Bidirectional Contrastive Split Learning for Visual Question Answering

Yuwei Sun and Hideya Ochiai

The University of Tokyo, RIKEN AIP

Introduction

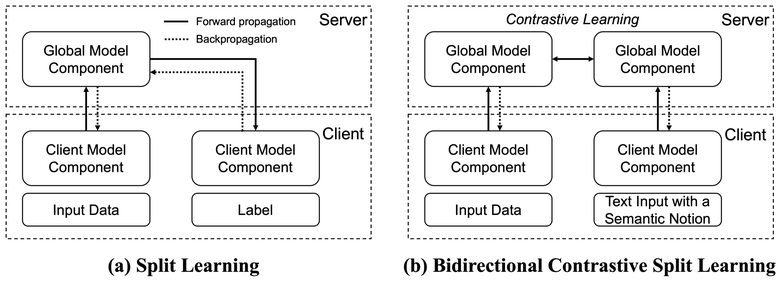
- Propose Bidirectional Contrastive Split Learning (BiCSL) to address the decentralized learning of multi-modal models.
- BiCSL can achieve competitive performance compared to a centralized method, while ensuring privacy protection and robustness against adversarial attacks.

Decentralized Visual Question Answering

Key idea: The collected vast amount of user data for training raises critical privacy concerns. Decentralized Visual Question Answering depends on learned client model weight sharing. However, sharing a complete model results in adversarial attacks and inefficient training due to constrained client resources.

Methods	Shared Data	Shared Model	Learning Framework	Loss Function
MMNas	✓	✓	Single fusion	Cross entropy
QICE	✓	✓	Single fusion	Contrastive loss
aimNet	×	✓	Federated Learning	Cross entropy
BiCSL (Ours)	×	×	Split Leaning	Contrastive loss

Table 1. BiCSL does not require sharing training data or models and is a self-supervised method without the need for training labels.



(a) Split Learning **(b) Bidirectional Contrastive Split Learning**

Figure 1. Conventional Split Learning vs. BiCSL: (a) Split Learning utilizes numeric one-hot vectors of answer labels for training, based on a unidirectional process that requires sequential processing of components resulting in longer waiting time. (b) BiCSL employs lexical semantic notions of answer texts and a bidirectional process that enables concurrent processing of model components.

- A model is divided into three components: a global component f_g and two client components $\{f_{c,1}, f_{c,2}\}$.
- The activations are sent via a forward path $f_{c,1} \rightarrow f_g \rightarrow f_{c,2}$.
- The gradients are computed via an inverse path $f_{c,1} \leftarrow f_g \leftarrow f_{c,2}$.
- Client update gradients are averaged and distributed to clients for the update of their local components.

$$\delta\theta_{c,1} = \frac{1}{K}\sum_{k\in K}\delta_k\theta_{c,1},\ \delta\theta_{c,2} = \frac{1}{K}\sum_{k\in K}\delta_k\theta_{c,2},\ \delta\theta_g = \frac{1}{K}\sum_{k\in K}\delta_k\theta_g$$

.

Bidirectional Contrastive Split Learning

Key idea: A multi-modal model is decoupled into representation modules and a contrastive module for inter-module gradients and inter-client weight sharing.

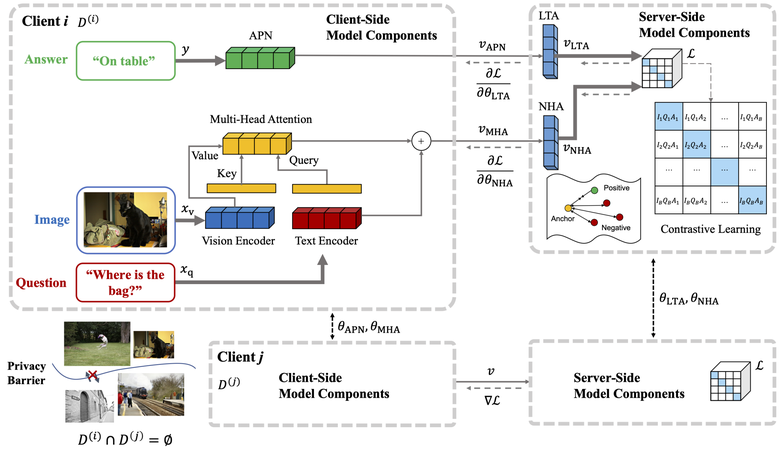


Figure 2. BiCSL comprises three key components: cross-modal learning, an answer projection network (APN) for semantic understanding of answers, and two adapter networks (LTA and NHA) for contrastive learning of different model component outputs.

- APN projects the text answer y into a feature vector $v_{\text{APN}} \in \mathbb{R}^S$.
- Different existing VQA architectures can be employed for the cross-modal learning.
- Contrastive learning aims to disentangle similar and dissimilar pairs of data points within a batch input B: $\{v_{\text{LTA},j} \mid j = i\}$ as the positive pair and the irrelevant LTA outputs $\{v_{\text{LTA},j} \mid j \neq i\}_{j=1}^{B}$ as the negative pairs, given $v_{\text{NHA},i}$.
- Every epoch t, aggregating model updates $\theta_{t+1}^{(k)} - \theta_t^{(k)}$ from different clients $k \in \{1, 2, \ldots, K\}$ enhances the generality of the global model.

$$\delta\theta_t = \frac{1}{K}\sum_{k\in\{1,2,\ldots,K\}}(\theta_{t+1}^{(k)} - \theta_t^{(k)}),\ \theta \in \{\theta_{\text{APN}}, \theta_{\text{MHA}}, \theta_{\text{NHA}}, \theta_{\text{LTA}}\}$$

.

Experiments

VQA Models	Contrastive learning (%)			
	Overall	Yes/No	Number	Other
BAN	36.23 ± 0.53	66.90 ± 0.71	12.71 ± 0.32	19.11 ± 0.47
BUTD	45.08 ± 0.64	75.82 ± 0.82	29.27 ± 0.53	25.86 ± 0.41
MFB	46.98 ± 0.58	73.95 ± 0.77	32.81 ± 0.49	30.20 ± 0.38
MCAN-s	53.18 ± 0.61	81.06 ± 0.78	41.95 ± 0.46	34.93 ± 0.35
MCAN-l	53.32 ± 0.55	81.21 ± 0.73	42.66 ± 0.39	34.90 ± 0.42
MMNas-s	51.54 ± 0.57	78.06 ± 0.79	39.76 ± 0.44	34.46 ± 0.36
MMNas-l	53.82 ± 0.53	80.06 ± 0.72	42.86 ± 0.37	36.75 ± 0.39

VQA Models	BiCSL (%)			
	Overall	Yes/No	Number	Other
BAN	35.11 ± 0.68	63.84 ± 0.54	11.06 ± 0.25	19.61 ± 0.36
BUTD	40.96 ± 0.76	66.98 ± 0.62	13.34 ± 0.35	28.74 ± 0.47
MFB	42.43 ± 0.72	68.65 ± 0.58	23.33 ± 0.41	27.52 ± 0.52
MCAN-s	48.42 ± 0.68	74.93 ± 0.54	30.88 ± 0.37	32.89 ± 0.49
MCAN-l	48.44 ± 0.62	77.44 ± 0.48	30.72 ± 0.32	32.01 ± 0.44
MMNas-s	45.14 ± 0.69	70.55 ± 0.53	28.04 ± 0.39	30.33 ± 0.48
MMNas-l	49.89 ± 0.61	74.85 ± 0.47	36.88 ± 0.34	34.33 ± 0.41

Table 2. In BiCSL, each client trains a contrastive learning-based model on its local dataset. The global model learns the entire data distribution of clients through weight sharing.

Figure 3. By introducing perturbations into images and malicious tokens at the end of questions, the combined multi-modal Trojans aim to compromise a VQA model, triggering it to produce incorrect answers.

BiCSL maintained stronger robustness against such attacks than the single fusion and split learning methods.

Self-supervised learning on input data increases the difficulty of generating effective Trojans for the attack.

Decentralized learning avoids sharing the entire model, and incomplete information about the target model degrades the success rate of attacks.

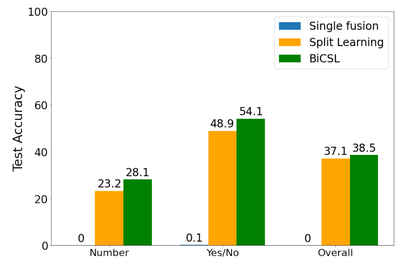


Figure 4. VQA task performance under the Trojan attack.

Future Research

Leverage approaches like differential privacy to secure the activation and gradient sharing between modules.

This motivates research in robust learning for decentralized multi-modal models.

Decentralized general intelligence capable of continual learning.