UniCon: Unidirectional Split Learning with Contrastive Loss for Visual Question Answering

Yuwei Sun* and Hideya Ochiai

ywsun@g.ecc.u-tokyo.ac.jp

Motivation

This work aims to bridge the gap between the prerequisite large-scale training and the constraint of data sharing due to confidentiality. We propose the Unidirectional Split Learning with Contrastive Loss (UniCon) to tackle Visual Question Answering (VQA) tasks for distributed data silos. UniCon trains a global model by transferring knowledge from different clients' local tasks learning refined cross-modal representations via contrastive learning. This work is the first study of VQA under the constraint of data confidentiality using self-supervised learning.

Introduction

- The real-world deployment of VQA in safety-critical applications such as healthcare needs to address the robust architecture design. Previous studies do not touch on the privacy of VQA with multi-modality data.
- The semantic notions of answers are usually not well correlated with the inputs reducing the generality of the trained model to unknown samples.

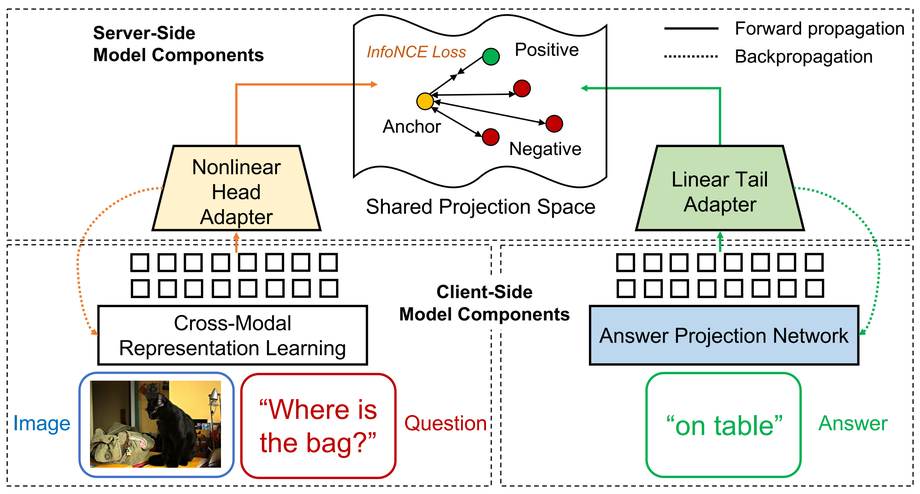
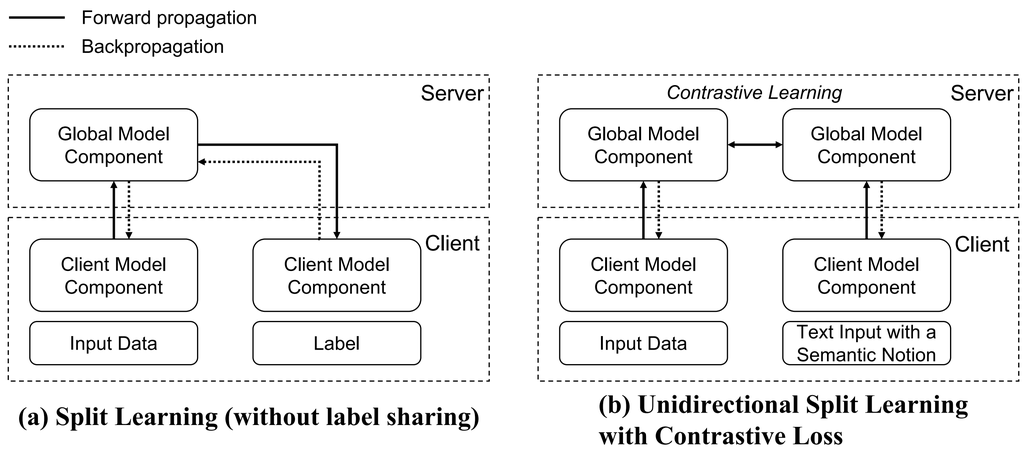


Figure 1: The overall architecture of Unidirectional Split Learning with Contrastive Loss (UniCon).

Methodology

Split Learning for Visual Question Answering

- The goal of the VQA model is to predict the correct answer y given the input pair of an image and a question (x_i, x_q).
- Split Learning (SL) guarantees the confidentiality of a client by splitting a complete model into different components, such that the input data and the complete model are not shared during the training.



(a) Split Learning (without label sharing)

(b) Unidirectional Split Learning with Contrastive Loss

Figure 2: The comparison between Split Learning and UniCon.

Unidirectional Split Learning with Contrastive Loss

We propose the use of contrastive loss in Split Learning to correlate vision contents and language semantic notions such that each model component learns better-refined cross-modal representations for the VQA tasks. Moreover, the direction of the propagation is unidirectional in UniCon, which allows the components to be computed simultaneously.

- The Answer Projection Network (APN) f_{APN} embeds the answer language contexts y into a feature vector $v_{\text{APN}} \in \mathbb{R}^P$.
- We propose the use of two adapter networks to project the outputs from different model components into the shared projection space.
- We replace a VQA model's output layer with the Nonlinear Head Adapter (NHA) network f_{NHA} that projects the high-level cross-modal representations $v_{\text{NHA}} \in \mathbb{R}^S$. The Linear Tail Adapter (LTA) projects the low-level representations v_{APN} of APN into the shared projection space $v_{\text{LTA}} \in \mathbb{R}^S$.

Learning with Information Noise Contrastive Estimation loss

- The relevant NHA and LTA outputs of the same input triplets within one training batch are employed as positive pairs. $\{(v_{\text{NHA},j}, v_{\text{LTA},j})\}_{j=1}^{B}$ where B is the batch size.
- Given a NHA output $v_{\text{NHA},i}$, any irrelevant LTA outputs $\{v_{\text{LTA},j} | j \neq i\}_{j=1}^{B}$ are employed as the negative keys of the NHA output.
- Train the model by aligning the knowledge between the component outputs in positive pairs while discouraging the similarity between the outputs in negative pairs.

$$\mathcal{L} = -\sum_{i=1}^{B} \log \frac{\exp(v_{\text{NHA},i} \cdot v_{\text{LTA},i}/\tau)}{\sum_{j=1}^{B} \mathbb{1}_{[j \neq i]} \exp(v_{\text{NHA},i} \cdot v_{\text{LTA},j}/\tau)}.$$

Results

- Extensive experiments with seven state-of-the-art VQA models.
- Contrastive learning-based models were trained on the entire training set.
- UniCon was trained on the two complementary subsets of the training set by reusing knowledge from the different local tasks, without disclosing training data and models of clients.

VQA Models	**Contrastive learning-based VQA**				**UniCon** (%)			
	Overall	Yes/No	Number	Other	Overall	Yes/No	Number	Other
BAN	36.23	66.90	12.71	19.11	35.11	63.84	11.06	19.61
BUTD	45.08	75.82	29.27	25.86	40.96	66.98	13.34	28.74
MFB	46.98	73.95	32.81	30.20	42.43	68.65	23.33	27.52
MCAN-s	53.18	81.06	41.95	34.93	48.42	74.93	30.88	32.89
MCAN-l	53.32	**81.21**	42.66	34.90	48.44	**77.44**	30.72	32.01
MMNas-s	51.54	78.06	39.76	34.46	45.14	70.55	28.04	30.33
MMNas-l	**53.82**	80.06	**42.86**	**36.75**	**49.89**	74.85	**36.88**	**34.33**

Table 1: Performance evaluation for different models.

We further studied the relation between the increasing clients $K = \{1, 2, 3, 4, 5\}$ and the model performance. The kth client has a sample size of $N^{(k)}$. $\sum_{k=1}^{K} N^{(k)} = N$, where N is the size of the entire training set.

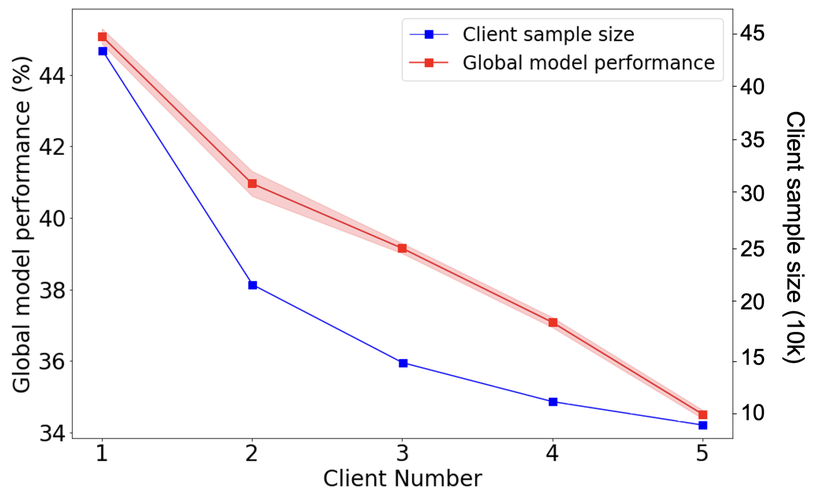


Figure 3: The tradeoff between the client number and the model performance.

Conclusion

We proposed the Unidirectional Split Learning with Contrastive loss (UniCon) to facilitate the robust learning of VQA under the constraint of data confidentiality by aligning model component outputs using contrastive learning. We presented an in-depth evaluation with the VQA-v2 dataset and a wide range of state-of-the-art models. The results showed the efficacy of UniCon in the different learning settings.