Federated Learning for Vision-and-Language Grounding Problems

Fenglin Liu[1], Xian Wu[3], Shen Ge[3], Wei Fan[3] and Yuexian Zou[1,2]

[1] Peking University, China

[2] Peng Cheng Laboratory, China

[3] Tencent, China

CONTENTS

1 Introduction

Vision-and-Language Grounding Problems

Vision-and-Language Grounding Problems, such as image captioning and visual question answering (VQA), 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.

Image Captioning

- ✓ A group of people of Asian descent watch a street performer in a wooded park area.
- ✓ A large crowd of people surround a colorfully dressed street entertainer.
- ✓ A crowd of people watching a balloon twister on a beautiful day.
- ✓ A crowd of people are gathered outside watching a performer.
- ✓ A crowd is gathered around a man watching a performance.

Visual Question Answering

Who is wearing glasses?

man

woman

Image Captioning and Visual Question Answering

Image Captioning

- ✓ A group of people of Asian descent watch a street performer in a wooded park area.
- ✓ A large crowd of people surround a colorfully dressed street entertainer.
- ✓ A crowd of people watching a balloon twister on a beautiful day.
- ✓ A crowd of people are gathered outside watching a performer.
- ✓ A crowd is gathered around a man watching a performance.

In image captioning, an intelligence system takes an **image** as input and generates a **description** in natural language form.

Visual Question Answering

VQA is a more challenging problem takes an extra **question** into account and requires the model to give an **answer** depending on both the **image** and the **question**.

Motivations

- 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. However, some approaches are trained under the condition of sharing all downstream task data, which may cause data leakage.

- In recent years, federated learning 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

Solutions

- We can treat each of vision-and-language grounding tasks 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.

- We propose a bonding framework (**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 fine-grained 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. At the same time, our approach avoids data leakage.

- In implementation, we design an **A**ligning, **I**ntegrating and **M**apping **Net**work (**aimNet**) as the centralized model to better learn fine-grained image representations in the federated learning framework.

Contributions

- 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.

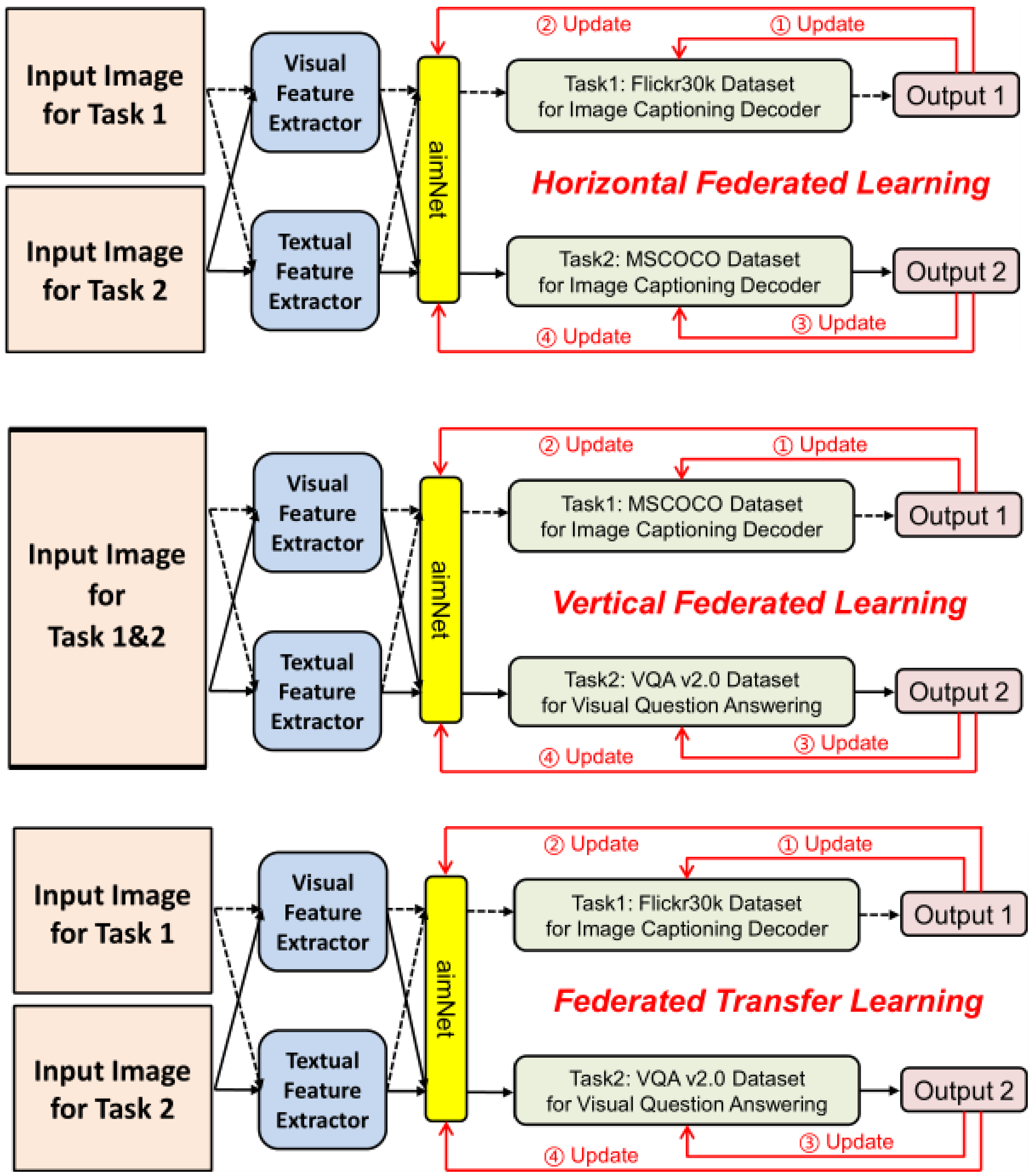
- We implement the centralized model in our framework as the designed **A**ligning, **I**ntegrating and **M**apping **Net**work (**aimNet**), which converts the extracted visual and textual features from image to fine-grained image representations, effectively and automatically.

- We validate our approach on three federated learning settings. The proposed approach outperforms previous on the MSCOCO image captioning dataset, Flickr30k image captioning dataset and VQA v2.0 dataset

2 Approach

北京大學
PEKING UNIVERSITY

Overview



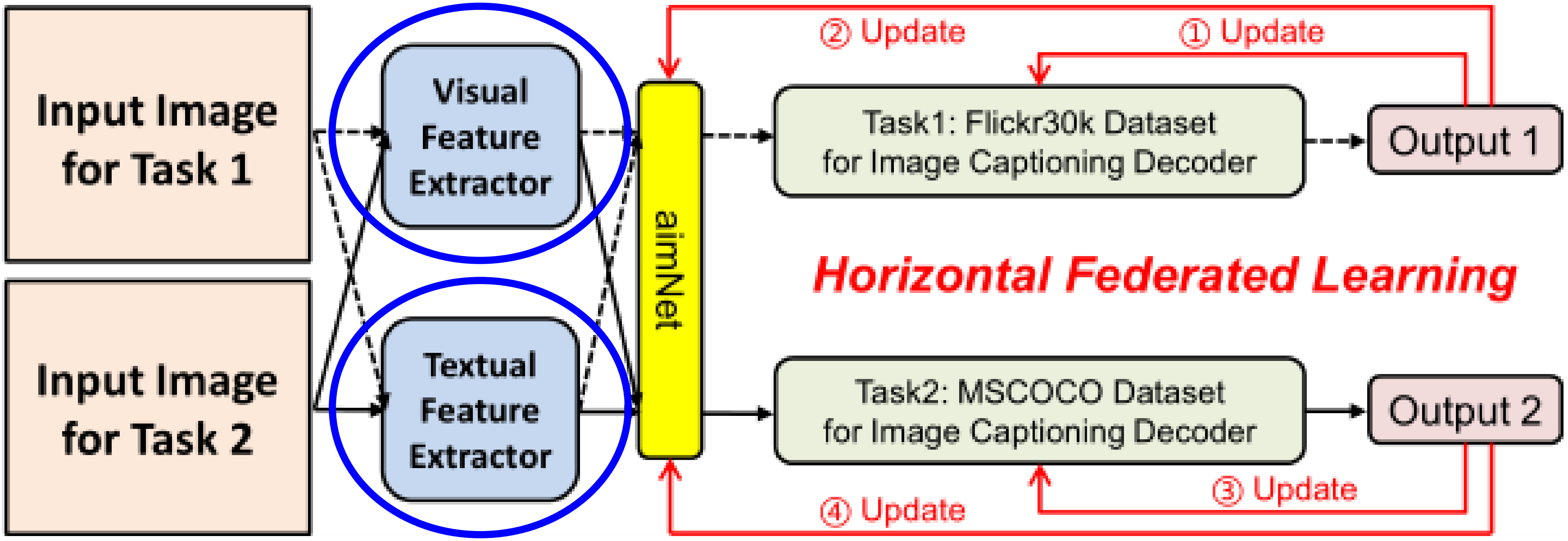
北京大學 PEKING UNIVERSITY

Visual and Textual Features

$\vec{I}$ Visual Feature Extractor:

Faster-RCNN

Anderson et al., 2018: Bottom-up and top-down attention for image captioning and VQA . In CVPR 2018.



$\vec{T}$ Attribute Word Extractor:

Multiple Instance Learning

Zhang et al., 2006: Multiple instance boosting for object detection. In NIPS 2006.

Fang et al., 2015: From captions to visual concepts and back. In CVPR2015

Background: Multi-Head Attention

Scaled Dot-Product Attention (Att):

$$\mathrm{Att}(\mathrm{Q},\mathrm{K},\mathrm{V})_i = \mathrm{softmax}\left(\frac{\mathrm{QW}_i^{\mathrm{Q}}(\mathrm{KW}_i^{\mathrm{K}})^T)}{\sqrt{d_k}}\right)\mathrm{VW}_i^{\mathrm{V}}$$

Multi-Head Attention (MHA):

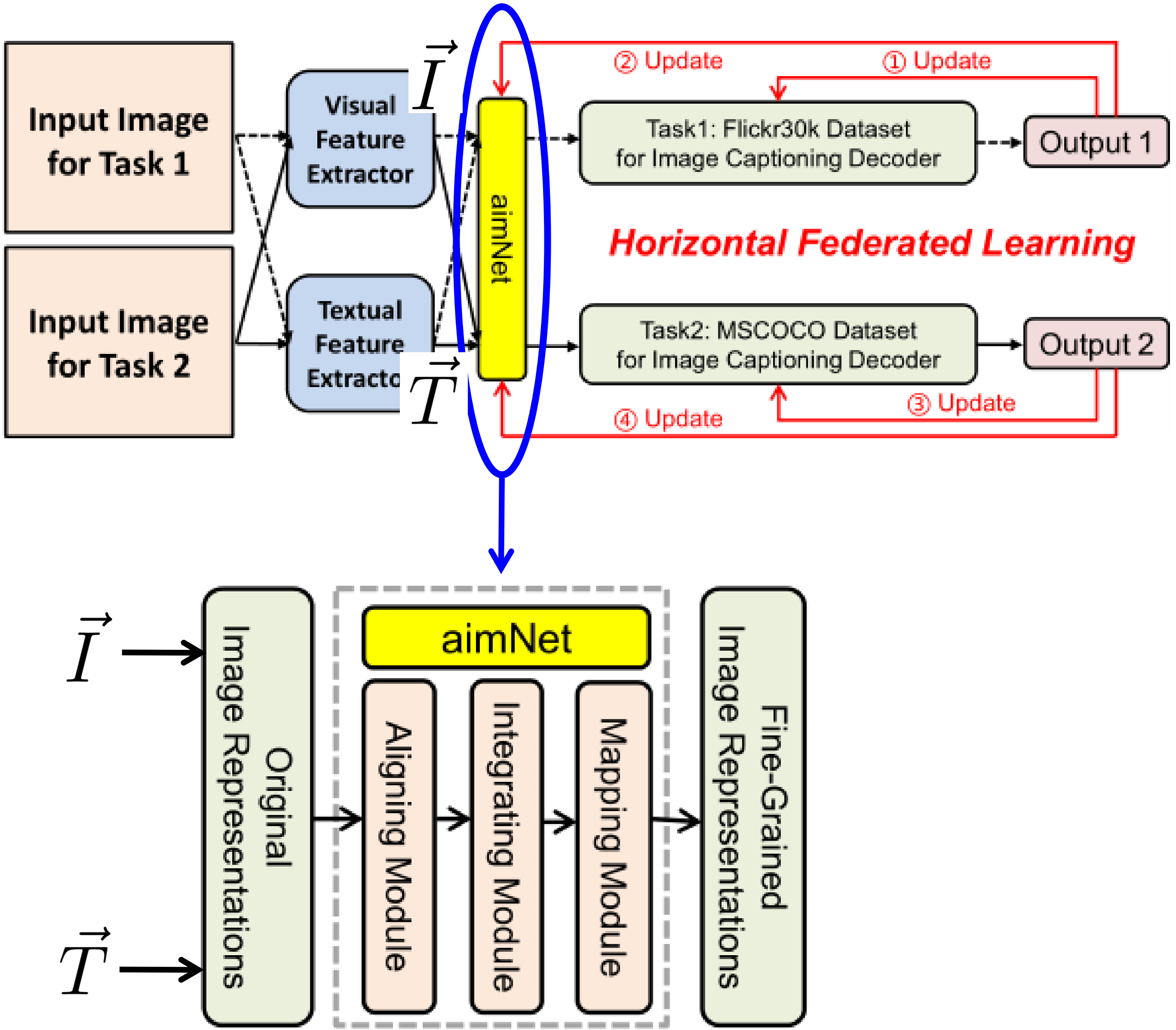
$$\mathrm{MHA}(\mathrm{Q},\mathrm{K},\mathrm{V}) = [\mathrm{Att}_1;\mathrm{Att}_2;\ldots;\mathrm{Att}_k]\mathrm{W}_k$$

Feed-Forward Network (FFN):

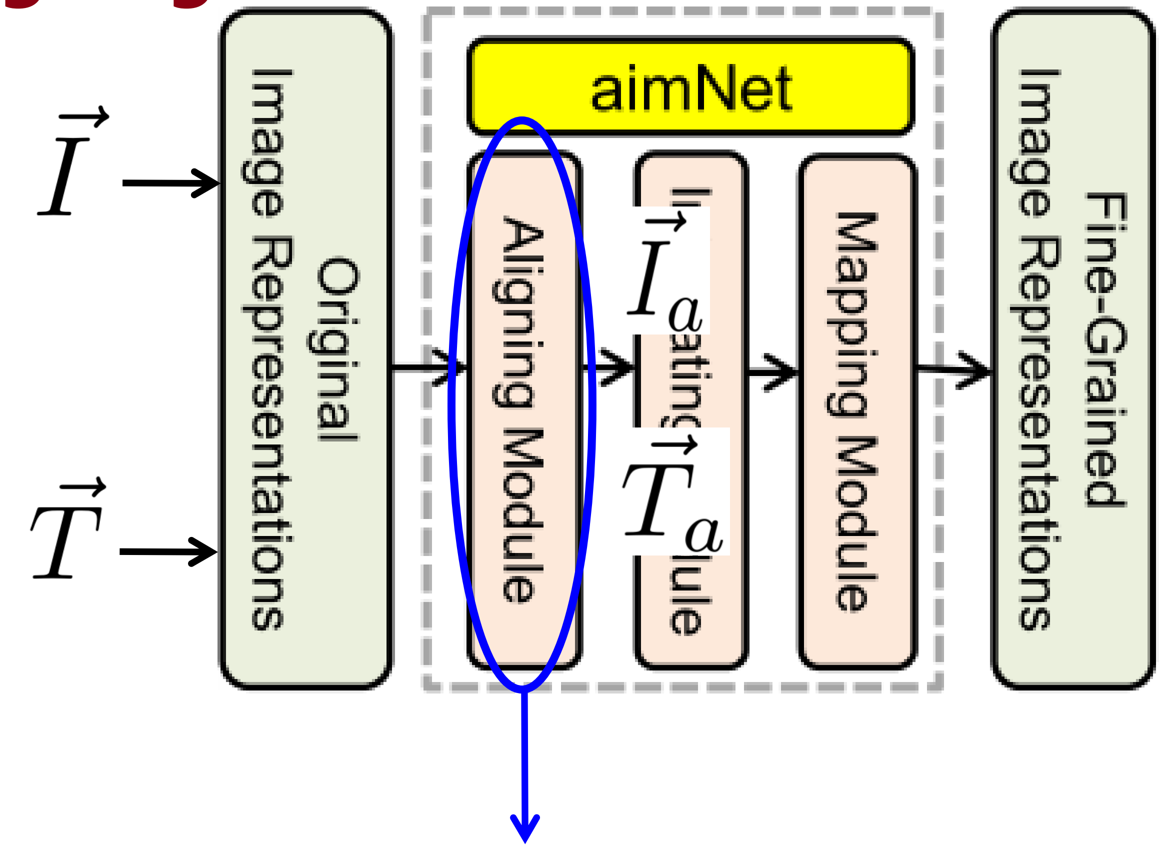
$$\mathrm{FFN}(x) = \max(0, x\mathrm{W}_{\mathrm{f}} + b_{\mathrm{f}})\mathrm{W}_{\mathrm{ff}} + b_{\mathrm{ff}}$$

The multi-head attention and feed-forward network are followed by a series of operations of shortcut connection, dropout, and layer normalization

Aligning, Integrating and Mapping Network



aimNet: Aligning Module



$$\vec{I}_a = \mathrm{FFN}(\mathrm{MHA}(\vec{I}, \vec{T}, \vec{T}))$$

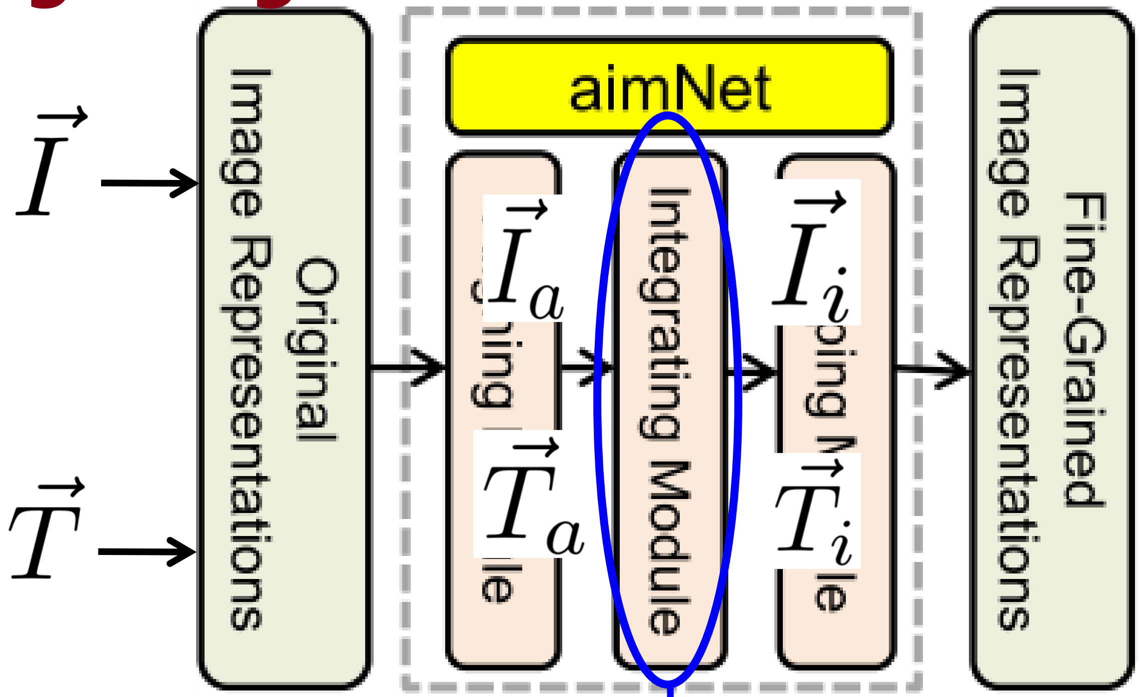
$$\vec{T}_a = \mathrm{FFN}(\mathrm{MHA}(\vec{T}, \vec{I}, \vec{I}))$$

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)

Liu et al., 2019: Aligning visual regions and textual concepts for semantic-grounded image representations. In NeurIPS 2019.

aimNet: Integrating Module



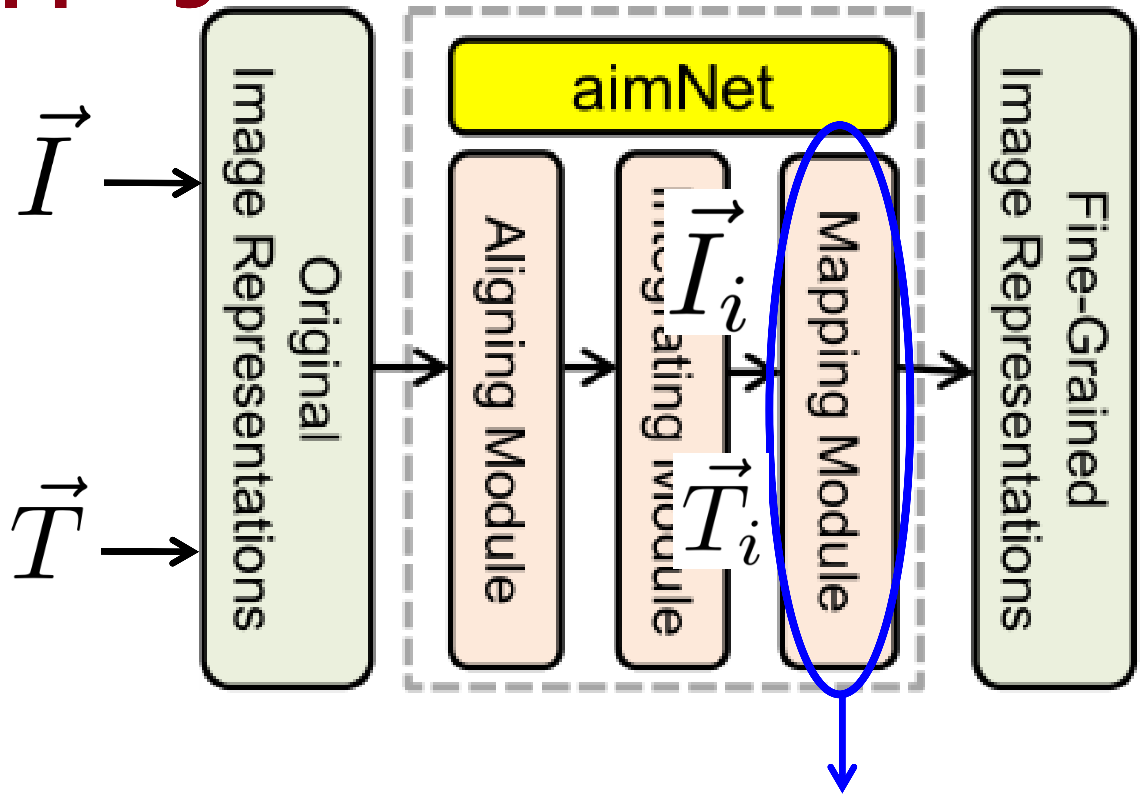
When describe an image, we often focus on one specific region and seek for other regions that often appears in the neighborhood of that region.

$$\vec{I}_i = \mathrm{FFN}(\mathrm{MHA}(\vec{I}_a, \vec{I}_a, \vec{I}_a))$$

$$\vec{T}_i = \mathrm{FFN}(\mathrm{MHA}(\vec{T}_a, \vec{T}_a, \vec{T}_a))$$

Liu et al., 2019: Exploring and distilling cross-modal information for image captioning. In IJCAI 2019.

aimNet: Mapping Module

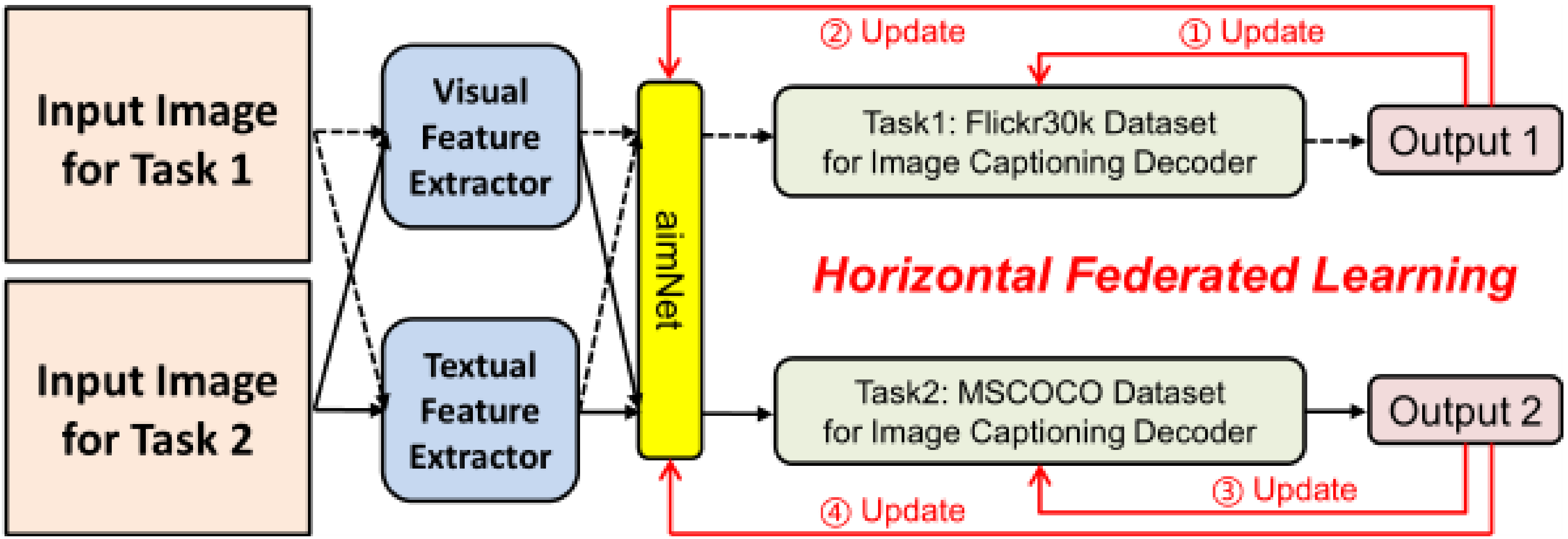


Different tasks have different data spaces, so we need to map the fine-grained image representations into the task space.

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

$$\mathrm{LayerNorm}(\mathrm{Mapping}(\vec{I}_i)+\mathrm{Mapping}(\vec{T}_i)).$$

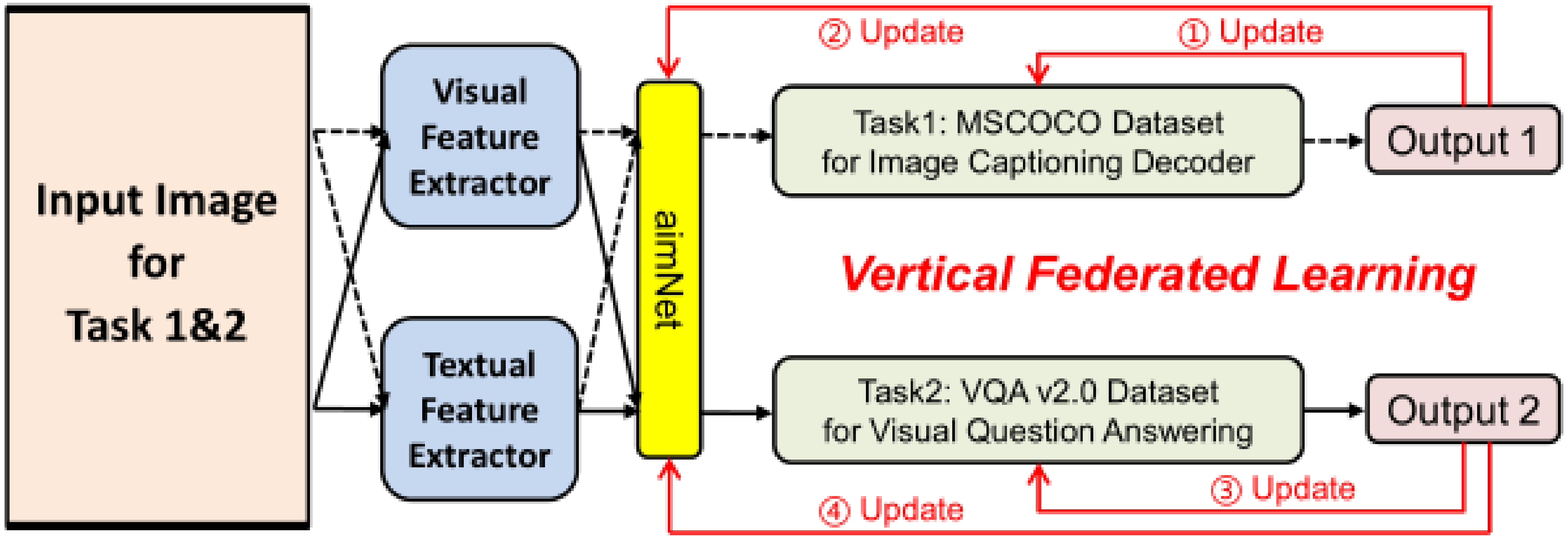
Implementation: Horizontal Federated Learning



Horizontal Federated Learning:
For example, two banks in two different cities may have **different users**, but they share the **same business**.

Implementation:
MSCOCO Image Captioning and Flickr30k Image Captioning
Same business: same task (generate captions)
Different users: different input images

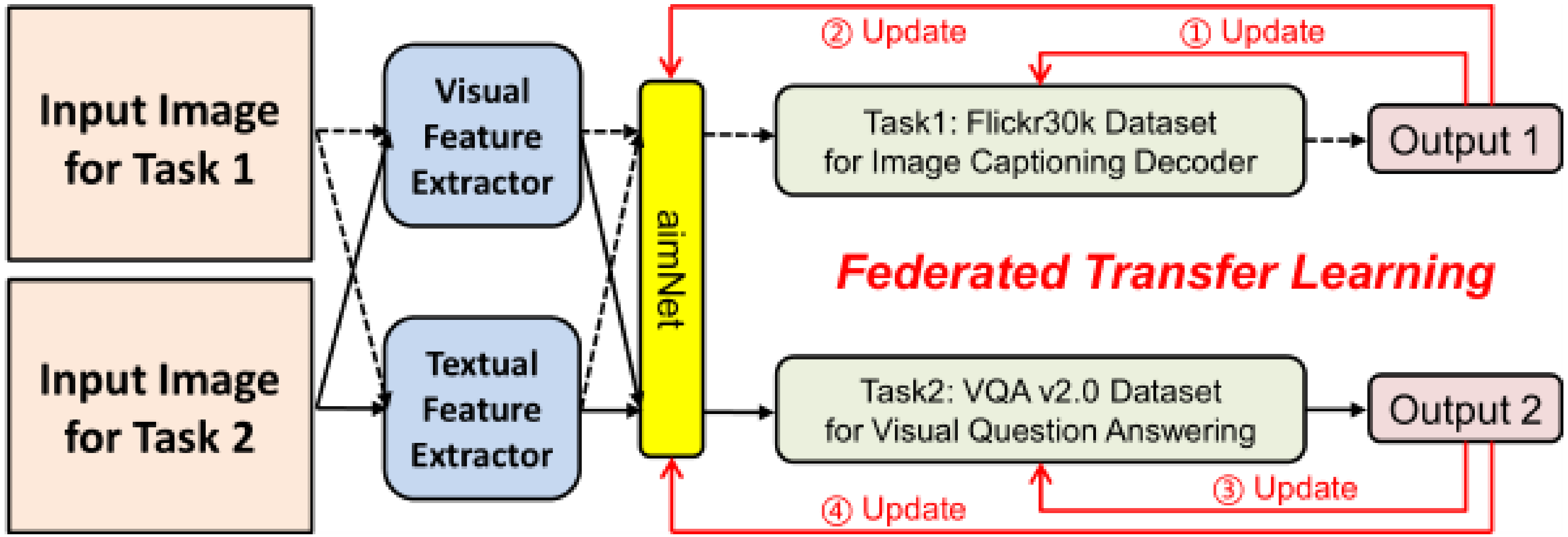
Implementation: Vertical Federated Learning



Vertical Federated Learning:
For example, two different companies in the same city, one is a bank, and the other is an insurance company, have **different business**, but **the intersection of their user space may be large**.

Implementation:
MSCOCO Image Captioning and VQA v2.0
Different business: different task
Same users: they share most of the input images

Implementation: Federated Transfer Learning



Federated Transfer Learning:
Consider the following situation, a bank is located in United States, and an insurance company is located in Europe. They have **different business**, **the intersection of their user space may be small**.

Implementation:
Flickr30k Image Captioning and VQA v2.0
Different business: different task
Different users: different input images

3 Experiments

Experiments: Image Captioning

Dataset

Microsoft COCO(MSCOCO) and Flickr30k

- ✓ Sparrow bird on branch, with beak inspecting leaves on branch.
- ✓ A bird sitting on the branch of a tree near leaves.
- ✓ A bird that is sitting in a tree.
- ✓ A bird sitting on a branch of a tree.
- ✓ A bird that is on a small branch of a tree.

Evaluation Metrics

- ✓ **CIDEr**
- ✓ **SPICE**
- ✓ BLEU
- ✓ METEOR
- ✓ ROUGE

SPICE and **CIDEr** are **customized metrics** for evaluating image captioning systems.

Experiments: Visual Question Answering (VQA)

VQAv2.0 dataset

Experiments: Horizontal Federated 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**

It enjoys an increase of
14% in CIDEr score.

Experiments: Horizontal Federated 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**

Flickr30k (small dataset):
An increase of **14%**

>

MSCOCO (large dataset):
An increase of **4%**

北京大學 PEKING UNIVERSITY

Experiments: Vertical Federated Learning

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**

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

Experiments: Vertical Federated Learning

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**

It 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.

Experiments: Federated Transfer Learning

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**

Our approach can still bring improvements to the strong baselines under the federated transfer learning settings

4 Conclusion

Conclusion

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

Thank you!

If you have any questions about the paper, you can send an email to fenglinliu98@pku.edu.cn