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

Yuwei Sun
The University of Tokyo
RIKEN AIP
ywsun@g.ecc.u-tokyo.ac.jp

Hideya Ochiai
The University of Tokyo
ochiai@g.ecc.u-tokyo.ac.jp

Abstract

Visual question answering (VQA) that leverages multi-modality data has attracted intensive interest in real-life applications, such as home robots and clinic diagnoses. Nevertheless, one of the challenges is to design robust learning for different client tasks. This work aims to bridge the gap between the prerequisite of large-scale training data and the constraint of client data sharing mainly due to confidentiality. We propose the Unidirectional Split Learning with Contrastive Loss (UniCon) to tackle VQA tasks training on distributed data silos. In particular, UniCon trains a global model over the entire data distribution of different clients learning refined cross-modal representations via contrastive learning. The learned representations of the global model aggregate knowledge from different local tasks. The comprehensive experiments with five state-of-the-art VQA models on the VQA-v2 dataset demonstrated the efficacy of UniCon, achieving an accuracy of 49.89% in the validation set of VQA-v2. This work is the first study of VQA under the constraint of data confidentiality using self-supervised learning.

1 Introduction

The real-world deployment of multi-modal machine learning (MMML) in safety-critical applications such as healthcare needs to address a variety of model vulnerabilities for robust architecture design. Nevertheless, in practice, previous studies do not touch on the privacy of VQA model training using multi-modality data. Intuitively, methods like federated learning (FL) can retain some data confidentiality via locally trained model parameter sharing instead of raw data sharing [McMahan et al., 2017], however, FL is vulnerable to model poisoning attacks [Hitaj et al., 2017]. Moreover, for the Visual Question Answering (VQA) tasks, one challenge is that most models are supervised where the one-hot vectors of answer classes are employed to perform a multi-class classification [Anderson et al., 2018, Kim et al., 2018]. Therefore, the semantic notions of answers are usually not well correlated with the inputs.

We tackle VQA as a self-supervised learning task instead of a multi-class classification task. The proposed method learns refined cross-modal representations from different question-image-answer triplets without the supervision of labels, such that the representations from relevant triplets stay close while the representations from irrelevant triplets are far apart (Fig. 1). This is different from contrastive learning using augmentations but learning the positive pairs between the different modality inputs instead. We termed the proposed framework **Uni**directional Split Learning with **Con**trastive Loss (**UniCon**). The main contributions of this work are the following: 1) This work studies VQA under the constraint of data confidentiality using Split Learning. 2) This paper demonstrates the contrastive learning of model components for aligning different modality representations and training a global VQA model.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

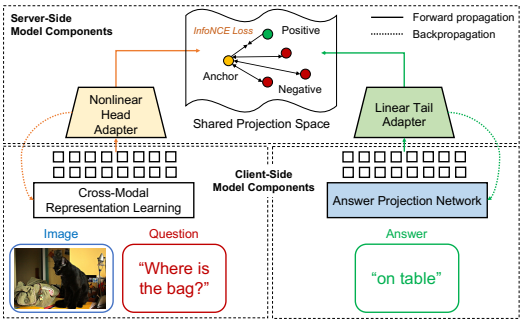

Figure 1: The overall architecture of UniCon.

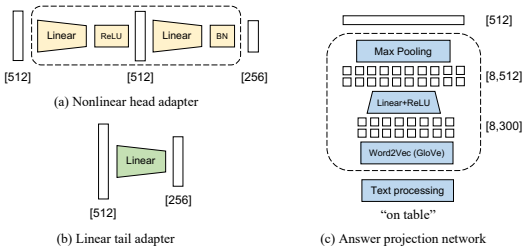

Figure 2: The architectures of the Nonlinear Head Adapter (NHA), Linear Tail Adapter (LTA), and Answer Projection Network (APN).

2 Related Work

Visual Question Answering (VQA) is the task to answer a natural language question according to the contents of a presented image. VQA is actively studied [Anderson et al., 2018, Kim et al., 2018, Yu et al., 2020] with recent years' progress in the attention mechanism [Vaswani et al., 2017]. Nevertheless, the vast majority of VQA studies so far are based on the modality network fusion methods [Yang et al., 2016, Kim et al., 2017] where the VQA task is considered a multi-class classification. Such an assumption hinders the understanding of the semantic notions embedded in the natural language answers. Moreover, in practice, these studies do not touch on the privacy of large-scale multi-modal model training in VQA.

Recent research in multi-modal machine learning has emerged in self-supervised learning (SSL)[Chopra et al., 2005, Chen et al., 2020, Zbontar et al., 2021]. In particular, contrastive learning is commonly used to learn a shared embedding space from unlabelled data, in which similar sample pairs stay close to each other while dissimilar ones from different pairs are far apart. The most recent work has also explored the implementation of SSL to VQA such as Question-Image Correlation Estimation [Zhu et al., 2020] and Video Question Answering using SSL [Kim et al., 2021]. These methods above adopt a two-step training strategy, i.e., the cross-modality pretraining with SSL and the fusion network fine-tuning. In contrast, we train a VQA model end-to-end based on the contrastive learning of different model components and determine a prediction with the similarity measurement between modalities.

3 Methods

3.1 Split Learning for Visual Question Answering

Visual Question Answering (VQA) is the task to answer questions according to given image contents. The VQA problem is usually considered a supervised learning task with a fixed list of C possible answer options. In particular, let f_{VQA} be the VQA model that takes as the input the pair of an image $x_i \in \mathbb{R}^V$ and a question $x_q \in \mathbb{R}^Q$ and outputs an answer $\hat{y} \in \{y_1, y_2, ..., y_C\}$ where $y_c \in \mathbb{R}^A$. The goal of the VQA model is to predict the correct answer y given the input pair $(x_i, x_q) \in D$ where D is the dataset. $\hat{y} = \arg\max_y p(y|x_i, x_q; f_{\text{VQA}})$ where $p(\cdot|\cdot)$ is the conditional probability.

Split Learning (SL) splits a complete model into different parts and trains a global model via interactive representation and gradient sharing. This architecture could guarantee the confidentiality of a client since both data and the model are not shared for training. Notably, a complete model is split into three different components, i.e., a global component f_g and two client components $\{f_{c,1}, f_{c,2}\}$. We assume there are K clients. The kth client has its own dataset $D^{(k)} := \{(x_{i,j}, x_{q,j}, y_j)\}_{j=1}^{N^{(k)}}$ where $N^{(k)}$ is the sample size of dataset $D^{(k)}$. Here, $\cup_{k=1}^{K} D^{(k)} = D$, $D^{(i)} \cap D^{(j)} = \emptyset \ \forall i \neq j$, and $\sum_{k=1}^{K} N^{(k)} = N$ where N is the sample size. We suppose the model components of clients share the same architecture and clients cannot share data due to confidentiality.

3.2 Unidirectional Split Learning with Contrastive Loss

3.2.1 Semantic Notion Understanding with Answer Projection Network

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 representations for the VQA tasks. We devise an Answer Projection Network (APN) f_{APN} to embed the answer language contexts y into a feature vector $v_{\text{APN}} \in \mathbb{R}^P$. APN comprises three different building blocks including a text preprocessing module, the Word2Vec using GloVe [Pennington et al., 2014], and a linear projection layer.

3.2.2 Adapter Networks and the Shared Projection Space

We propose the use of two adapter networks to project the outputs from different model components into a shared projection space. A nonlinear projection head on more complex representations can improve the performance while for simpler modality representations it is not beneficial to use the nonlinear projection [Chen et al., 2020, Alayrac et al., 2020]. In particular, we replace a VQA model f_{VQA}'s output layer with the Nonlinear Head Adapter (NHA) network f_{NHA} that projects the high-level cross-modal representations into the shared projection space $v_{\text{NHA}} \in \mathbb{R}^S$. Similarly, we devise the Linear Tail Adapter (LTA) to project the low-level representations v_{APN} of APN $v_{\text{LTA}} \in \mathbb{R}^S$. Note that v_{LTA} and v_{NHA} have the same dimension of S.

3.2.3 Learning with Information Noise Contrastive Estimation loss

The Information Noise Contrastive Estimation (InfoNCE) loss is commonly used for contrastive learning [Oord et al., 2018] to identify the positive sample from a set of unrelated negative samples. Notably, UniCon employs the relevant NHA and LTA outputs in the shared projection space of the same input triplets within one training batch as positive pairs. $\{(v_{\text{NHA},j}, v_{\text{LTA},j})\}_{j=1}^{B}$ where B is the sample size of the training batch. In contrast, 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. Then, we train the model by aligning the knowledge between the component outputs in positive pairs while discouraging the similarity between the outputs in negative pairs (Fig. 1). We formulate the loss $\mathcal{L}$ as follows

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

where τ is the temperature parameter and $\mathbb{1}_{[j \neq i]}$ is an indicator function: 1 if $j \neq i$, 0 otherwise.

For the K clients training on their local dataset $D^{(k)}$, the model aggregation of different components trained on each local dataset is aimed to improve the generality of the models. Each client k updates the local components and the server updates the global components with the aggregated update gradients. We formulate the aggregation as follows

$$\delta\theta_t = \frac{1}{K} \sum_{k \in \{1,2,\ldots,K\}} (\theta_{t+1}^{(k)} - \theta_t^{(k)}),$$

where θ is the component parameters from $\{\theta_{\text{APN}}, \theta_{\text{VQA}}, \theta_{\text{NHA}}, \theta_{\text{LTA}}\}$, where θ_{VQA} is the parameters of the VQA model f_{VQA} with the output layer removed.

3.3 Proposed Metric with Representation Similarity Measurement

We evaluate the product similarity scores between the representations v_{NHA} of an image and question input pair (x_i, x_q) from the hold-out validation dataset D_{val} of VQA and the representations $v_{\text{LTA},c}$ of C answer options $y_c \in \{y_1, y_2, \ldots, y_C\}$, where $v_{\text{LTA},c}$ denotes the representation of answer option y_c. Then, we select the answer with the highest similarity with the input as the predicted answer$\hat{y}$. We formulate the accuracy by the following

$$\hat{y} = \arg\max_c v_{\text{NHA}} \cdot v_{\text{LTA},c}, \ \text{ValAcc} = \frac{\sum_{(x_i,x_q,y) \in D_{\text{val}}} \mathbb{1}\{\hat{y} = y\}}{|D_{\text{val}}|}.$$

VQA Models	**Contrastive learning-based VQA**				**UniCon** (without data and model sharing) (%)			
	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 of the contrastive learning-based VQA and UniCon when applying different VQA models.

4 Experiments

4.1 Implementation Details

Our method is evaluated on the benchmark dataset VQA-v2 [Agrawal et al., 2017] and we report the results on its validation split. We studied our approach with the following state-of-the-art Visual Question Answering models: (1) Multi-modal Factorized Bilinear (MFB) [Yu et al., 2017]; (2) Bottom-Up and Top-Down attention mechanism (BUTD) [Anderson et al., 2018]; (3) Bilinear Attention Networks (BAN) [Kim et al., 2018]; (4) Multimodal neural architecture search (MMNas) [Yu et al., 2020]; (5) Modular Co-Attention Network (MCAN) [Yu et al., 2019b]. We used PyTorch and the OpenVQA platform [Yu et al., 2019a] to implement the VQA models. We set the hyperparameters of different VQA models to their author-recommended values. We employed the following architectures for the NHA, LTA, and APN, respectively (Fig. 2). We employed a batch size of 128, a total epoch of 20 (693400 steps), the Adam optimizer, and an initial learning rate of 0.0001 with a linear warmup of 10K steps and a decay rate of 0.2 at the epochs 10 and 15. For the InfoNCE loss, we adopted a temperature of 0.07 as in [Patrick et al., 2020].

4.2 Empirical Results

We performed extensive experiments based on the five state-of-the-art VQA models above. In particular, for MMNas and MCAN, we further considered the different model complexities. The results of the contrastive learning-based VQA are shown in Table 1. The empirical results demonstrate that the contrastive learning-based approach can be effectively applied to various VQA models. BAN showed the worst performance, particularly in counting the number. MMNas-l showed the best overall performance of 53.82% outperforming the other models in counting the number (Number) and answering the contents (Other). Nevertheless, MCAN-l performed the best in the Yes/No questions.

In reality, it is difficult to train a model with large-scale training data in VQA. In this regard, we evaluated the efficacy of UniCon for knowledge transfer in a two-client setting. We divided the training set into two subsets as the local datasets of the two clients. We show the numerical results in Table 1. There exists a trade-off between the model performance and using split learning for confidentiality. Compared to the overall accuracy of 53.82% of MMNas-l in the standalone training over the entire training set, UniCon obtained a slightly decreased accuracy of 49.89%. Nevertheless, when data sharing becomes an obstacle, the proposed approach can benefit the model training by leveraging other clients' tasks. UniCon allows clients to train over the entire data distribution without data and model sharing.

5 Conclusion

We proposed the Unidirectional Split Learning with Contrastive loss (UniCon) for Visual Question Answering, which learns refined cross-modal representations by aligning different component outputs and leveraging the knowledge of client tasks. We evaluated the efficacy on five different VQA models based on the VQA-v2 dataset. Extensive experiments showed the effectiveness and universality of our approach that can be applied to different VQA models. This work can be extended by considering a broader list of answer options using prompt engineering [Gao et al., 2021]. We hope that this work will motivate future research in robust learning of personal multimodal models.

References

A. Agrawal, J. Lu, S. Antol, and et al. VQA: visual question answering - www.visualqa.org. In *Int. J. Comput. Vis.*, volume 123, pages 4–31, 2017.

J. Alayrac, A. Recasens, R. Schneider, and et al. Self-supervised multimodal versatile networks. In *NeurIPS*, 2020.

P. Anderson, X. He, C. Buehler, and et al. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

T. Gao, A. Fisch, D. Chen, and et al. Making pre-trained language models better few-shot learners. In *ACL/IJCNLP*, 2021.

B. Hitaj, G. Ateniese, and F. Pérez-Cruz. Deep models under the GAN: information leakage from collaborative deep learning. In *ACM Conference on Computer and Communications Security*, 2017.

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

K. Kim, M. Heo, S. Choi, and B. Zhang. Deepstory: Video story QA by deep embedded memory networks. In *IJCAI*, 2017.

S. Kim, S. Jeong, E. Kim, I. Kang, and N. Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video QA. In *AAAI*, 2021.

B. McMahan, E. Moore, D. Ramage, and et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS*, 2017.

A. Oord, Y. Li, O. Vinyals, and et al. Representation learning with contrastive predictive coding. In *arXiv Preprint*, 2018.

M. Patrick, Y. M. Asano, R. Fong, and et al. Multi-modal self-supervision from generalized data transformations. In *arXiv preprint*, 2020.

J. Pennington, R. Socher, C. D. Manning, and et al. Glove: Global vectors for word representation. In *EMNLP*, 2014.

A. Vaswani, N. Shazeer, N. Parmar, and et al. Attention is all you need. In *NeurIPS*, 2017.

Z. Yang, X. He, J. Gao, and et al. Stacked attention networks for image question answering. In *CVPR*, 2016.

Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, 2017.

Z. Yu, Y. Cui, Z. Shao, P. Gao, and J. Yu. Openvqa. `https://github.com/MILVLG/openvqa`, 2019a.

Z. Yu, J. Yu, Y. Cui, and et al. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019b.

Z. Yu, Y. Cui, J. Yu, and et al. Deep multimodal neural architecture search. In *ACM Multimedia*, 2020.

J. Zbontar, L. Jing, I. Misra, and et al. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

X. Zhu, Z. Mao, C. Liu, P. Zhang, B. Wang, and Y. Zhang. Overcoming language priors with self-supervised learning for visual question answering. In *IJCAI*, 2020.