SELF-PACED CURRICULUM LEARNING FOR VISUAL QUESTION ANSWERING ON REMOTE SENSING DATA

Zhenghang Yuan[1,2], *Lichao Mou*[1,2], *Xiao Xiang Zhu*[1,2]

[1] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

Answering questions with natural language by extracting information from image has great potential in various applications. Although visual question answering (VQA) for natural image has been broadly studied, VQA for remote sensing data is still in the early research stage. For the same remote sensing image, there exist questions with dramatically different difficulty-levels. Treating these questions equally may mislead the model and limit the VQA model performance. Considering this problem, in this work, we propose a self-paced curriculum learning (SPCL) based VQA model with hard and soft weighting strategies for remote sensing data. Like human learning process, the model is trained from easy to hard question samples gradually. Extensive experimental results on two datasets demonstrate that the proposed training method can achieve promising performance.

Index Terms— visual question answering (VQA), self-paced curriculum learning (SPCL), remote sensing, deep learning

1. INTRODUCTION

Recently, novel tasks such as image captioning and visual question answering (VQA) have been developed for understanding and analyzing Earth observation data in a multi-modal way. These tasks need to take multi-modal knowledge into account, involving both computer vision (CV) and natural language processing (NLP) research fields. Among them, VQA is an important component of computer-aided systems and receives increasing attention in recent years. Given an image and the corresponding natural language questions about this image, VQA system aims to provide correct answers to the questions [1].

VQA is a challenging task, which needs to learn multi-modal feature representations from both image and language jointly. For one thing, VQA system should be able to process visual information well to understand the image. For another thing, it also needs to reason over natural language and then answer the question according to multi-modal features [2]. The great success of deep learning has enabled remarkable achievements in CV and NLP. These advances also make it possible to understand the image and natural language at a higher-semantic level.

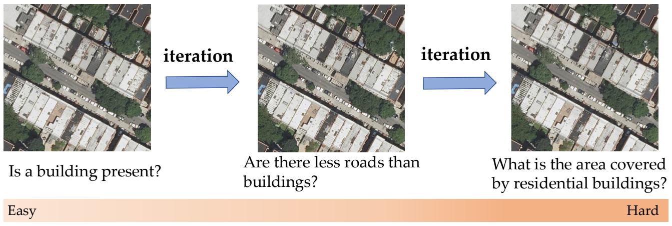
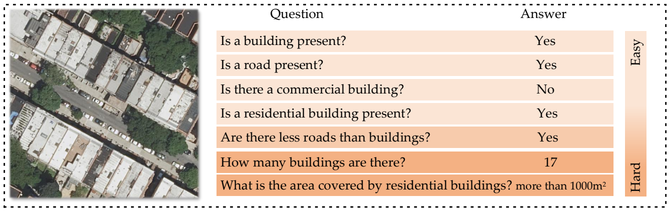

iteration
iteration
Is a building present?
Are there less roads than buildings?
What is the area covered by residential buildings?
Easy
Hard

Fig. 1. Motivation of the proposed method. We aim to train the VQA model from easy questions first and include hard ones gradually.

Attention mechanisms [3, 4] and multi-modal feature learning [5] methods are widely employed in many CV and NLP tasks. VQA for natural image has also gained benefits from these methods in recent years. However, VQA for remote sensing data is still in the early research stage and needs more exploration. Lobry et al. [6] first introduced the task of VQA for remote sensing data (RSVQA) and created two remote sensing-oriented datasets via the data from OpenStreetMap and pre-defined templates. Their work paves the way for the remote sensing imagery based VQA task and provides inspiration for later researchers.

A common rule of human learning process is to learn from easy samples first and then include hard samples gradually, as shown in Fig. 1. Inspired by this, self-paced curriculum learning (SPCL) has been proposed, which takes both prior knowledge before training and feedback during training into account [7, 8]. However, its effectiveness is still under explored for VQA task on remote sensing data. As displayed in Fig. 2, for the same remote sensing image, there exist questions with dramatically different difficulty-levels. Since VQA task involves learning both visual and language concepts, learning easy and difficult questions simultaneously may make the model confused and then limit the performance.

Considering this problem, in this work, we propose a SPCL based VQA model to mimic the human learning process. Specifically, we intend to model the incrementality and the



Question	Answer
Is a building present?	Yes
Is a road present?	Yes
Is there a commercial building?	No
Is a residential building present?	Yes
Are there less roads than buildings?	Yes
How many buildings are there?	17
What is the area covered by residential buildings?	more than 1000m²

Fig. 2. There exist questions with clearly different difficulty-levels for the same remote sensing image.

cumulative nature of human learning process in RSVQA task. By designing SPCL strategy, the proposed method can start learning from easy questions and gradually to hard questions. Experiments on two datasets have demonstrated the effectiveness of the proposed method. Overall, the contributions can be summarized as follows:

1. In order to incorporate prior guidance in the learning process, a curriculum learning method is designed for RSVQA task by measuring the difficulty of different question types. Specifically, the difficulty of the question is mainly defined by the length of the question and the prior weight.
2. SPCL with hard and soft weighting strategies is studied for RSVQA task, where model can receive both prior knowledge and dynamical learning progress information. This enables a more effective training process by learning from easy questions to hard ones gradually.
3. Extensive experiments on two datasets are conducted and the results demonstrate the effectiveness of the proposed SPCL method for RSVQA task.

2. METHODOLOGY

The overall architecture of the proposed method for RSVQA task is shown in Fig. 3. First, multi-modal features are extracted from two types of inputs, including the given image and the corresponding question. Then, visual features and language features are fused to get the multi-modal representation. In this work, we formulate the RSVQA task as a classification problem. Therefore, the answer is predicted via a classifier in the final step.

2.1. Self-paced Learning for RSVQA

In the feature learning part, the input image and question are transformed into visual and language features, respectively. Supposing that the i^{th} input image is $\mathbf{x}_i$ and the i^{th} input question is $\mathbf{q}_i$, the two types of extracted features are then fused together to form the multi-modal feature representation. At last, self-paced learning (SPL) [9] is employed to train the model by gradually including from easy to complex question samples.

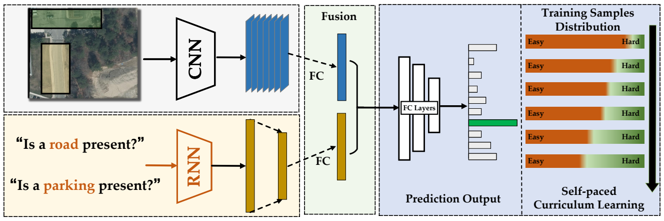


Fig. 3. Main architecture of the proposed method. Three parts are included: 1) visual and language feature learning part; 2) multi-modal feature fusion part; 3) answer prediction and SPCL training part.

Self-paced learning is a joint learning objective, which can control the training process and optimize the target task jointly in a unified framework. Specifically, SPL introduces adaptive weight for each training sample to realize an importance sampling strategy during the training process. Let $\mathbf{v} = [v_1, v_2, ..., v_N]$ be the weight vector for each sample from N training questions. The SPL loss function can be formulated as:

$$\min_{\mathbf{w},\mathbf{v}} \mathbb{E}(\mathbf{w}, \mathbf{v}, \lambda) = \sum_{i=1}^{N} v_i L\left(y_i, g\left(\mathbf{x}_i, \mathbf{q}_i, \mathbf{w}\right)\right) + f(\mathbf{v}; \lambda), \quad (1)$$

where $\mathbf{w}$ represents the weight of the network. $\mathbf{v}$ reflects the importance of samples. y_i is the training label of answer corresponding to the input image $\mathbf{x}_i$ and question $\mathbf{q}_i$. $L\left(y_i, g\left(\mathbf{x}_i, \mathbf{q}_i, \mathbf{w}\right)\right)$ is the loss function between the ground truth answer y_i and the predicted answer $g\left(\mathbf{x}_i, \mathbf{q}_i, \mathbf{w}\right)$. λ can control the learning pace, which can also be interpreted as the "age" of the deep model.

$f(\mathbf{v}; \lambda)$ is called the self-paced function or self-paced regularizer, which can control the learning scheme. During each training iteration, the weight vector $\mathbf{v}$ is updated by optimizing the self-paced function. In this work, we adopt the hard and soft regularizer for SPL. We further compare the performance of them on RSVQA task in the experiment section. Specifically, we define $f(\mathbf{v}; \lambda)$ for the hard and soft regularizer as the following equations:

$$\begin{aligned} &\textbf{Hard}: f = -\lambda \sum_{i=1}^{n} v_i, \mathbf{v} \in \{0, 1\}^N, \\ &\textbf{Soft}: f = \lambda \left(\frac{1}{2}\mathbf{v}^2 - \mathbf{v}\right), \mathbf{v} \in (0, 1)^N. \end{aligned} \quad (2)$$

Equ. 1 is a biconvex optimization problem with two disjoint blocks of variables. Alternative convex search is usually used to solve it. When the weight $\mathbf{w}$ is fixed, the global opti-

Table 1. Ablation Study on Low Resolution Dataset. Both the Mean Value and the Standard Deviation are Reported.

Types	Baseline	SPL(**Hard**)	SPL(**Soft**)	SPCL(**Hard**)	SPCL(**Soft**)
Count	72.35% (0.31%)	72.68% (0.28%)	**72.80%** (0.48%)	72.48% (0.41%)	72.59% (0.06%)
Presence	88.81% (0.02%)	88.81% (0.09%)	89.04% (0.12%)	88.74% (0.12%)	**89.68%** (0.22%)
Comparison	87.74% (0.29%)	87.30% (0.18%)	89.47% (0.11%)	87.51% (0.18%)	**89.97%** (0.05%)
Rural/Urban	82.67% (1.16%)	84.00% (1.00%)	83.67% (0.58%)	**85.33% (0.58%)**	83.67% (0.58%)
Average Accuracy	82.80% (0.35%)	83.20% (0.24%)	83.74% (0.27%)	83.52% (0.20%)	**83.97%** (0.10%)
Overall Accuracy	83.29% (0.16%)	83.36% (0.15%)	84.47% (0.46%)	83.39% (0.07%)	**84.67%** (0.07%)

mum $\mathbf{v}^*$ for the hard regularizer can be calculated by:

$$v_i^* = \begin{cases} 1, & L\left(y_i, g\left(\mathbf{x}_i, \mathbf{q}_i, \mathbf{w}\right)\right) \leq \lambda, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Similarly, the global optimum $\mathbf{v}^*$ for the soft regularizer can be computed by:

$$v_i^* = \begin{cases} -\frac{L}{\lambda} + 1, \text{ if } L\left(y_i, g\left(\mathbf{x}_i, \mathbf{q}_i, \mathbf{w}\right)\right) \leq \lambda, \\ 0, \text{ otherwise.} \end{cases} \tag{4}$$

Then, the model weight $\mathbf{w}$ is updated with the fixed $\mathbf{v}$. The values of loss function for easy question are usually smaller than those for hard ones. If the loss value L is smaller than λ, the corresponding questions will be considered as relatively easy samples to train the model with higher priority. As the value of λ increases, the model will involve relatively harder questions with larger loss values. During the training process, λ, i.e., the "age" of the deep model increases gradually along with the training iteration. In this work, we record the maximum and minimum loss value of the epoch $t-1$, and use these two values to update λ. Specifically, λ can be computed as:

$$\lambda = (\max(L^{t-1}) - \min(L^{t-1})) \cdot K + \min(L^{t-1}), \tag{5}$$

where K is used to adjust the value of λ.

2.2. Self-paced Curriculum Learning for RSVQA

Although SPL is useful for learning samples gradually from easy to hard ones, it does not incorporate prior guidance in the learning process. At the beginning, the network weights are randomly initialized, and the loss values of easy and hard examples may not be accurate to determine the true difficulty level of each question. Thus, incorporating prior knowledge is necessary to deal with this situation [10].

Different types of questions tend to have different difficulty-levels. For instance, the question "How many buildings are there?" is obviously much harder than the question "Is a building present?". This is mainly because counting is a more difficult task than the classification task. In addition, longer questions are usually more difficult than shorter ones. Inspired by this observation, we design an effective curriculum learning method based on question length (QL) and prior weight for different question types.

$$\min_{\mathbf{w},\mathbf{v}} \mathbb{E}(\mathbf{w}, \mathbf{v}, \lambda, \Psi) = \sum_{i=1}^{N} v_i L\left(y_i, g\left(\mathbf{x}_i, \mathbf{q}_i, \mathbf{w}\right)\right) + f(\mathbf{v}; \lambda),$$
$$\text{s.t. } \mathbf{v} \in \Psi, \tag{6}$$

where $\Psi = \left\{\mathbf{v} \mid \mathbf{a}^T\mathbf{v} \leq c\right\}$ is the pre-defined curriculum region, where the weight vector $\mathbf{v}$ is initialized.

In practical, we can define a_i by a ranking function $a_i = W(q_i) * QL(q_i)$, where $W(q_i)$ is the pre-defined prior weight for different question types. $QL(q_i)$ indicates the length of question, which is calculated by summing the number of words in the question and divided by the max question length. As presented in Equ. 6, the curriculum region can be obtained with the defined ranking function $\mathbf{a}$ and a constant value c.

3. EXPERIMENTS AND DISCUSSION

The proposed method is evaluated on two public RSVQA datasets from [6]: the Low Resolution (LR) and the High Resolution (HR) dataset. There are 772 images of size 256×256 with 77,232 questions and answers in the LR dataset. HR dataset contains 10,659 images of size 512×512 with 1,066,316 questions and answers.

Since the test set in LR dataset is not publicly released, we use cross-validation for the performance evaluation. For the HR dataset, we follow the settings in [6] for evaluation. Adam optimizer with the initial learning rate 1e-5 is used for training. The batch size for both datasets is 70. Besides, 150 epochs are used to train the model in the LR dataset and 35 epochs are used in the HR dataset. Note that curriculum learning is only used to initialize the weights in the first 15 epochs.

We take the method of [6] as the baseline, and replace the traditional cross entropy loss with the proposed loss function. Detailed accuracy of different question types, average accuracy and overall accuracy are used as the evaluation measurements. Each model is trained 3 times. The mean value and the standard deviation are reported for both datasets.

The experimental results of the LR dataset are shown in Table 1. For the LR dataset, there are four question types:

Table 2. Experiment Comparision on the Test Set 1 of HR Dataset. Both the Mean Value and the Standard Deviation are Reported.

Types	Baseline	SPCL(**Soft**)
Count	68.63% (0.11%)	**68.91%** (0.03%)
Presence	90.43% (0.04%)	**90.66%** (0.08%)
Comparison	88.19% (0.08%)	**89.07%** (0.27%)
Area	85.24% (0.05%)	**85.66%** (0.26%)
Average Accuracy	83.12% (0.03%)	**83.57%** (0.11%)
Overall Accuracy	83.23% (0.02%)	**83.69%** (0.11%)

Count, Presence, Comparison and Rural/Urban, and the corresponding prior weights $W(q_i)$ are $\{\text{Count} : 4.0, \text{Presence} : 1.0, \text{Comparison} : 3.0, \text{Rural/Urban} : 1.0\}$. Table 1 shows that compared with the baseline method, using SPL is effective for performance improvement of VQA task. Moreover, "SPL (Soft)" can achieve better performance than "SPL (Hard)". The results show that jointly using prior knowledge (curriculum learning) and adaptive re-weighting (SPL) can further enhance the performance. Moreover, we find that using soft weights $\mathbf{v} \in (0, 1)^N$ is more effective than hard ones $\mathbf{v} \in \{0, 1\}^N$.

Table 2 shows the comparison results on the HR dataset. For the HR dataset, there are four question types: Count, Presence, Comparison and Area, and the corresponding prior weights $W(q_i)$ are set as $\{\text{Count} : 4.0, \text{Presence} : 1.0, \text{Comparison} : 3.0, \text{Area} : 4.0\}$. From the results we can see that the proposed "SPCL (Soft)" can effectively enhance the performance of RSVQA for different question types. By simply replacing the traditional cross entropy loss with SPCL loss, the proposed method can achieve better performance. This reveals that the training strategy "learning from easy to hard" is effective for RSVQA task.

4. CONCLUSION

In this paper, we propose a SPCL based VQA model with hard and soft weighting strategies for remote sensing data to train the model from easy to hard question samples. SPCL based VQA model can take both prior knowledge before training and dynamic feedback during training into account. Prior knowledge contains the difficulty-levels of the questions including the length and the prior weight. Dynamic feedback means that when the loss function is less than a certain value, the sample will participate in training. Ablation studies and comparisons with the baseline method are conducted on the LR and HR datasets. The experimental results demonstrate that the proposed method can achieve promising performance.

5. ACKNOWLEDGMENT

This work is supported by the China Scholarship Council.

6. REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[2] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Underst.*, vol. 163, pp. 21–40, 2017.

[3] Yuansheng Hua, Lichao Mou, and Xiao Xiang Zhu, "Relation network for multilabel aerial image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4558–4572, 2020.

[4] Lichao Mou and Xiao Xiang Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 110–122, 2019.

[5] Zhitong Xiong, Yuan Yuan, and Qi Wang, "Ask: Adaptively selecting key local features for rgb-d scene recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2722–2733, 2021.

[6] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[7] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann, "Self-paced curriculum learning," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, 2015, pp. 2694–2700.

[8] Mrinmaya Sachan and Eric Xing, "Easy questions first? a case study on curriculum learning for question answering," in *Proceedings of the Association for Computational Linguistics*, 2016, pp. 453–463.

[9] M Pawan Kumar, Benjamin Packer, and Daphne Koller, "Self-paced learning for latent variable models.," in *NIPS*, 2010, vol. 1, p. 2.

[10] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.