that our proposed CDCIN achieves excellent performance on few-shot VQA and outperforms state-of-the-art methods on three widely used benchmark datasets.

Introduction

With the booming development of deep learning in recent years, all kinds of Visual Language Processing (VLP) tasks have attracted widespread attention from researchers, such as image captioning (Pan et al. 2020), visual entailment (Tran et al. 2022), Visual Question Answering (VQA) (Jiang et al. 2020a; Penamakuri et al. 2023; Dancette et al. 2023) and so on. As an important topic in VLP, VQA is a typical cross-modal problem that needs to analyze visual content and question semantics simultaneously. Currently, it is typically viewed as a classification problem where the goal is to predict the accurate answer given a pair of images and questions.

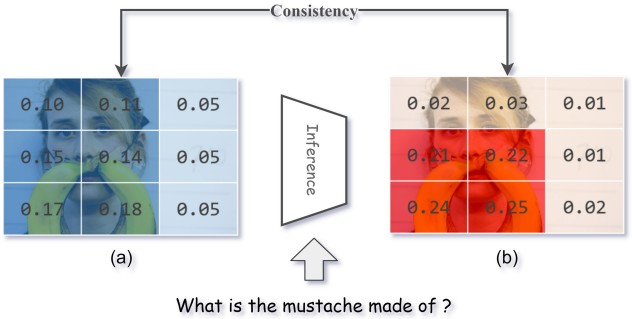

Figure 1: Visual information distribution of different inference stages. The shade of the color means that the information in that region is likely to be caught by the model.

The early joint embedding models (Fukui et al. 2016; Kim et al. 2016; Ben-Younes et al. 2019) for VQA focused on the fusion of multimodal features and cross-modal interactions, most of which were realized through attention mechanism (Yu et al. 2018; Kim, Jun, and Zhang 2018; Guo, Yao, and Chu 2023) and graph neural network (Huang et al. 2020; Li et al. 2019). These classical VQA methods are generally trained on a large amount of labeled multimodal data and ignore the sparsity problem in most categories caused by the diversity of multimodal data. As a machine learning method that aims to recognize new concepts with few samples, few-shot learning caters to the characteristics of the VQA task and trains an effective model with a small amount of labeled data. For this reason, some methods (Dong et al. 2018; Yin et al. 2021) are proposed, which attempt to apply few-shot learning to solve few-shot cross-modal learning. However, these methods cannot effectively deal with cross-modal information inference and constrain multimodal feature distribution, which limits the performance of few-shot VQA.

Cross-modal semantic inference is capable of facilitating joint reasoning based on correlation analysis between modalities, which is important for few-shot VQA. For a given image and question, the operation of few-shot VQA can be divided into two steps: understanding and reasoning. The performance of "inference" becomes crucial for answer prediction when the process of "understanding" can be well completed by existing encoders such as ViT (Dosovitskiy

*These authors contributed equally.
†It is on behalf of the corresponding author.

et al. 2021), SwinT (Liu et al. 2021b) and so on. Cross-modal "inference" aims to explore the relationships among multimodal data and use one modality to guide the filtering and enhancement of features in another modality.

The general few-shot VQA methods barely have the ability to perform cross-modal semantic inference, hence the semantics of the question cannot be effectively used to guide the visual encoding, resulting in visual encoding that may focus on unimportant regions. For example, the information distribution in the Figure 1 (a) means that the general model cannot accurately explore the relationship between the question word "mustache" and the visual object without cross-modal inference. It captures the blue region associated with the concept "person". Obviously, the blue region contains noise, which makes the spatial distribution of features far from the corresponding categories. To address this problem, the model requires learning how to discriminate redundant information and adjust the feature distribution to a more reasonable state. As shown in the Figure 1 (b), if the model catches critical visual region "mustache" covered by the red and removes the unrelated region, it will narrow the distance of the multimodal features among the samples of the same category in the feature space.

Based on the above analysis, we propose a Cross-modal feature Distribution Calibration Inference Network (CDCIN) for few-shot VQA in this paper. A novel concept of visual information entropy is defined, which is a statistical form similar to information entropy in information theory. It defines the entropy (average self-information) of visual information discrete source under the condition of a given question and reflects the spatial information distribution of visual features guided by the question. We also propose a Visual Information Entropy Calibration Module (VIECM) to realize the alignment of visual information entropy, in which the consistency of the visual information entropy from pre-inference (before multimodal relationship interaction) and post-inference (after multimodal relationship interaction) is used to realize information filtering and feature distribution calibrating. Additionally, we also propose an inference enhancement pre-training strategy, which strengthens the representation ability of multimodal features by pre-training on a classic VQA paradigm. Extensive experiments on three benchmark datasets demonstrate that our proposed CDCIN performs excellently and outperforms state-of-the-art approaches.

Related Work

Although visual question answering is a hot topic in compute vision, few-shot cross-modal learning only recently starts to attract more attention and still has great research potential. In this section, we will discuss some works related to our method from the perspectives of visual question answering, few-shot learning and few-shot visual question answering.

Visual Question Answering

Visual question answering is a challenging cross-modal analysis task since it requires establishing relationships between visual and textual modalities to achieve cross-modal semantic reasoning. Most of the current methods (Zhou et al. 2015; Yu et al. 2017; Ding et al. 2022) focused on improving the fusion strategy of visual and textual features to achieve good performance. These methods usually ignore the significant guidance of question semantic information for image understanding and are not good at relationship reasoning. To address the insufficient of the above methods, some methods (Xu and Saenko 2016; Anderson et al. 2018; Pan et al. 2022) attempt to utilize the attention mechanism to reinforce cross-modal interaction. Although the above attention related models can realize certain multimodal semantic inference, it is difficult for them to reach high-level reasoning with limited semantic interactions. Therefore, some VQA works (Huang et al. 2020; Jing et al. 2022; Cao et al. 2019) are devoted to enhancing the reasoning ability of attention networks to achieve more deeply cross-modal information interaction and reasoning. The above methods can obtain good performance on classical VQA, but they suffer severe failures when encountering certain classes with small data sizes. This suggests that it is meaningful and promising work to deal with few-shot VQA by utilizing few-shot learning methods.

Few-shot Learning

Few-shot learning is an idea of solving problems, which teaches the model how to learn and enables the model to recognize new concepts with few samples. Many few-shot learning works (Jiang et al. 2020b; Zhang et al. 2022; Li, Wang, and Hu 2021) use metric-based approaches to generate prototype representations for fast training of classifiers. Some researchers (Yang et al. 2021; Ma et al. 2020) attempt to use different calibration methods to reinforce the performance of few-shot learning. While numerous existing few-shot learning methods are applicable to multimodal data, a comprehensive examination of the interplay between modalities is often lacking in many of these approaches. When dealing with cross-modal tasks that require deep interaction between visual and semantic, these methods expose their insufficiencies. Therefore, there is still significant room for the development of few-shot learning on cross-modal reasoning tasks.

Few-shot Visual Question Answering

Few-shot visual question answering aims to train an excellent VQA model with limited data, which requires outstanding cross-modal reasoning and powerful feature representation ability. Dong et al. (Dong et al. 2018) introduced few-shot learning to VQA and image captioning and proposed a fast parameter adaptation method to train the joint image-text learner. Yin et al. (Yin et al. 2021) propose a two-stage network, where each stage is responsible for intra-modal or inter-modal relation capture. They extract features at different levels by constructing visual feature maps and semantic relationship maps by a multi-layer attention mechanism.

At present, the research results on few-shot VQA are relatively few, and the studies have not attracted widespread attention. These two papers mentioned above are considered pioneering work on few-shot VQA. Although these works have solved the problems of few-shot VQA to some extent,

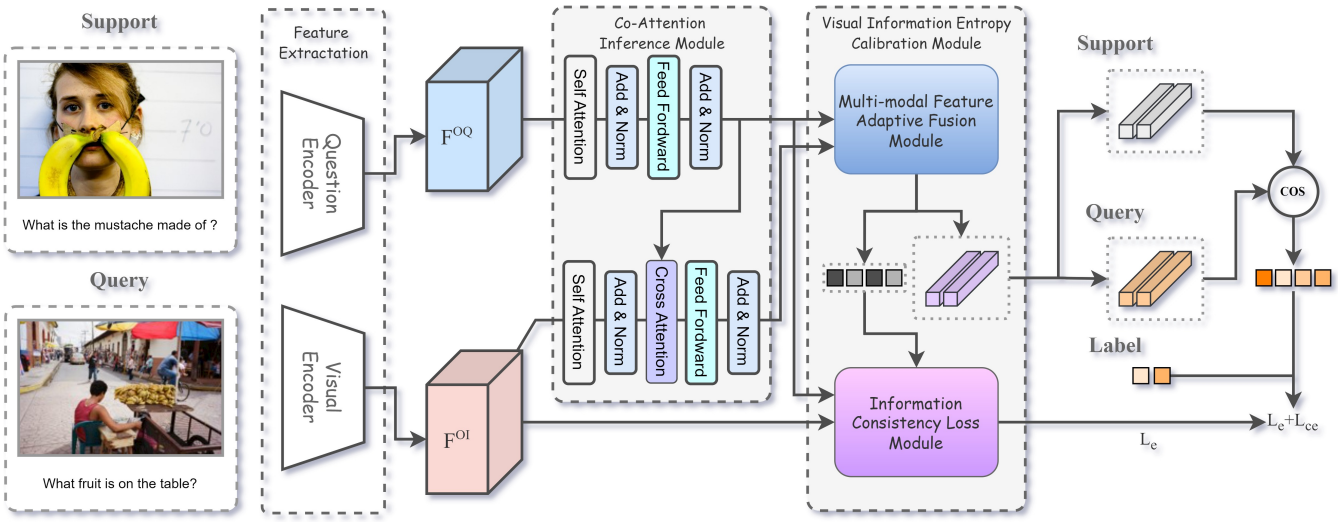

Figure 2: The framework of the cross-modal feature distribution calibration inference network. Given the support and query sets, the CDCIN extracts visual and textual features and feeds them into CAIM to model cross-modal interaction. The distribution of visual information before and after inference is aligned in the VIECM.

they fail to achieve efficient cross-modal inference to adjust the integrated multi-modal feature distribution. This leads to the dispersion of the feature distribution for each class, with considerable distances from the class center, affecting the overall performance adversely. Regarding the issue above, we propose a Cross-modal feature distribution calibration inference network for few-shot VQA that aligns the visual information entropy to enhance the ability of feature distribution calibration. Extensive experiments on widely used benchmark datasets demonstrate that the performance our method surpasses state-of-the-art few-shot VQA methods by a large margin.

Methodology

In this section, we will describe the problem definition of few-shot VQA and introduce our proposed CDCIN and pre-training method in detail.

Problem Statement

For each task τ, the few-shot VQA is formulated as a N-way K-shot classification problem with N classes sampled from the answer set and K examples per class. The answer set contains the labels of the corresponding samples. If the number of query examples for each class is M, we can get a query set $\{Q\}$ with $N \times M$ samples and a support set $\{S\}$ with $N \times K$ samples. A training sample is a triplet containing an image, a question, and an answer, so there are three query subsets $\{Q^I\}$, $\{Q^T\}$, $\{Q^A\}$ and support subsets $\{S^I\}$, $\{S^T\}$, $\{S^A\}$ actually. We combine samples from the same modality into the image set $\{I\} = \{S^I, Q^I\}$, the question set $\{T\} = \{S^T, Q^T\}$ and the answer set $\{A\} = \{S^A, Q^A\}$.

In this work, the features from $\{I\}^{N\times K+N\times M}$ and $\{T\}^{N\times K+N\times M}$ are extracted through visual embedding $\psi(\cdot;\theta_\psi)$ and text embedding $\phi(\cdot;\theta_\phi)$ neural networks. During the classification phase, the fused multimodal features are divided into corresponding support sets $\{S^{multi}\}^{N\times K}$ and query sets $\{Q^{multi}\}^{N\times M}$, and the support sets are taken as input to train the CDCIN by minimizing the loss over the corresponding query sets.

Cross-Modal Feature Distribution Calibration Inference Network

In this paper, we focus on studying deep cross-modal inference and helping the model filter out redundant information to calibrate the spatial distribution of multimodal features for few-shot VQA. To this end, we propose a Cross-modal feature Distribution Calibration Inference Network (CDCIN), as illustrated in the Figure 2. In the CDCIN, a new concept called visual information entropy is proposed to reflect the spatial distribution of visual information, which assists in excluding irrelevant information. Specifically, CDCIN mainly includes a Co-Attention Inference Module (CAIM) and a Visual Information Entropy Calibration Module (VIECM). In order to strengthen the representations of multiple modalities, we also design an Inference Enhancement Pre-training strategy.

In the CDCIN, the word tokens of the question are embedded by the pre-trained GolVe (Pennington, Socher, and Manning 2014), and the visual features are extracted by the Swin-Transformer (Liu et al. 2021b). Then the visual and question features are simultaneously input to the CAIM to mine cross-modal fine-grained interaction. The integrated features will be sent to the multimodal feature adaptive fusion module in the VIECM to generate visual information entropy and the multimodal feature vector in the later reasoning stage. Finally, the information distribution alignment is achieved through the information consistency loss.

Co-Attention Inference Existing few-shot learning methods (Ye et al. 2020; Zhang et al. 2022) usually do not consider cross-modal inference, resulting in insufficient reasoning ability. To completely realize the reasoning process of VQA under the few-shot approach, we introduce a Co-Attention Inference Module (CAIM). It is mainly composed of masked multi-head self-attention and cross attention and can realize the fine-grained interactions between images and questions.

The CAIM deeply explores the correlation between visual and textual features. Specifically, the text encoder is responsible for generating transitional features ready for cross-modal interactive operations, which mainly achieves through multi-head self-attention. Given a key and value of dimension d_k and query of dimension d_v, the attended features are obtained as follows:

$$\mathrm{SA}(Q, K, V) = \mathrm{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d}})V \tag{1}$$

where $K \in \mathbb{R}^{n\times d}$, $Q \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{n\times d}$ are key, query and value respectively.

The multi-head self-attention splits the features into parallel "heads". Each head independently performs the dot-production. The calculation is given by:

$$T = \mathrm{MHSA}(Q, K, V) = [h_i, h_2, ..., h_t]W_h \tag{2}$$

$$h_i = \mathrm{SA}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where T is attended question features, W_i^Q, W_i^K, W_i^V are the projection matrices, and $W_h \in \mathbb{R}^{h\times d_h\times d}$. d_h is the dimension of each head.

The essential component of the CAIM is the image encoder containing cross attention, which promotes the query of critical visual information according to the question, and is beneficial for the cross-modal information interaction.

Visual Information Entropy Calibration In order to maintain the consistency of visual information entropy to calibrate the feature distribution, we propose a new Visual Information Entropy Calibration Module (VIECM), which contains two sub-modules: the multimodal feature adaptive fusion module and information consistency loss module, as illustrated in the Figure. 3. The features generated by CAIM are fed into the multimodal feature adaptive fusion module to learn the multimodal features and the later visual information entropy. The original visual features extracted by Swin Transformer are fed into the information consistency loss module to compute the former visual information entropy. Finally, consistency loss is calculated from the former and later visual information entropy.

Multimodal Feature Adaptive Fusion: Few-shot learning methods (Chen et al. 2021) usually perform average pooling on the corresponding features to compute the similarity between the support set and the query set, which always results in omitting critical information. The information represented by the features $F^{n\times d}$ in the dimension of n is quite different from each other. Therefore, we propose a multimodal adaptive feature fusion module $\xi(\cdot;\theta_\xi)$, which assigns reliable weights to features of each dimension of n.

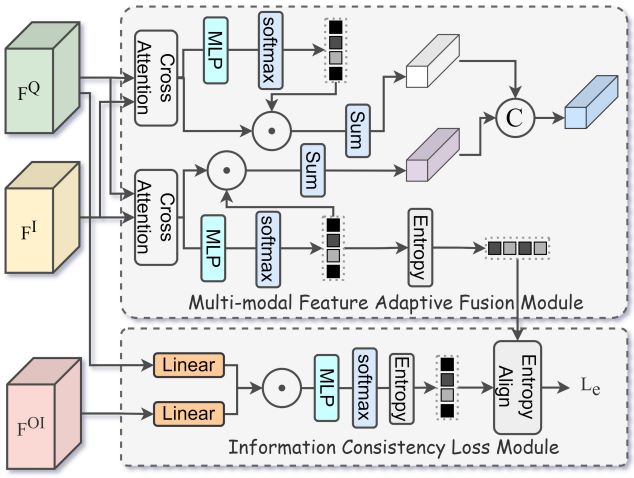

Figure 3: The illustration of the Visual Information Entropy Calibration Module.

The visual features $F^I \in \mathbb{R}^{n\times d}$ inferred by the CAIM interact deeply with the question features in the cross attention. After that, adaptive fusion weight α^I is obtained through multi-layer perception and softmax function. Then the visual features are summed up according to their relative weights to generate a flattened visual vector $f^I \in \mathbb{R}^{1\times d}$.

$$F^I = Norm(\mathrm{MHSA}(F^I, F^Q, F^Q) + F^I) \tag{4}$$

$$\alpha^I = softmax(\mathrm{MLP}(F^I)) \tag{5}$$

$$f^I = \sum_{j=1}^{n} \alpha_j^I \odot F_j^I \tag{6}$$

The flattened question features $f^Q = \xi(F^Q;\theta_\xi)$ and visual features are concatenated to compute similarity.

The weights $\alpha^I \in \mathbb{R}^{n\times 1}$ generated in the flattening stage represent the distribution of visual features after reasoning. The features with high weight have strong representation capabilities. By adjusting the weights, we can get an accurate feature distribution. To this end, we convert them into visual information entropy δ^l and feed them into the information consistency loss module.

$$\delta^l = \Delta(\alpha^I) = \sum_{i=1}^{n} \alpha_i^I log(\alpha_i^I) \tag{7}$$

Information Consistency Loss: We have discussed that there are discrepancies in information distribution at different stages of the inference network. In order to ease this gap, we utilize a loss function L_e in the information consistency loss module to constrain the information distribution of these two stages.

In the multimodal feature adaptive fusion module, we calculate the later visual information entropy δ^l for information distribution alignment. In the information consistency loss module, we generate adaptive weight from original visual features $F^{OI} \in \mathbb{R}^{n\times d}$ according to question features and converted it into visual information entropy by $\Delta(\cdot)$.

$$\delta^f = \Delta(softmax(\mathrm{MLP}(W_I F^{OI} \odot W_Q F^Q))) \tag{8}$$

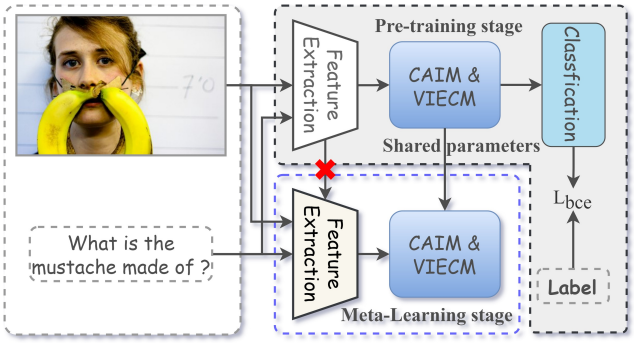

Figure 4: The illustration of the Inference Enhancement Pre-training. The components of the same color share parameters.

where W_I, W_Q are the projection matrices. Our method is trained to minimize the difference in visual information entropy between stages of inference to reinforce the convergence of the feature distribution. We define an information consistency loss function to calculate the squared difference of later and former visual information entropy, denoted as L_e:

$$L_e = \sum_{i=1}^{N} (\delta^l - \delta^f)^2 \tag{9}$$

$$L_t = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}) + \lambda_t L_e \tag{10}$$

where λ_t is the trade-off of the strength of information distribution alignment, and L_t is the joint loss function.

Inference Enhancement Pre-training

Some few-shot learning methods (Liu et al. 2021a) pre-train the image encoder on the base class to enhance the representation ability of features. In few-shot VQA, it is difficult to achieve significant success just by pre-training the visual encoder. What distinguishes few-shot VQA from other few-shot learning tasks is that it requires the network to interact across modalities. We design an inference enhancement pre-training strategy as illustrated in the Figure. 4 to enhance the fine-grained interaction for generating better features.

We divide the whole CDCIN into two parts: feature extraction $\zeta(\cdot;\theta_\zeta)$ and inference network $\mu(\cdot;\theta_\mu)$. In the pre-training stage, we adopt the traditional classification setting of VQA to train the inference network. For a given training set $\Upsilon^{train} = \{(I_i, Q_i, A_i)|1 \le i \le n\}$, the target is training the parameters θ of CDCIN to predict the answers A, taking images I and questions Q as input.

$$\theta = arg\max_{\theta} \sum_{i=1}^{n} logP(y_i = A_i|f(I_i, Q_i|\theta)) \tag{11}$$

The parameters of the inference network in the meta-learning stage $\mu(\cdot;\theta_\mu)$ are shared with that in the pre-training stage, while the parameters of the feature extractor $\zeta(\cdot;\theta'_\zeta)$ are not. This avoids overfitting the feature extractor during the pre-training stage and preserves the reasoning ability of the network.

Experiments

To evaluate the effectiveness of CDCIN for few-shot VQA, we conduct a series of experiments on datasets based on widely used Toronto COCO-QA (Ren, Kiros, and Zemel 2015), Visual Genome-QA (Krishna et al. 2017) and VQA v2 (Goyal et al. 2017), including the quantitative analysis, qualitative analysis, ablation studies.

Dataset and Implementation Details

Datasets. Before training the network, we preprocess Toronto COCO-QA, Visual Genome-QA and VQA v2 for few-shot VQA. Different from these works (Dong et al. 2018; Yin et al. 2021), we fully considerate the imbalance problem of the datasets and construct three balanced few-shot datasets, named FS COCO-QA, FS VG-QA, and FS VQA. We abide by the following rules to clean the data: (1) the occurrence of each word is not less than 3; (2) the number of samples in each class is more than 30 and less than 60; (3) there is no duplication of images in all examples pairs. Finally, we randomly select 60% samples of the final set as the training set, 20% as the valid set, and the rest as the test set. The details of datasets and implementation can be seen in supplementary materials.

Comparison Experiments

To demonstrate the effectiveness of the proposed CDCIN, we compare it with several few-shot VQA methods. The results are illustrated in Table 1 and 2.

All experiments in the Table 1 were conducted on the datasets cleaned by the paper (Yin et al. 2021). We compare the performance of the proposed CDCIN with existing few-shot VQA methods, such as FPAIT (Dong et al. 2018) and HGAT (Yin et al. 2021). From the experimental results, the CDCIN outperforms all the methods in the Table 1. Compared with HGAT, the state-of-the-art algorithm, the accuracy of our method on Toronto COCO-QA improves by 15.95%, 13.77%, 21.17%, and 20.99% under the settings of 5-way-1-shot, 5-way-5-shot, 10-way-1-shot and 10-way-5-shot respectively. The performance on Visual Genome-QA improved by 6.68%, 7.44%, 11.79% and 16.66%. The excellent performance on two benchmark datasets shows that our method effectively implements multimodal information interaction and captures the commonality between modalities. The visual information entropy alignment enables the model to obtain an accurate spatial distribution of visual feature, which enhances the filtering ability of irrelevant information, thus converging the multimodal feature distribution and improving the classification performance.

We also investigate the impact of different visual backbones on the performance of CDCIN, which is illustrated in the Table 1. Three visual feature extractors are utilized in the CDCIN, i.e, ResNet12, Vit-S, and Swin-T. Among them, Swin-T and Vit-S perform better than ResNet12. Because the pixel of images inputted into Swin Transformer-Tiny and

Method	Toronto COCO-QA				Visual Genome-QA			
	5 way accuracy		10 way accuracy		5 way accuracy		10 way accuracy	
	1 shot	5shot	1 shot	5shot	1 shot	5shot	1 shot	5shot
FPAIT-CNN(Dong et al. 2018)	59.38	71.92	45.11	60.20	75.49	79.12	61.66	67.62
FPAIT+CLT(Dong et al. 2018)	60.61	72.17	46.37	60.92	75.05	79.28	60.82	67.48
Relation Net(Sung et al. 2018)	61.75	71.89	45.60	60.13	77.21	80.72	63.14	68.10
EGNN(Kim et al. 2019)	62.21	73.41	46.99	60.01	77.67	83.26	64.07	70.87
HGAT-Res12-Res12(Yin et al. 2021)	63.13	75.41	48.10	61.50	79.56	86.10	66.62	72.13
GCN-Res12⋆(Satorras and Estrach 2018)	64.20	76.85	51.36	66.12	69.14	79.49	56.91	70.12
FEAT-Res12⋆(Ye et al. 2020)	62.67	75.18	49.53	64.56	66.95	78.12	55.65	69.36
ProtoNet-Res12⋆(Chen et al. 2021)	63.08	78.40	50.74	67.63	70.16	82.82	59.00	75.61
CDCIN-Res12	72.26	84.18	59.70	75.65	83.69	91.89	74.84	85.72
CDCIN-Vit	76.52	87.44	65.49	79.46	85.83	93.10	78.18	88.25
CDCIN-Swin	**79.08**	**89.48**	**69.27**	**82.49**	**86.24**	**93.54**	**78.41**	**88.79**

Table 1: Comparison of accuracy on Toronto COCO-QA, Visual Genome-QA. The ⋆ represents that we extend these few-shot learning methods for few-shot VQA and the bolded data indicate the best results under this experimental setup.

Method	FS COCO-QA				FS VG-QA				FS VQA			
	5 way		10 way		5 way		10 way		5 way		10 way	
	1 shot	5shot	1 shot	5shot	1 shot	5shot	1 shot	5shot	1 shot	5shot	1 shot	5shot
FEAT-Res12⋆(Ye et al. 2020)	59.90	72.62	46.81	61.58	69.65	80.39	57.33	70.70	60.88	74.03	50.13	64.97
GCN-Res12⋆(Satorras and Estrach 2018)	61.84	74.72	47.69	62.26	70.71	81.63	58.47	71.24	63.20	76.25	51.81	67.05
MatchingNet-Res12⋆(Vinyals et al. 2016)	61.50	75.36	47.32	64.05	70.45	83.87	57.96	75.14	63.07	80.74	51.01	72.68
ProtoNet-Res12⋆(Chen et al. 2021)	61.64	75.76	47.52	64.83	71.19	83.97	60.06	74.73	65.80	80.51	54.59	73.41
CDCIN-Res12	66.11	79.43	52.36	68.83	80.52	89.35	70.22	82.16	78.75	88.85	69.50	82.43
CDCIN-Vit	71.42	83.77	59.48	74.21	82.84	90.67	72.88	83.47	80.01	89.41	70.58	83.04
CDCIN-Swin	**74.30**	**86.02**	**63.12**	**77.94**	**83.43**	**91.59**	**73.79**	**85.38**	**80.23**	**90.14**	**71.88**	**84.67**

Table 2: The accuracy of comparison on FS COCO-QA, FS VG-QA and FS VQA. The ⋆ represents that we extend these few-shot learning methods for few-shot VQA and the bolded data indicate the best results under this experimental setup.

Vision Transformer-Small is 224×224 much larger than that of ResNet which is resized into 84×84. Swin Transformer uses shifted window self-attention to reduce computational cost and reinforce the local receptive field, making the accuracy of CDCIN-Swin 2.88%, 2.25%, 3.64%, and 3.73% higher than that of CDCIN-Vit.

In order to eliminate the effect of data imbalance, we carefully cleaned up the data in three benchmark datasets and constructed three balanced datasets, which were introduced in the section "dataset and implementation". We reproduced some excellent few-shot learning methods and extended them to few-shot VQA. The experimental results are shown in the Table 2. We achieve two conclusions from these experimental results. (1) The performance of all the models on the cleaned datasets degrades. The accuracy of CDCIN-Swin dropped by 4.78%, 3.46%, 6.15%, and 4.55% under all of the experimental settings on COCO-QA. The reason is that we balanced the data classes and constrained the sample number of each class to not exceed 60. In this way, the phenomenon that examples of a certain category are frequently sampled will be removed, which enhances the inference capability of this kind of questions. (2) Our proposed CDCIN still outperforms all the other methods on the balanced datasets and reaches the state-of-the-art. Compared to ProtoNet, which performs the best among all reproduced methods equipped with ResNet12, the accuracy of CDCIN-Res12 increases by 4.47%, 3.67%, 4.84%, and 4.00% under all of the experimental settings on FS COCO-QA. These boosts demonstrate that the CDCIM can build deep interactions between modalities to improve the performance of inference. What's more, our best model, CDCIN-Swin, improves the accuracy by 12.66%, 10.26%, 15.6%, and 13.11% under all of the experimental settings.

Ablation Study

We conduct a series of ablation experiments to verify the effectiveness of each proposed module of the CDCIN. Table 3 presents the results of the ablation experiments on Toronto COCO-QA, which use ResNet-12 to extract visual features. The baseline is a simple network that only consists of a visual encoder and a question encoder, i.e. Case 1.

According to Table 3, all the proposed components and methods are helpful in improving the accuracy of few-shot VQA. Case 2 is equipped with CAIM, which provides multimodal fine-grained information interaction for the baseline and enables the model to learn to reason. Therefore,

Case	CAIM	VIECM	Pre	5 way		10 way	
				1 shot	5 shot	1 shot	5 shot
1				63.08	78.40	50.74	67.63
2	✔			65.94	79.69	53.32	69.04
3	✔	✔		69.92	82.45	57.70	73.42
4	✔		✔	70.32	82.35	57.72	73.05
5	✔	✔	✔	**72.26**	**84.18**	**59.70**	**75.65**

Table 3: Ablation Study of accuracy on Toronto COCO-QA. The bolded data indicate the best results under this experimental setup.

compared with Case 1, the accuracy of Case 2 in each experimental setting increased by 2.86%, 1.29%, 2.58%, and 1.41%. On the basis of Case 2, VIECM is introduced to constitute Case 3 and achieves the improvements of 3.98%, 2.76%, 4.38%, and 4.38%. This shows that VIECM maintains the consistency between the pre- and post-inference information distribution, calibrating the feature distribution and strengthening inference. We also apply the traditional VQA paradigm to train CAIM and VIECM on the base class of Toronto COCO-QA and fine-tune the pre-trained parameters, namely Case 5. This model obtains the improvements of accuracy by 7.24%, 4.13%, 6.98%, and 5.42%, which confirms that our proposed pre-training strategy can effectively preserve the reasoning ability learned during the pre-training stage.

Qualitative Analysis

In order to clearly demonstrate the whole procedure that the proposed CDCIN calibrates the feature distribution, we visualize the multimodal feature distribution before and after inference. Figure 5 shows an example of the feature distribution of pre- and post-inference, which comes from a 5-way 5-shot VQA task on the FS VQA test set. The figure (a) represent the multimodal feature distribution without passing through the inference network. When the CDCIN has not undergone deep interaction, which only learns the common representations of similar samples and fails to capture the critical information accurately, making the boundaries between feature distribution blurred. The figure (b) represent that the model has completed the multimodal information interaction and calibrated the distribution of features. The constrained features have their distributions converged toward the prototype since they are localized to important information by the model, enhancing the classification performance.

The main target of CDCIN is to achieve feature distribution calibration by aligning the information entropy between pre- and post-inference. The Figure 5 shows the visual information entropy visualization of two test set samples from the FS VQA dataset. The figure "Base" refers to the model that has not equipped with the VIECM. While "Base" has the ability to capture essential information in the image (e.g., the body of the motorcycle), it still suffers from other redundant information (e.g., trails and children). Our proposed method that calibrates the information distribution can not only retain the original key information but also search for auxiliary

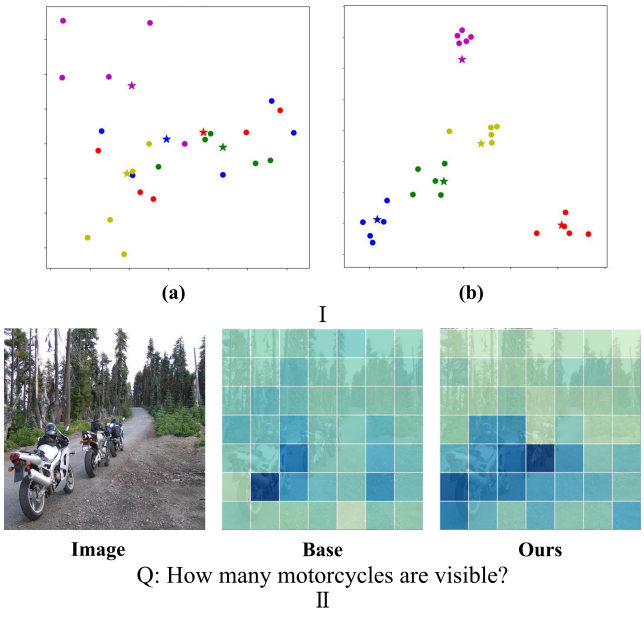

Figure 5: Figure I is the feature distribution of a few samples under the setting of 5-way 5-shot. Note that "⋆" means the prototype of each class and "o" means the feature distribution of each query sample. Figure II is the visualization of visual information entropy generated by CDCIN-Swint.

information (e.g., the wheels of the motorcycle). This allows the CDCIN to dynamically adjust the feature distribution according to the information calibration, thus improving the classification performance.

Conclusion

In this paper, we propose a cross-modal feature distribution calibration inference network for few-shot VQA, in which a novel visual information entropy is proposed to represent the spatial distribution of visual information and used to calibrate the distribution of multimodal features. Initially, visual information entropy is obtained by joint computation of original visual features and question queries, which means the visual distribution that has not yet interacted with textual features. Later, visual information entropy is generated in the multimodal feature adaptive fusion module representing the result of inference. Two kinds of visual information entropy are all sent to the information consistency loss module for distribution alignment and feature distribution calibration, thus realizing more accurate answer prediction. We conduct extensive experiments on three benchmark datasets and achieve excellent performance, surpassing the state-of-the-art methods by a large margin.

Acknowledgments

This study was funded by the Natural Science Foundation of Shanghai, China (grant number 22ZR1418400).

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Ben-Younes, H.; Cadene, R.; Thome, N.; and Cord, M. 2019. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8102–8109.

Cao, Q.; Liang, X.; Li, B.; and Lin, L. 2019. Interpretable visual question answering by reasoning on dependency trees. *IEEE transactions on pattern analysis and machine intelligence*, 43(3): 887–901.

Chen, Y.; Liu, Z.; Xu, H.; Darrell, T.; and Wang, X. 2021. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9062–9071.

Dancette, C.; Whitehead, S.; Maheshwary, R.; Vedantam, R.; Scherer, S.; Chen, X.; Cord, M.; and Rohrbach, M. 2023. Improving Selective Visual Question Answering by Learning from Your Peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24049–24059.

Ding, Y.; Yu, J.; Liu, B.; Hu, Y.; Cui, M.; and Wu, Q. 2022. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5089–5098.

Dong, X.; Zhu, L.; Zhang, D.; Yang, Y.; and Wu, F. 2018. Fast parameter adaptation for few-shot image captioning and visual question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, 54–62.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. arXiv:1606.01847.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Guo, Q.; Yao, K.; and Chu, W. 2023. Switch-BERT: Learning to Model Multimodal Interactions by Switching Attention and Input. arXiv:2306.14182.

Huang, Q.; Wei, J.; Cai, Y.; Zheng, C.; Chen, J.; Leung, H.-f.; and Li, Q. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7166–7176.

Jiang, H.; Misra, I.; Rohrbach, M.; Learned-Miller, E.; and Chen, X. 2020a. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10267–10276.

Jiang, W.; Huang, K.; Geng, J.; and Deng, X. 2020b. Multi-scale metric learning for few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3): 1091–1102.

Jing, C.; Jia, Y.; Wu, Y.; Liu, X.; and Wu, Q. 2022. Maintaining Reasoning Consistency in Compositional Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5099–5108.

Kim, J.; Kim, T.; Kim, S.; and Yoo, C. D. 2019. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11–20.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*.

Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, J.; Wang, Z.; and Hu, X. 2021. Learning intact features by erasing-inpainting for few-shot classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8401–8409.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10313–10322.

Liu, X.; Tian, X.; Lin, S.; Qu, Y.; Ma, L.; Yuan, W.; Zhang, Z.; and Xie, Y. 2021a. Learn from Concepts: Towards the Purified Memory for Few-shot Learning. In *IJCAI*, 888–894.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ma, Y.; Liu, W.; Bai, S.; Zhang, Q.; Liu, A.; Chen, W.; and Liu, X. 2020. Few-shot Visual Learning with Contextual Memory and Fine-grained Calibration. In *IJCAI*, 811–817.

Pan, H.; He, S.; Zhang, K.; Qu, B.; Chen, C.; and Shi, K. 2022. AMAM: an attention-based multimodal alignment model for medical visual question answering. *Knowledge-Based Systems*, 255: 109763.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10971–10980.

Penamakuri, A. S.; Gupta, M.; Gupta, M. D.; and Mishra, A. 2023. Answer Mining from a Pool of Images: Towards Retrieval-Based Visual Question Answering. arXiv:2306.16713.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

Satorras, V. G.; and Estrach, J. B. 2018. Few-shot learning with graph neural networks. In *International conference on learning representations*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Tran, Q.-T.; Tran, T.-P.; Dao, M.-S.; La, T.-V.; Tran, A.-D.; and Dang Nguyen, D. T. 2022. A Textual-Visual-Entailment-based Unsupervised Algorithm for Cheapfake Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7145–7149.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 3630–3638.

Xu, H.; and Saenko, K. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, 451–466. Springer.

Yang, S.; Wu, S.; Liu, T.; and Xu, M. 2021. Bridging the gap between few-shot and many-shot learning via distribution calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9830–9843.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8808–8817.

Yin, C.; Wu, K.; Che, Z.; Jiang, B.; Xu, Z.; and Tang, J. 2021. Hierarchical graph attention network for few-shot visual-semantic learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2177–2186.

Yu, Z.; Yu, J.; Fan, J.; and Tao, D. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 1821–1830.

Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; and Tao, D. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12): 5947–5959.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2022. Deepemd: Differentiable earth mover's distance for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, B.; Tian, Y.; Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2015. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*.