Federated Learning for Vision-and-Language Grounding Problems

Fenglin Liu,[1] **Xian Wu,**[3] **Shen Ge,**[3] **Wei Fan,**[3] **Yuexian Zou**[1,2*]
[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Tencent, Beijing, China
fenglinliu98@pku.edu.cn, {kevinxwu, shenge, davidwfan}@tencent.com, zouyx@pku.edu.cn

Abstract

Recently, vision-and-language grounding problems, e.g., image captioning and visual question answering (VQA), has attracted extensive interests from both academic and industrial worlds. However, given the similarity of these tasks, the efforts to obtain better results by combining the merits of their algorithms are not well studied. Inspired by the recent success of federated learning, we propose a federated learning framework to obtain various types of image representations from different tasks, which are then fused together to form fine-grained image representations. The representations merge useful features from different vision-and-language grounding problems, and are thus much more powerful than the original representations alone in individual tasks. To learn such image representations, we propose the Aligning, Integrating and Mapping Network (aimNet). The aimNet is validated on three federated learning settings, which include horizontal federated learning, vertical federated learning, and federated transfer learning. Experiments of aimNet-based federated learning framework on two representative tasks, i.e., image captioning and VQA, demonstrate the effective and universal improvements of all metrics over the baselines. In image captioning, we are able to get 14% and 13% relative gain on the task-specific metrics CIDEr and SPICE, respectively. In VQA, we could also boost the performance of strong baselines by up to 3%.

Introduction

Recently, there is a surge of research interests in vision-and-language grounding tasks such as image captioning (Chen et al. 2015) and visual question answering (VQA) (Antol et al. 2015). In image captioning, an intelligence system takes an image as input and generates a description in natural language form. VQA is a more challenging problem that takes an extra question into account and requires the model to give an answer depending on both the image and the question. The deep neural networks (Xu et al. 2015; Anderson et al. 2018; Kim, Jun, and Zhang 2018) have achieved great success in advancing the state-of-the-arts of image captioning and VQA.

*Corresponding Author.

Despite the impressive results, most of the existing deep learning based frameworks focus on individual tasks. If these problems are considered together, different knowledge from different tasks could be learned jointly, and there are high chances to promote the performance of each task. To achieve this goal, a multi-task learning framework has been proposed for vision-and-language grounding tasks (Li et al. 2018; 2017). However, some approaches are trained under the condition of sharing all downstream task data, which may cause data leakage. For example, Li et al. (2018) requires a question-answer pair and an identical image as input to train the visual question answering and visual question generation tasks together. This approach won't work when cross-task datasets are different.

In recent years, federated learning (McMahan et al. 2017; Konecný et al. 2016a; 2016b; Yang et al. 2019) has been proposed as an alternative machine learning setting. The goal is to train a high quality centralized model based on datasets that are distributed across multiple clients without sharing the clients' data. For the vision-and-language grounding tasks, inspired by the success of federated learning, we can treat each of them as an individual client, enabling the design of a federated learning framework with a centralized model. Such design establishes a bond among different tasks to learn various types of knowledge, with the advantage to improve the performance of each downstream task while preventing data leakage.

In this paper, we mainly study on two representative vision-and-language grounding tasks, i.e., image captioning and VQA. We bridge the gap between these two tasks with a federated learning framework (McMahan et al. 2017), which allows the sharing of the obtained fine-grained image representations instead of direct task data. To this end, we design an aimNet as the centralized model in the federated learning framework, which consists of an aligning module, an integrating module and a mapping module, the sketch of which are shown in Figure 1. The aligning module builds aligned image representations by conducting mutual attention (Liu et al. 2019b) over the extracted visual and textual features. The resulting image representations form a clearer semantic description of salient image regions, benefiting the downstream tasks (Su et al. 2019; Liu et al.

2019b) by injecting more semantic information. Next, the integrating module focuses on integrating visual and textual features via a self-attention mechanism (Vaswani et al. 2017), which captures the groupings of salient regions and the collocations of attributes, generating aspect-describing image representations (Liu et al. 2019c). The explored spatial and relational representations of the image serve as a powerful basis for image captioning task (Yao et al. 2018; Liu et al. 2019c). Finally, the mapping module consists of a two-layer non-linear layer, which are used to map the learned fine-grained image representations to the feature domain of specific tasks. Our modules fully exploited all effective information in the image, and pass it to the decoder as input to generate meaningful sentences or give an accurate answer to the question. Experiments on two image captioning datasets and a VQA dataset validate the motivations and corroborate the effectiveness of our approach.

Overall, our main contributions are as follow:

- We propose a federated learning framework. By generating fine-grained image representations, our framework improves the performance on a variety of vision-and-language grounding problems, without the sharing of downstream task data.
- We implement the centralized model in our framework as the designed Aligning, Integrating and Mapping Network (aimNet), which converts the extracted visual and textual features from image to fine-grained image representations, effectively and automatically.
- We validate our approach on three federated learning settings. Extensive experiments on the MSCOCO image captioning dataset, Flickr30k image captioning dataset and VQA v2.0 dataset demonstrate the effectiveness and the universality of our approach.

Related Work

The related work are introduced from four aspects: 1) Vision-and-Language Grounding Problems; 2) Federated Learning; 3) Multi-Task Learning and 4) Image Representations.

Vision-and-Language Grounding Problems Vision-and-language grounding problems, which including image captioning (Chen et al. 2015), visual question answering (Antol et al. 2015) and image caption retrieval (Nam, Ha, and Kim 2017), and others, have drawn remarkable attention in both natural language processing and computer vision. These tasks combine image and language understanding together at the same time, are tough yet practical. However, these studies deal with one single task at one time. In this paper, we propose a bonding framework of different tasks to further improve the performance of each task. At the same time, our approach avoids data leakage.

Federated Learning Recently, McMahan et al. (2017), Konecný et al. (2016a) and Konecný et al. (2016b) propose the concept of federated learning, which can be divided into three categories, i.e., horizontal federated learning, vertical federated learning and federated transfer learning, based on the distribution characteristics of the data. Due to space limit, please refer Yang et al. (2019) for detailed explanations. Federated learning poses new statistical and systems challenges in training machine learning models over distributed networks of devices, and is a key learning scenario in large-scale applications. In that scenario, a centralized model is trained based on data originated from a large number of clients, which may be phones, other mobile devices, or sensors. In this paper, in order to transfer the success of federated learning, we treat each vision-and-language grounding task as a client and implement the centralized model with the proposed aimNet.

Multi-Task Learning It is worth noticing that the goal of the proposed framework is similar to multi-task learning (Caruana 1997). Li et al.; Li et al. (2017; 2018) have achieved early successes in vision-and-language grounding tasks. However, in their approach, an input is shared by all the tasks, thus they must create a dataset specially designed for multi-task learning, where multiple objectives are given to the identical inputs. Recently, Nguyen and Okatani (2019) tried to relax the reliance on the dedicated dataset for multiple tasks. However, the original text information about the image is still needed as one input (e.g., the captions in image caption retrieval and the questions in VQA). As a result, the model won't work when applied to tasks where the inputs are only images, e.g., image captioning and visual storytelling.

Image Representations For a variety of vision-and-language grounding problems, an important goal is to understand the image despite their different application scenarios, which justifies the acquisition of fine-grained image representations. In the literature, to represent images, visual features extracted by CNNs or Region-CNNs are most-widely used (Xu et al. 2015; Anderson et al. 2018), while textual features consisting of semantic concept vectors are also proposed (Fang et al. 2015). However, relationships among the individual parts of representations are not explicitly defined, which in fact should be essential to a deep understanding of images. Recently, some works (Yao et al. 2018; Liu et al. 2019b; 2019c) explored the visual relationships among the individual parts of representations, which provides a solid basis for downstream vision-and-language grounding tasks. Specifically, Yao et al. (2018) and Liu et al.; Liu et al. (2019b; 2019c) attempted to use graph networks and attention mechanism to explore visual relationships, respectively. The graph-based approaches explicitly model the spatial and semantic relationships of image information, while the attention-based methods accomplish that in implicit ways.

Approach

This section includes three parts: 1) The visual and textual features extractor; 2) The designed centralized model - aimNet and 3) The implementations in three federated learning settings. We first introduce the visual and textual feature extractors. We then discuss the aligning, integrating and mapping network (aimNet) in detail (see Figure 1). Finally, we describe the three federated learning settings (see Figure 2, 3 and 4).

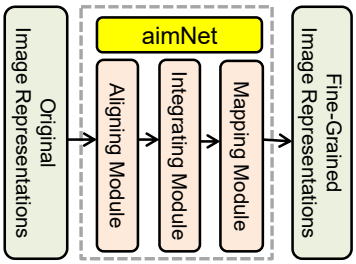

Figure 1: The overall framework of the proposed centralized model aimNet, which consists of an aligning module, an integrating module and a mapping module. The proposed aimNet could acquire fine-grained image representations in an effective and automatic manner for improved downstream tasks.

Visual and Textual Features

For vision-and-language grounding tasks, visual features extracted by deep CNNs are most-widely used. In our experiments, we utilize the RCNN-based features extracted from the Faster R-CNN (Ren et al. 2015). We denote the extracted visual features as $\vec{I} = \{\vec{i}_1, \vec{i}_2, \ldots, \vec{i}_N\} \in \mathbf{R}^{N \times d}$. Considering the limited expression capacity of the visual features (Wu et al. 2016), textual features have been used to provide explicit high-level information of an image (Fang et al. 2015; Wu et al. 2016). In implementation, the textual features are generated by the predicted semantic concepts in the image, and we adopt a weakly-supervised approach of Multiple Instance Learning (Zhang, Platt, and Viola 2006) to build the semantic concepts extractor, following Fang et al. (2015). We denote the textual features as $\vec{T} = \{\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_M\} \in \mathbf{R}^{M \times d}$, which are the word embeddings for a list of semantic concepts. The concepts could be objects (e.g. *dog*, *frisbee*), attributes (e.g. *off*, *electric*), or relationships (e.g. *holding*, *flying*). The textual features represents the image from a semantic perspective, providing a powerful bias for vision-and-language tasks.

Aligning, Integrating and Mapping Network

In this section, we first introduce the basic module of the proposed approach. Then we will introduce the proposed aligning module, integrating module, and mapping module in detail.

Basic Module In order to extract the relationship between the intra-modality and inter-modality of visual features and textual features, we adapt Multi-Head Attention (MHA) and Feed-Forward Network (FFN)[1] (Vaswani et al. 2017), which compute the association weights between different features. The attention mechanism allows probabilistic many-to-many relations instead of monotonic relations, as in Xu et al. (2015). We take advantage of the multi-head attention to implement the idea of aligning visual and textual features for semantic-based representations, as well as learning the groupings of salient region and the collocations of attributes for aspect-based representations.

[1]Please refer to Vaswani et al. (2017) for the detailed introduction of Multi-Head Attention and Feed-Forward Network.

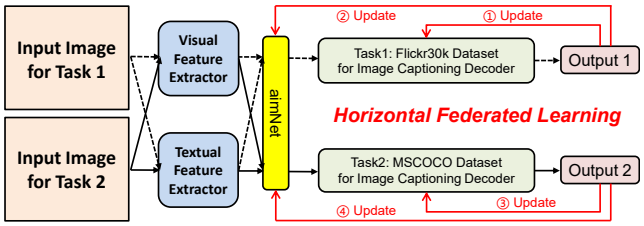

Figure 2: Under the setting of horizontal federated learning, we apply our framework on Flickr30k and MSCOCO image captioning datasets, where the two tasks share the same task objective ("generate captions") with different input images. For federated learning based training, we first run an instance on the task 1, and update the task-specific decoder of task 1 (update 1) and then the aimNet (update 2) based on the training loss. After that, we run an instance of task 2, and similarly update the task-specific decoder (update 3) and aimNet (update 4). In all, the aimNet is able to obtain various types of image representations from different tasks, and the learned fine-grained image representations are much more powerful than the original representations alone in separate tasks. Besides, it is worth noticing that the source information between different tasks is not shared. The running instance of each task runs independently during the training and inference stage.

Aligning Module To represent visual features in a more meaningful way, we need to find the most relevant semantic concepts from the textual features to summarize the properties of the visual features. Similarly, we need to provide visual references for textual features to reduce semantic ambiguity (e.g., the word *mouse* can either refer to a mammal or an electronic device), via providing an image area that is consistent with current semantic concept (Liu et al. 2019b).

According to the attention theorem, we can adapt the following formula to simulate the above process:

$$\vec{I}_a = \text{FFN}(\text{MHA}(\vec{I}, \vec{T}, \vec{T})) \quad (1)$$

$$\vec{T}_a = \text{FFN}(\text{MHA}(\vec{T}, \vec{I}, \vec{I})) \quad (2)$$

Through the alignment of visual and textual features, we can get the semantic-based image representations (Liu et al. 2019b), providing a powerful basis for vision-and-language grounding tasks, especially for answering the question about images in VQA (Su et al. 2019).

Integrating Module When describe an image, we often focus on one specific region and seek for other regions that often appears in the neighbourhood of that region. Following this, the integrating module is supposed to learn those spatially or semantically related objects from an inherent group that we attend to. In this case, each specific region vector in visual features $\vec{I}$ and each specific concept vector in textual features $\vec{T}$ should pay attention to the entire visual features $\vec{I}$ and textual features $\vec{T}$, respectively, to find the most related image regions and attribute collocations and generate

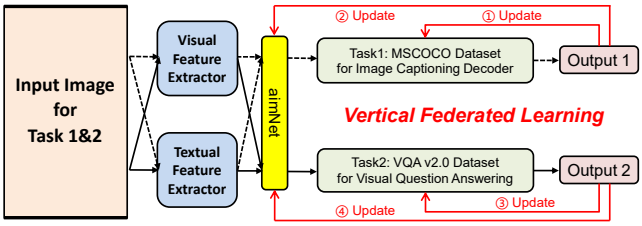

Figure 3: We implement vertical federated learning setting on MSCOCO image captioning dataset and VQA v2.0 dataset. Although their downstream tasks are different, they share most of the input images.

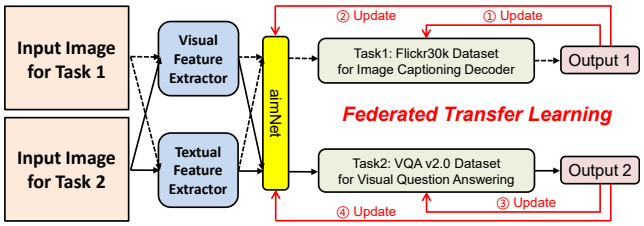

Figure 4: In federated transfer learning, we conduct the implementation on the Flickr30k image captioning dataset and VQA v2.0 dataset. They have not only different input images but also different downstream tasks.

aspect-based image representations $\vec{I}_i$ and $\vec{T}_i$. This process can be defined as follow:

$$\vec{I}_i = \text{FFN}(\text{MHA}(\vec{I}_a, \vec{I}_a, \vec{I}_a)) \quad (3)$$

$$\vec{T}_i = \text{FFN}(\text{MHA}(\vec{T}_a, \vec{T}_a, \vec{T}_a)) \quad (4)$$

Through the formula, in the visual domain, the integrating module learns salient region groupings and integrates naturally related image regions for a higher-level representation of the image. In the textual domain, it learns attribute collocations and have the ability of considering associations and collocations during sentence phrasing. The acquired aspect-based image representations are super beneficial for image captioning task (Yao et al. 2018; Liu et al. 2019c).

Mapping Module Different tasks have different data spaces, so we need to map the fine-grained image representations into the task space, and allow the proposed framework to adapt to different tasks. In order to do so, we introduce the mapping module, which is defined as:

$$\text{Mapping}(x) = \tanh(x\text{W}_\text{m} + b_\text{m})\text{W}_\text{mm} + b_\text{mm} \quad (5)$$

where W_m and W_mm denote matrices for linear transformation; and b_m and b_mm represent the bias terms. For each downstream task applied in the framework, we apply a mapping module to map the fine-grained image representations learned from aimNet to the task space, i.e., $\text{LayerNorm}(\text{Mapping}(\vec{I}_i)+\text{Mapping}(\vec{T}_i))$, where the LayerNorm stands for layer normalization (Ba, Kiros, and Hinton 2016). In this way, we inject rich information into the task space, so the mapped fine-grained image representations are expected to be a better start for downstream tasks.

Implementation

In this section, we briefly introduce the three federated learning settings in our implementation.

Horizontal Federated Learning Horizontal federated learning is known as sample-based federated learning, which applies in the cases where datasets share the same feature space while holding different samples. For example, two banks in two different cities may have different users, but their feature spaces may be the same because they share the same business.

In our applied scenario, we treat two different image captioning datasets as two banks coming with different users, because they have the same "business" (generate captions) but different "users" (input images). We implement this scenario on MSCOCO and Flickr30k image captioning datasets, as shown in Figure 2.

Vertical Federated Learning In contrast to horizontal federated learning, vertical federated learning is a feature-based learning, which is applicable in the scenarios where two datasets share the same users but differ in feature space. For example, consider two different companies in the same city, one is a bank, and the other is an insurance company. Their user sets are likely to contain most of the residents in that area, so the intersection of their user space may be large. However, due to the largely different business of the two companies, their feature spaces may be quite different.

Similarly, in applying this case, we treat two different downstream tasks as two different companies, enjoying the same "users" (input images). As shown in Figure 3, we choose the MSCOCO image captioning dataset and VQA v2.0 dataset to implement the scenario. The reason is that most of the input images in VQA v2.0 dataset are collected from the MSCOCO dataset.

Federated Transfer Learning Federated transfer learning applies to the scenarios that the two datasets differ not only in samples but also in feature space. Consider the following situation, a bank is located in United States, and an insurance company is located in Europe. Due to geographical restrictions and business differences, the intersection between the user groups and the feature spaces of the two companies will be rather limited.

In order to simulate the above scenario in our implementation, we treat the image captioning task on Flickr30k dataset and VQA task on VQA v2.0 dataset as the aforementioned bank and insurance company, respectively.

Experiments

In this section, we first describe three benchmark datasets and experimental settings for image captioning and VQA tasks, and some widely-used evaluation metrics. Next, we present our evaluation of the proposed framework on three federated learning settings, i.e., horizontal federated learning, vertical federated learning and federated transfer learning.

Table 1: Evaluation of the proposed framework on the Flickr30k and MSCOCO image captioning datasets under the horizontal federated learning setting. B-4, M, C and S are short for BLEU-4, METEOR, CIDEr and SPICE, respectively. All values are reported in percentage (%). As we can see, the horizontal federated learning (HFL) promotes the baselines in all metrics, proving the effectiveness to learn various of knowledge from different tasks in our proposed federated framework.

Training Datasets	Flickr30k	B-4	M	C	S	Training Datasets	MSCOCO	B-4	M	C	S
Spatial (Lu et al. 2017)											
Flickr30k	Baseline	26.7	21.0	57.1	14.6	MSCOCO	Baseline	33.5	26.9	109.8	20.0
Flickr30k+MSCOCO	HFL	**27.8**	**21.9**	**63.3**	**16.5**	Flickr30k+MSCOCO	HFL	**35.1**	**27.6**	**114.9**	**20.5**
NBT (Lu et al. 2017)											
Flickr30k	Baseline	27.8	21.7	60.2	15.6	MSCOCO	Baseline	34.9	27.4	110.7	19.9
Flickr30k+MSCOCO	HFL	**29.6**	**22.3**	**68.4**	**16.6**	Flickr30k+MSCOCO	HFL	**35.9**	**27.7**	**115.2**	**20.6**

Datasets, Metrics and Baselines

We evaluate our framework on image captioning and VQA.

In image captioning, our reported results are evaluated on the popular MSCOCO image captioning dataset (Chen et al. 2015) and the Flickr30k image captioning dataset (Young et al. 2014). The datasets contain 123,287 images and 31,783 images, respectively, with 5 sentences paired to each image. To make fair comparisons, we use the widely-used splits (Karpathy and Li 2015) to report our results. There are 5,000 images each in the validation set and the test set for MSCOCO, and 1,000 images as for Flickr30k. Following common practice (Lu et al. 2017; Liu et al. 2018), we report results with the help of the MSCOCO captioning evaluation toolkit (Chen et al. 2015), which includes the evaluation metrics SPICE (Anderson et al. 2016), CIDEr (Vedantam, Zitnick, and Parikh 2015), METEOR (Banerjee and Lavie 2005) and BLEU (Papineni et al. 2002). Among them, SPICE and CIDEr are customized metrics for evaluating image captioning systems, based on scene-graph matching and n-gram matching, respectively. We conduct the experiments on two strong baselines, i.e., Spatial (Lu et al. 2017) and NBT (Lu et al. 2018), with cross-entropy optimization.

In VQA, we evaluate the framework on VQA v2.0 dataset, where the images are collected from the MSCOCO dataset (Lin et al. 2014). VQA 2.0 is split into train, validation and test-standard sets. There are 82,783, 40,504 and 81,434 images, (443,757, 214,354 and 447,793 corresponding questions) in the training, validation and test set, respectively. The questions are categorized into three types, namely Yes/No, Number and other categories. Each question is accompanied with 10 answers composed by the annotators. Answers with the highest frequency are treated as the ground-truth. We choose BUTD (Anderson et al. 2018) and BAN (Kim, Jun, and Zhang 2018) for comparison. The former is the winner of VQA challenge 2017 and the latter is the state-of-the-art on VQA v2.0. Following common practice, these VQA models are trained on the training and validation splits plus extra Visual Genome dataset (Krishna et al. 2017) (We ensure that any images found in test set of either datasets of image captioning and VQA v2.0 are avoided be contained in the training split in both datasets.). The reported accuracies are calculated by the standard VQA metric (Antol et al. 2015).

Settings

For fair comparisons, we use the RCNN-based image features provided by Anderson et al. (2018), which uses Faster R-CNN to detect objects. For textual concepts, we use the textual concepts prediction model pre-trained by Fang et al. (2015) for 1,000 words. The caption/question words and the textual concept words share the same embeddings. For our proposal, d stands for the hidden/model size of the baseline decoder. The number of both extracted visual and textual features are 36, which means $N = M = 36$. Following Vaswani et al. (2017), we set the number of attention heads to 8 and the feed-forward network dimension to 2048.

For equipping with our aimNet in baseline models, i.e., using the fine-grained image representations learned by aimNet in baseline models, we replace the original features with the refined features directly since our features are considered to be more powerful. Also our aimNet does not make any changes in the number or the size of original feature vectors (each of them can be seen as a weighted average of the original features). We preserve the original settings for all baselines, and our framework is end-to-end trainable.

Experimental Results

In this section, we briefly introduce the three federated learning settings in our experiments, followed by the discussion of the experimental results.

Horizontal Federated Learning As mentioned above, we experiment with horizontal federated learning setting on MSCOCO and Flickr30k image captioning datasets, the results are shown in Table 1. As we can see, all baselines enjoy comfortable improvements on all metrics. Especially, applied with with the proposed approach, all the models enjoy a relative increase of 11%$\sim$14% in performance of CIDEr score on Flickr30k dataset. It is worth noticing that the federated learning with Flickr30k and MSCOCO is more beneficial for the smaller dataset (Flickr30k) than the larger one (MSCOCO), e.g., $60.2 \rightarrow 68.4$ in CIDEr vs. $110.7 \rightarrow 115.2$ in CIDEr (NBT), which indicates that the model with small dataset can learn more useful knowledge from that with big dataset.

Vertical Federated Learning In vertical federated learning setting, we experiment on two different downstream

Table 2: Performance on MSCOCO dataset and VQA v2.0 dataset under vertical federated learning setting.

Datasets	Methods	C	S	Datasets	Methods	test-std
Spatial				*BUTD*		
MSCOCO	Baseline	109.8	20.0	VQA	Baseline	67.5
+ VQA	+ BUTD	115.4	20.7	+ MSCOCO	+ Spatial	69.1
+ VQA	+ BAN	**116.1**	**20.8**	+ MSCOCO	+ NBT	**69.3**
NBT				*BAN*		
MSCOCO	Baseline	110.7	19.9	VQA	Baseline	69.8
+ VQA	+ BUTD	116.3	21.0	+ MSCOCO	+ Spatial	70.4
+ VQA	+ BAN	**117.5**	**21.2**	+ MSCOCO	+ NBT	**70.6**

Table 3: Results of the Flickr30k dataset and VQA v2.0 dataset under the federated transfer learning setting.

Datasets	Methods	C	S	Datasets	Methods	test-std
Spatial				*BUTD*		
Flickr30k	Baseline	57.1	14.6	VQA	Baseline	67.5
+ VQA	+ BUTD	**61.2**	15.3	+ Flickr30k	+ Spatial	68.7
+ VQA	+ BAN	60.7	**15.4**	+ Flickr30k	+ NBT	**68.8**
NBT				*BAN*		
Flickr30k	Baseline	60.2	15.6	VQA	Baseline	69.8
+ VQA	+ BUTD	64.2	15.8	+ Flickr30k	+ Spatial	70.1
+ VQA	+ BAN	**64.8**	**16.1**	+ Flickr30k	+ NBT	**70.2**

tasks with the most of the same input images. As shown in Table 2, our approach successfully boosts all baselines, with the most significant improvement up to relatively 6% and 3% in terms of SPICE for image captioning and accuracies for VQA, verifying the effectiveness of our approach. Specifically, we achieve the best performance on the MSCOCO and VQA v2.0 datasets in all of our experiments. Vertical federated learning allows the sharing of most input images, which directly helps the baseline models to learn a broader knowledge of the identical images.

Federated Transfer Learning As shown in Table 3, our approach can still bring improvements to the strong baselines under the federated transfer learning settings, proving the effectiveness and the generalization ability of the proposed framework. Besides, Table 3 shows similar phenomenon as in Table 1, i.e., federated learning framework is more beneficial for the smaller dataset rather than the larger dataset. Nonetheless, both of them can get performance improvements from our approach.

We further compare our framework with the multi-task framework recently proposed in Nguyen and Okatani (2019) on VQA v2.0 dataset. Specifically, Nguyen and Okatani (2019) adopt the DCN (Nguyen and Okatani 2018) as their VQA decoder, and their performance of DCN is promoted from 68.9% to 69.6% overall accuracy (+0.7% overall accuracy gain). In our work, our best results (see Table 2) show that our framework can promote the performance of BAN, which is an even stronger baseline than DCN, from 69.8% to 70.6% overall accuracy (+0.8% overall accuracy gain). Moreover, we attempt to experiment with the DCN (joint training with NBT) on vertical federated learning setting in our framework for fair comparisons. The results show that our framework performs better than Nguyen and Okatani (2019) (70.1% overall accuracy vs. 69.6% overall accuracy), which strongly demonstrated the performance promoting capability of our proposed framework.

In all, our framework successfully promotes all baseline models in all metrics across the board, regardless of their downstream tasks. In image captioning, it has brought improvements up to 14% and 13% in terms of CIDEr and SPICE, respectively. In VQA, an overall improvement up to 3% is achieved when applying our framework to the baselines. These results validate that our framework generalizes well to different tasks and indicates its effectiveness in learning fine-grained image representations for vision-and-language grounding tasks.

Analysis

In this section, we first give some intuitive examples to demonstrate the strength of our approach. Next, we analyze the contribution of each component in the proposed method. The following analyses are conducted on two baselines, i.e., Spatial on MSCOCO image captioning dataset and BUTD on VQA v2.0 dataset.

Qualitative Analysis

In Figure 5, we list some intuitive examples to show the differences between the models. We can see that the aligning module learns to extend its focus of a specific object and look for related attributes, which assists the image captioning baseline models to generate captions that are more detailed in attributes and colors. The aligning module also helps the VQA baseline models to answer the questions more accurately, such as "what" and "where" object that belong to the *"Other"* category. The integrating module performs well in integrating related objects in the image, which results in more comprehensiveness in objects for image captioning, as well as more accurate answer generation, especially in answering the *"Number"* category. The Full Model (w/ Aligning + Integrating) helps the baselines to maintain good balances. By including more objects as well as many informative and detailed attributes, such as the quantity and the color, the image captioning models could generate the captions with best quality.

Quantitative Analysis

In Table 4, we conduct the ablation analysis to investigate the contribution of each component in the proposed aimNet. From the table, we can see that the aligning module could achieve greater improvements on the VQA than the integrating module. This could be illustrated in Figure 6, because the semantic-based image representations learned by the aligning module is more informative than the aspect-based features learned by integrating module. However, in image captioning tasks, the integrating module could bring more increase in scores than the aligning module, which

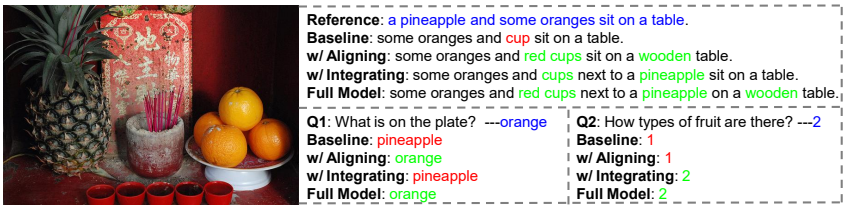

Figure 5: Examples of the generated captions / predicted answers by different methods. The color *Blue* denotes the ground truth, the color *Green* denotes the examples when image captioning model generates better captions than the baseline or when the VQA model gives the correct answer, while *Red* denotes unfavorable results.

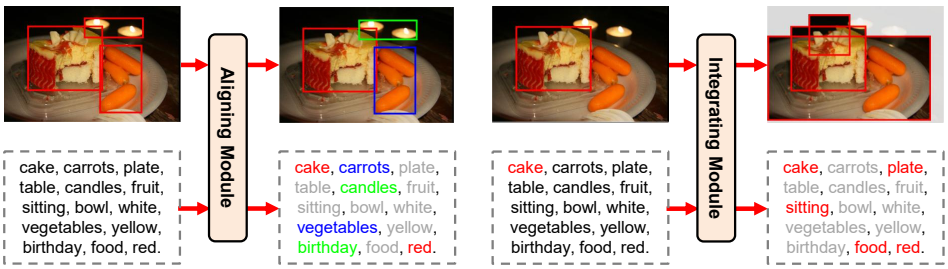

Figure 6: Illustration of the semantic-based image representations (left) and aspect-based image representations (right) learned by the proposed aligning module and integrating module, respectively. For the left plot, we show the process of aligning three typical visual features with textual features. For the right plot, the example comes from the input is *cake*.

Table 4: Ablation analysis of aimNet. We perform the analysis for the Spatial and BUTD model on the MSCOCO and VQA v2.0. VFL stands for vertical federated learning. Bold numbers are best **before** applying full model and VFL.

Methods	MSCOCO		VQA
	C	S	test-dev
Baseline	109.8	20.0	67.3
w/ Aligning	112.7	20.2	**68.1**
w/ Integrating	**113.1**	**20.4**	67.8
w/ Aligning + Integrating (Full Model)	114.2	20.5	68.3
Full Model w/ VFL	116.1	20.8	69.0

is due to the fact that the integrating modules are better at exploring the visual and textual relationships, making the image captioning decoder to generate more comprehensive and accurate captions. In our aimNet, all components bring about fine-grained image representations, yet from different perspectives to the model. As a result, their advantages are unified to produce abundant and enriched image information. By doing this, we could achieve a deep image understanding, producing an overall improvement regardless of the downstream vision-and-language grounding tasks. The introduction of the federated learning framework further improves the performance.

Visualization

Figure 6 shows an example of semantic-based and aspect-based image representations learned by the aligning module and integrating module according to the attention weights in the Multi-Head Attention, respectively. Please view in color. As we can see, the aligning module provides a clearer semantic information of image by aligning the visual and textual features. The integrating module generates intrinsic combinations among the individual parts of representations, which models the spatial and semantic relationships of image regions. The semantic-based and aspect-based image representations are both beneficial for deep and semantic understanding of images, which provides a solid bias for downstream vision-and-language grounding tasks.

Conclusion

We propose a federated learning framework and an Aligning, Integrating and Mapping Network (aimNet), which extract the fine-grained image representations by bonding different downstream vision-and-language tasks while avoid the data sharing of the downstream tasks. The proposed federated framework with aimNet is validated by experiments on three federated learning settings. Extensive experiments on two representative tasks show that our approach successfully boosts all baselines in all metrics, demonstrating the effectiveness and universality of our approach.

Acknowledgments

This paper was partially supported by National Engineering Laboratory for Video Technology - Shenzhen Division, Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing). Special acknowledgements are given to Aoto-PKUSZ Joint Lab for its support. We thank all the anonymous reviewers for their constructive comments and suggestions.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic propositional image caption evaluation. In *ECCV*.

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 VQA. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: visual question answering. In *ICCV*.

Ba, L. J.; Kiros, R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Banerjee, S., and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*.

Caruana, R. 1997. Multitask learning. *Machine Learning*.

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

Fang, H.; Gupta, S.; Iandola, F. N.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; Zitnick, C. L.; and Zweig, G. 2015. From captions to visual concepts and back. In *CVPR*.

Karpathy, A., and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kim, J.; Jun, J.; and Zhang, B. 2018. Bilinear attention networks. In *NeurIPS*.

Konecný, J.; McMahan, H. B.; Ramage, D.; and Richtárik, P. 2016a. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

Konecný, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016b. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Li, F. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In *ICCV*.

Li, Y.; Duan, N.; Zhou, B.; Chu, X.; Ouyang, W.; Wang, X.; and Zhou, M. 2018. Visual question generation as dual task of visual question answering. In *CVPR*.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: common objects in context. In *ECCV*.

Liu, F.; Ren, X.; Liu, Y.; Wang, H.; and Sun, X. 2018. simNet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In *EMNLP*.

Liu, F.; Gao, M.; Zhang, T.; and Zou, Y. 2019a. Exploring semantic relationships for image captioning without parallel data. In *ICDM*.

Liu, F.; Liu, Y.; Ren, X.; He, X.; and Sun, X. 2019b. Aligning visual regions and textual concepts for semantic-grounded image representations. In *NeurIPS*.

Liu, F.; Ren, X.; Liu, Y.; Lei, K.; and Sun, X. 2019c. Exploring and distilling cross-modal information for image captioning. In *IJCAI*.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *CVPR*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*.

Nam, H.; Ha, J.; and Kim, J. 2017. Dual attention networks for multimodal reasoning and matching. In *CVPR*.

Nguyen, D., and Okatani, T. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*.

Nguyen, D., and Okatani, T. 2019. Multi-task learning of hierarchical vision-language representation. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. BLEU: a Method for automatic evaluation of machine translation. In *ACL*.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. VL-BERT: pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Wu, Q.; Shen, C.; Liu, L.; Dick, A. R.; and van den Hengel, A. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated machine learning: Concept and applications. *ACM TIST*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *ECCV*.

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. *TACL*.

Zhang, C.; Platt, J. C.; and Viola, P. A. 2006. Multiple instance boosting for object detection. In *NIPS*.