

# Training Recurrent Answering Units with Joint Loss Minimization for VQA

Hyeonwoo Noh

Bohyung Han

Dept. of Computer Science and Engineering, POSTECH, Pohang, Korea  
`{shgusdngogo, bhan}@postech.ac.kr`

## Abstract

We propose a novel algorithm for visual question answering based on a recurrent deep neural network, where every module in the network corresponds to a complete answering unit with attention mechanism by itself. The network is optimized by minimizing loss aggregated from all the units, which share model parameters while receiving different information to compute attention probability. For training, our model attends to a region within image feature map, updates its memory based on the question and attended image feature, and answers the question based on its memory state. This procedure is performed to compute loss in each step. The motivation of this approach is our observation that multi-step inferences are often required to answer questions while each problem may have a unique desirable number of steps, which is difficult to identify in practice. Hence, we always make the first unit in the network solve problems, but allow it to learn the knowledge from the rest of units by backpropagation unless it degrades the model. To implement this idea, we early-stop training each unit as soon as it starts to overfit. Note that, since more complex models tend to overfit on easier questions quickly, the last answering unit in the unfolded recurrent neural network is typically killed first while the first one remains last. We make a single-step prediction for a new question using the shared model. This strategy works better than the other options within our framework since the selected model is trained effectively from all units without overfitting. The proposed algorithm outperforms other multi-step attention based approaches using a single step prediction in VQA dataset.

## Introduction

Visual Question Answering (VQA) (Antol et al. 2015) is the problem to answer questions related to an input image. Contrary to traditional visual recognition tasks, which concentrate on a predefined problem such as object detection, scene classification, activity recognition, etc., VQA involves various recognition tasks at the same time and provides a unified approach to solve the problems defined by questions. Due to these reasons, VQA requires substantial amount of learning to capture information from images and understand questions.

Recently, deep learning based approaches using multiple reasoning steps with attention (Yang et al. 2016; Xiong, Mentity, and Socher 2016) have been proposed to improve the performance of VQA systems. After extract-

ing features from an image and a question using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), respectively, these methods iteratively generate answers by combining question feature with image feature of local area given by attention. This approach is natural as many questions in VQA conceptually require multiple steps for reasoning. For example, to answer the question like “what is in front of the giraffe?” the VQA model should be able to solve the subtasks in sequel as “finding a giraffe,” “identifying the region in front of the giraffe” and “classifying object in the specified region.”

One of the critical limitations of the existing methods is that the number of steps for reasoning is fixed to a predefined number. Even though the exact number of steps required to answer a question is difficult to figure out, it is reasonable to assume that different questions need the different numbers of steps to reach solutions. For example, a question “what color is the apple on the table?” requires more steps than another question “what color is the apple?” If the number of inference steps is not adaptive to questions, we may need to solve multiple subtasks in a single step or a single subtask in multiple steps, which are likely to degrade system performance by causing overfitting or underfitting of trained models.

Instead of presetting the number of steps, we make the model learn the number of steps for inference of each question implicitly by simply minimizing joint loss from multiple steps of our recurrent deep neural network without extra supervision. We believe that learning to predict with a proper number of steps not only enhance the overall accuracy but also help each answering unit to generalize better. However, identifying the optimal number of steps is an extremely difficult task, and we propose an indirect but simple solution to learn a unified model for inference. By analyzing the answers from each step, we find out that predictions with more steps tend to overfit to easier questions even if later steps could solve more complex questions better than the earlier ones. Based on this observation, we progressively stop backpropagating the loss from the overfitting units. Training is terminated when the accuracy of the first answering unit is saturated. With this training strategy, we empirically show that the single-step prediction based on the first unit, whose model is trained with joint loss from multiple steps, outperforms other training-testing variations, and the accuracy is further improved by integrating early-stopping training

strategy. The main contribution of the proposed algorithm is summarized below:

- We propose a novel architecture based on a deep recurrent neural network for VQA, which is composed of multiple answering units with shared parameters and optimized by minimizing joint loss from multiple units.
- We show that the VQA model involving multiple reasoning steps tends to overfit to easier questions quickly, and develop a unique training strategy to early-stop overfitting units progressively until the accuracy of the first unit is saturated.
- The proposed algorithm outperforms other multi-step attention based approaches using a single step prediction in VQA dataset.

The rest of the paper is organized as follows. We first review related work in Section and discuss our main idea with motivation in Section . Section describes the architecture to implement our idea and the training and testing methods of the proposed algorithm. We present experimental results of our algorithm in the standard public benchmark dataset in Section .

## Related Work

The VQA problem is first addressed by (Malinowski and Fritz 2014), which proposes a Bayesian framework to combine symbolic reasoning and uncertain visual information. (Zhou et al. 2015) has proposed a shallow neural network for VQA, which accepts the CNN features of images and the Bag-of-Words (BoW) representations of questions as its inputs. Recently, VQA algorithms are often formulated with deep neural networks since the techniques based on deep learning have straightforward end-to-end training procedure by back-propagation and present competitive performance in terms of accuracy. These approaches typically pose VQA problems as simple classification tasks based on the joint features from images and questions, where CNNs and LSTMs (Hochreiter and Schmidhuber 1997) are employed to encode images and questions, respectively (Ren, Kiros, and Zemel 2015; Malinowski, Rohrbach, and Fritz 2015) while only CNNs are used for both encoding and classification in (Ma, Lu, and Li 2016).

Several VQA systems (Yang et al. 2016; Xiong, Merity, and Socher 2016; Shih, Singh, and Hoiem 2016) attempt to identify relevant regions in the input image to answer questions, which is often performed by the soft attention mechanism (Xu et al. 2015). Specifically, (Shih, Singh, and Hoiem 2016) utilizes a single step attention to an object proposal for VQA while (Xiong, Merity, and Socher 2016) also employs multi-step attention based on the dynamic attention network (Kumar et al. 2016). Stacked attention network (Yang et al. 2016) is motivated by memory networks (Sukhbaatar et al. 2015; Kumar et al. 2016), and constructs two successive attention modules for multi-step inferences.

Several approaches employ the network architectures adaptive to input questions. Dynamic parameter prediction network (Noh, Seo, and Han 2016) determines the parameters of a fully connected layer in CNN given a question,

and constructs a unified model to handle various tasks related to input images. With similar motivation, neural module networks combine multiple module networks to construct a single deep network (Andreas et al. 2016a; Andreas et al. 2016b), where the module combinations are given by syntactic parsing of questions.

Deep neural networks are sometimes trained with multiple supervisions to facilitate training procedure, where losses are backpropagated from multiple branches. GoogLeNet (Szegedy et al. 2015) attaches auxiliary classifiers to a few intermediate layers of the network to provide additional supervision for training. Deeply supervised nets (Lee et al. 2015) has a companion objective function to avoid the vanishing gradient problem. Deeply recursive convolutional network (Kim, Lee, and Lee 2016) provides supervision to every recurrent convolutional layer and makes inference based on an ensemble of individual predictions. Our approach has something common with (Kim, Lee, and Lee 2016) in that supervision is given to each recurrent unit, but is differentiated since we analyze role of each supervision and propose a novel training strategy of early stopping.

## Overview

This section describes the general formulation of VQA and discusses our approach based on recurrent deep neural network with its motivation.

## Problem Formulation

We formulate VQA problem as a classification task. For a given image  $I$  and a question  $q$ , a VQA model predicts the best answer  $a^*$ , which is formally given by

$$a^* = \underset{a \in \Omega}{\operatorname{argmax}} p(a|I, q; \theta) \quad (1)$$

where  $\Omega$  denotes a set of all possible answers and  $\theta$  is a set of model parameters in the network. There exist a lot of different methods to implement this probabilistic framework, but we focus on the models based on deep neural networks, which are frequently employed for VQA problems in recent years.

The models based on deep neural networks are typically composed of three main components: image encoder, question encoder, and answering module. Image and question encoders extract image feature  $f_I$  from an input image  $I$  and question feature  $f_q$  from a question sentence  $q$ , respectively. Answering module takes extracted image and question features, and generates an answer by combining information from the image and the question. Therefore, the module should be able to perform various tasks defined by questions. In this work, we propose a deep neural network architecture and investigate a training strategy for the answering module.

## Main Idea

The motivation behind this work is our observation that solving a task defined by a question often requires the capability to solve a sequence of atomic subtasks. However, the kinds and numbers of the subtasks in the sequence vary in individual questions, and, in addition, the same subtasks may appear in any places within the sequence depending on the

question. Due to these reasons, it is extremely difficult to develop a unified algorithm to handle all the variations. Instead, to overcome the challenges indirectly, we design a novel neural network architecture for VQA and propose a training strategy with early-stopping based on a simple joint loss minimization.

The overall architecture of the proposed algorithm is illustrated in Figure 1. The main component of the proposed network is an answering unit. Each answering unit is capable of solving the full task; based on the features extracted from image and question, it predicts an answer and updates its hidden state. By concatenating multiple answering units sequentially, we infer a series of answers, which are predicted by the model integrating subtasks progressively and solve more and more composite tasks. This procedure is implemented as a recurrent neural network whose recurrent unit corresponds to each answering unit.

We need to train the proposed network so that it can answer a given question by solving a series of subtasks one by one using the answering units. The main challenge to this objective is that it is difficult to know the optimal number of steps to solve each problem since there are various question types with different complexity. To circumvent this problem, we always make the first unit in the network solve problems, but allow it to learn the knowledge from the rest of units by backpropagation unless it degrades the model. For the purpose, we simply provide the same supervision to every step in the unfolded recurrent neural network and optimize all the answering units with shared parameters jointly using the standard backpropagation technique for RNNs. We also propose a unique training technique with early-stopping strategy, which is useful to avoid overfitting in complex models among multiple answering units.

## Algorithm

This section discusses the proposed answering module and the procedure of training and testing. We provide more detailed description about the answering modules in the supplementary document.

### Answering Module

Answering module is a recurrent neural network, which is composed of multiple answering units with shared model parameters as illustrated in Figure 1. Using a given image feature map  $\mathbf{f}_I \in \mathbb{R}^{P \times L}$  (with  $P$  channels and  $L$  locations), a question feature  $\mathbf{f}_q \in \mathbb{R}^Q$  and an initial memory state  $\mathbf{h}_{\text{mem}}^0 \in \mathbb{R}^R$ , answering module provides a sequence of answer probability vectors  $\mathbf{a}^k \in \mathbb{R}^C$  ( $k = 1, \dots, K$ ), where  $K$  is the number of answering units. Formally, the  $k^{\text{th}}$  answering unit predicts an answer probability as

$$\mathbf{a}^k = \text{Answer}_k(\mathbf{f}_I, \mathbf{f}_q, \mathbf{h}_{\text{mem}}^{k-1}), \quad (2)$$

where  $\text{Answer}_k(\cdot)$  is the answering operation in the  $k^{\text{th}}$  answering unit.

The function  $\text{Answer}_k(\cdot)$  requires several internal operations to predict answer probability as illustrated in Figure 2. The first operation is to compute a subtask feature, which is

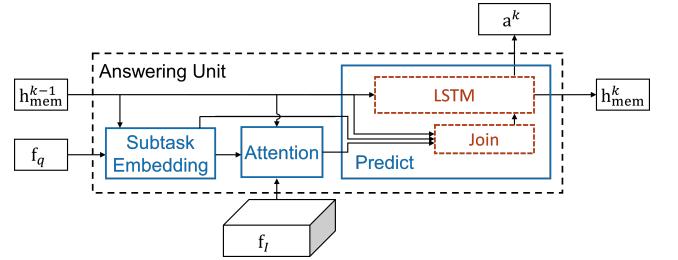


Figure 2: The illustration of the answering unit which comprises subtask embedding, attention and predict operation.

extracted based on the question and the history of performed subtasks in the previous steps as follows:

$$\mathbf{f}_{\text{sub}}^k = \text{Subtask}_k(\mathbf{f}_q, \mathbf{h}_{\text{mem}}^{k-1}; \theta_{\text{sub}}), \quad (3)$$

where  $\theta_{\text{sub}}$  is parameter for the subtask module.

This subtask feature is used to perform an attention operation, which finds a relevant location in an image feature map with its corresponding feature vector based on the key  $\mathbf{f}_{\text{sub}}^k$ . We employ soft attention mechanism (Xu et al. 2015) and the attention operation is formally given by

$$(\mathbf{f}_{\text{loc}}^k, \mathbf{f}_{\text{att}}^k) = \text{Attend}_k(\mathbf{f}_I, \mathbf{f}_{\text{sub}}^k, \mathbf{h}_{\text{mem}}^{k-1}; \theta_{\text{att}}), \quad (4)$$

where  $\theta_{\text{att}}$  is the parameters for soft attention, and  $\mathbf{f}_{\text{loc}}^k$  and  $\mathbf{f}_{\text{att}}^k$  denote attention probability map and attended feature, respectively.

The last operation is  $\text{Predict}_k(\cdot)$  function. It receives subtask feature, attention information, and the memory state in the previous step to produce the final answer and updates hidden state. This operation involves LSTM to update hidden state in memory and the answer probability is given by applying a softmax function. The prediction operation is formally defined as

$$(\mathbf{a}^k, \mathbf{h}_{\text{mem}}^k) = \text{Predict}_k(\mathbf{f}_{\text{sub}}^k, \mathbf{f}_{\text{loc}}^k, \mathbf{f}_{\text{att}}^k, \mathbf{h}_{\text{mem}}^{k-1}), \quad (5)$$

and the two outputs are specifically given by

$$\mathbf{a}^k = \text{softmax}(\mathbf{f}_{\text{sub}}^k, \mathbf{f}_{\text{loc}}^k, \mathbf{f}_{\text{att}}^k, \mathbf{h}_{\text{mem}}^k; \theta_{\text{soft}}) \quad (6)$$

$$\mathbf{h}_{\text{mem}}^k = \text{LSTM}(\{\mathbf{f}_{\text{sub}}^k, \mathbf{f}_{\text{loc}}^k, \mathbf{f}_{\text{att}}^k\}, \mathbf{h}_{\text{mem}}^{k-1}; \theta_{\text{LSTM}}), \quad (7)$$

where  $\theta_{\text{LSTM}}$  is the parameters for LSTM and  $\theta_{\text{soft}}$  is the parameters for softmax classifier.

To implement the Answering module, we need to learn the parameters for the all internal operations, and it is performed by the standard backpropagation technique. The detailed training procedure is described next.

### Training

We minimize joint loss from multiple steps by providing the ground-truth answer  $\mathbf{y} \in \{0, 1\}^C$  to every step  $k$  as a supervision. Loss from a single step is computed by a cross entropy between the answer probability  $\mathbf{a}^k$  and the ground-truth answer  $\mathbf{y}$ . Loss from each step is simply aggregated to compute the overall loss. Given image  $I_n$ , question  $q_n$ , and

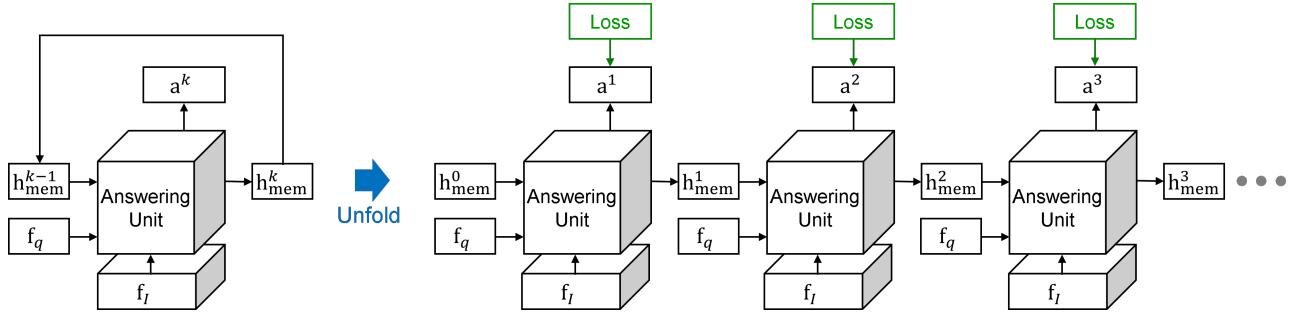


Figure 1: Overall architecture of the proposed network. The proposed network is a recurrent deep neural network, where each recurrent unit corresponds to a complete module for visual question answering. For training, we unfold the network to predict answer and give supervision for every steps. For testing, we use a single answering unit to answer a question about an image.

ground-truth label  $\mathbf{y}_n$  of the  $n^{\text{th}}$  training example, the joint loss function is formally given by

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K -\mathbf{y}_n^T \log (\mathbf{a}_n^k), \quad (8)$$

where

$$\mathbf{a}_n^k = \text{Answer}_k(\mathbf{f}_I(I_n; \theta_I), \mathbf{f}_q(q_n; \theta_q), \mathbf{h}_{\text{mem}}^{k-1}; \theta_{\text{ans}}). \quad (9)$$

Note that  $\theta_I$  and  $\theta_q$  denote parameters for image and question encoder, respectively. The model parameter for the answering module is given by  $\theta_{\text{ans}} = \{\theta_{\text{sub}}, \theta_{\text{att}}, \theta_{\text{soft}}, \theta_{\text{LSTM}}\}$ .

We train the model end-to-end by backpropagation, where the objective function based on cross entropy in Eq. (8) is minimized. Given an image  $I_n$  and a question  $q_n$ , image encoder and question encoder extract features,  $\mathbf{f}_I$  and  $\mathbf{f}_q$ , respectively. The extracted features are given to the individual unfolded answering units. As the answering module itself is a recurrent neural network, we compute the gradients of all the parameters in  $\theta_{\text{ans}}$  by backpropagation through time (BPTT). Note that the backpropagated loss to the inputs of answering module,  $\mathbf{f}_I$  and  $\mathbf{f}_q$ , is used to learn the parameters of the image and question encoders,  $\theta_I$  and  $\theta_q$ .

When the network is trained with the joint loss from multiple steps as in Eq. (8), we observe that models with multiple steps generally overfit to training data easily and show lower validation/testing accuracy than the model in the first step as illustrated in Figure 3. Note that this observation also coincides with (Yang et al. 2016), which states that using three or more attention does not improve the performance any further. We believe that the answering units in later steps have more model capacity and can fit to training data better while it loses the generality of models. Such overfitting tendency may ruin the models in the earlier steps since the losses in the later steps are propagated backwards until the first answering unit. To circumvent this issue, we stop backpropagating losses from the overfitted answering units progressively as soon as we identify their overfitting. In practice, we validate the model in every epoch and early-stop training from an overfitted answering unit if the validation accuracy of the answering unit drops more than a predefined threshold from its maximum value. According to our observation, the answering units in the later steps typically are terminated first

and the first answering unit always manages to stay alive. If validation dataset is not available, we early-stop training based on the formula for convenience, which is empirically determined as

$$t_{\text{stop}}^k = \text{round} \left[ (t_{\max} - t_{\min}) \frac{e^{\lambda(K-k)} - 1}{e^{\lambda(K-1)} - 1} + t_{\min} \right], \quad (10)$$

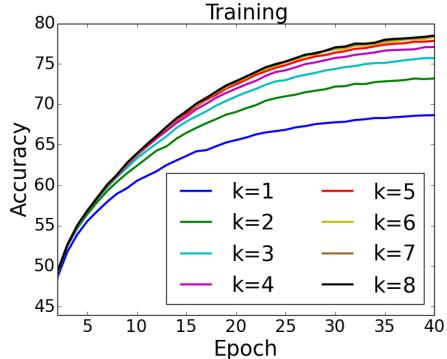
where  $t_{\min}$  is the number of epochs before the first early stopping,  $t_{\max}$  is the total number of epochs for training and  $t_{\text{stop}}^k$  is the number of epochs before training the  $k^{\text{th}}$  answering unit is terminated. The early stopping is scheduled by controlling the configuration parameter  $\lambda$ .

## Testing

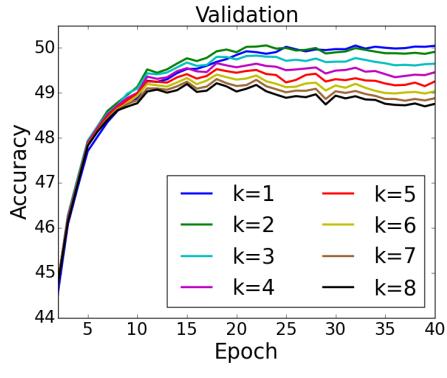
In testing time, we predict answers only from the first answering unit in our model since it has better generalization accuracy in our experiments. Although we believe that some complex tasks can be solved better in the later steps, it is very difficult to know which step is optimal to find the correct solution. Hence, selecting one of the answers from multiple answering units is not feasible in practice. Also, the ensemble of the answers from multiple units is not particularly better than the solution from the first unit. Note that, since the losses from the later steps are always backpropagated until the first answering module, it gradually learns how to solve more complex problems and does not tend to overfit by employing early-stopping strategy.

## Comparison to Memory Network based Approaches

At first glance, the proposed method resembles the VQA methods such as (Yang et al. 2016; Xiong, Merity, and Socher 2016), which are based on memory network (Sukhbaatar et al. 2015). However, we introduce a novel training strategy with early-stopping and a simple testing method with a single-step inference. This decision is based on our observations in VQA problems, but it is contradictory to (Yang et al. 2016; Xiong, Merity, and Socher 2016). The previous works (Yang et al. 2016; Xiong, Merity, and Socher 2016) state that multi-step training and testing without parameter sharing is advantageous, but our results suggest that multi-step training with parameter sharing and a single-step testing may be better in



(a) Training accuracy



(b) Validation accuracy

Figure 3: Training and validation accuracy curve for varying  $k$ . The prediction from the later step shows higher training accuracy but lower validation accuracy.

practice. This is partly supported by Figure 3 although the experiment setting is not exactly identical. Also, our experiment supports the claim that benefit from multi-step training with weight sharing is larger than that from multi-step training and testing without sharing weights.

## Experiment

### Dataset and Evaluation Metric

We train and test the proposed network in VQA dataset (Antol et al. 2015), which borrows images from MSCOCO dataset (Lin et al. 2014) and collects questions and answers via Amazon Mechanical Turk. The dataset consists of 248,349 questions for training, 121,512 questions for validation, and 244,302 questions for testing. For each image, 3 questions are asked and 10 independent answers are given to each question. There are two test datasets; we typically use *test-dev* split for the control experiments and *test-standard* for comparison with external algorithms.

Two tasks are defined on VQA dataset: open-ended task and multiple-choice task. The model has to predict an answer for an open-ended question without knowing predefined candidate answers, but select one of 18 candidate answers in multiple-choice task. In both cases, the answers given by a model are evaluated by the following metric reflecting human

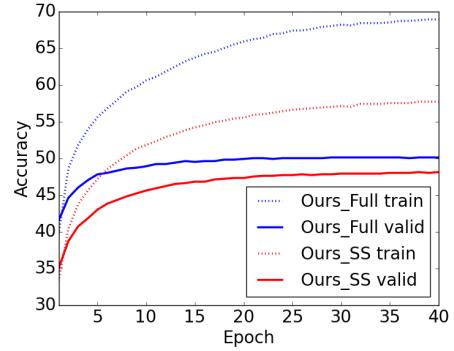


Figure 4: Training and validation accuracy of Ours\_SS and Ours\_Full.

consensus:

$$\text{Acc}_{\text{VQA}}(ans) = \min \left( \frac{\#\text{humans that said } ans}{3}, 1 \right) \quad (11)$$

where the score for a question is proportional to the number of matches with ground-truth answers, and the tested model receives full credit for each question if at least 3 people agree to the predicted answer.

### Implementation Details

We use VGG-16 net (Simonyan and Zisserman 2015) and ResNet-101 (He et al. 2016) as image encoders. After rescaling input images to  $448 \times 448$ , we extract the feature maps from the last pooling layer in VGG or the layer below global average pooling layer in ResNet. The question features are given by a 2-layer LSTM<sup>1</sup>; the final hidden and cell states in both layers are concatenated to be used as a question feature. The dimensionality of the attended image feature and the sub-task feature are both 512. For better generalization, we apply dropout with rate of 0.5 to the features for input images and questions in each answering unit independently. The vocabulary size of our model is 14,772 and the rest of words are converted to UNK tokens. The set of possible answers ( $\Omega$  in Eq. (1)) contains 1,000 answers with the highest frequencies among all answers in the training dataset.

We set the number of steps  $K$  to 8 for multi-step training, and determine the parameters in Eq. (10),  $t_{\max}$ ,  $t_{\min}$ , and  $\lambda$ , using validation set. We use Adam (Kingma and Ba 2015) for optimization. Learning rate for question encoder and answering modules are set to  $3 \times 10^{-3}$  and  $3 \times 10^{-4}$ , respectively, and the image encoder is not fine-tuned. Both learning rates are decayed in every epoch by the factor of 0.9. Additionally, we inject random noises sampled from Gaussian distribution to the gradient as suggested in (Neelakantan et al. 2016), where  $\eta$  is the number of training iterations. To alleviate the exploding gradient problem, we limit the magnitude of gradient vector to 0.1 by normalization. To facilitate reproduction, we make our code publicly available<sup>2</sup>.

<sup>1</sup>[https://github.com/VT-vision-lab/VQA\\_LSTM\\_CNN](https://github.com/VT-vision-lab/VQA_LSTM_CNN)

<sup>2</sup>[http://cvlab.postech.ac.kr/research/rau\\_vqa/](http://cvlab.postech.ac.kr/research/rau_vqa/)

Table 1: Single model performance on the VQA test-dev dataset of all compared algorithms including the variations of our algorithm. Asterisk (\*) denotes the concurrent submission with this paper.

	Open-Ended				Multiple-Choice			
	All	Y/N	Num	Others	All	Y/N	Num	Others
BoW Q (Antol et al. 2015)	48.1	75.7	36.7	27.1	53.7	75.7	37.1	38.6
I (Antol et al. 2015)	28.1	64.0	00.4	03.8	30.5	69.9	00.5	03.8
Bow Q+I (Antol et al. 2015)	52.6	75.6	33.7	37.4	59.0	75.6	34.4	50.3
LSTM Q (Antol et al. 2015)	48.8	78.2	35.7	26.6	54.8	78.2	36.8	38.8
LSTM Q+I (Antol et al. 2015)	53.7	78.9	35.2	36.4	57.2	79.0	35.8	43.4
(Antol et al. 2015)	57.8	80.5	36.8	43.1	62.7	80.5	38.2	53.0
(Andreas et al. 2016a)	54.8	77.7	37.2	39.3	-	-	-	-
(Zhou et al. 2015)	55.7	76.6	35.0	42.6	61.7	76.7	37.1	54.4
(Shih, Singh, and Hoiem 2016)	-	-	-	-	62.4	77.6	34.3	55.8
(Noh, Seo, and Han 2016)	57.2	80.7	37.2	41.7	62.5	80.8	38.9	52.2
(Andreas et al. 2016b)	57.9	80.5	37.4	43.1	-	-	-	-
(Yang et al. 2016)	58.7	79.3	36.6	46.1	-	-	-	-
(Wu et al. 2016)	59.2	81.0	38.4	45.2	-	-	-	-
(Xiong, Merity, and Socher 2016)	60.3	80.5	36.8	48.3	-	-	-	-
(Kim et al. 2016) ResNet*	61.7	82.3	38.8	49.3	-	-	-	-
(Lu et al. 2016) ResNet*	61.8	79.7	38.7	51.7	65.8	79.7	40.0	59.8
(Fukui et al. 2016) ResNet*	64.7	82.5	37.6	55.6	69.1	-	-	-
Ours_SS	59.0	78.3	37.1	47.6	64.0	78.4	38.2	57.5
Ours_MS	61.0	81.3	36.8	49.1	65.8	81.3	38.9	58.7
Ours_FULL	61.3	81.5	37.0	49.6	66.1	81.5	39.5	58.9
Ours_ResNet	63.3	81.9	39.0	53.0	67.7	81.9	41.1	61.5

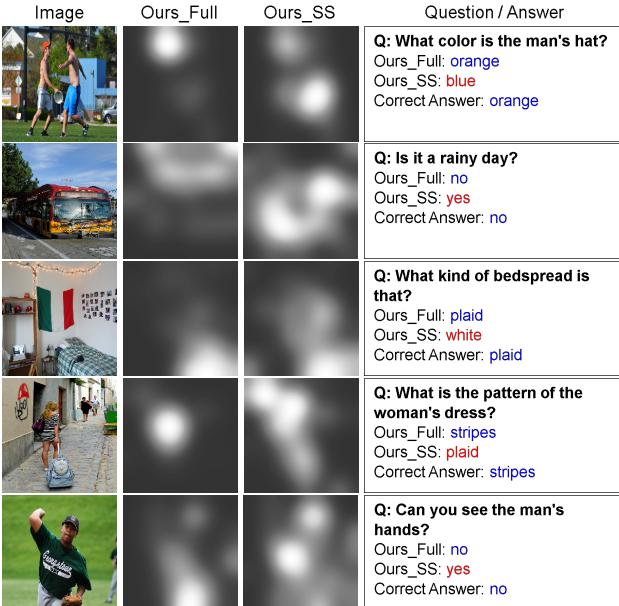


Figure 5: Qualitative comparisons of attention between Ours\_Full and Ours\_SS.

## Results and Analysis

To ensure the effectiveness of the multi-step training and early-stopping strategy, we perform several control experiments. The baseline is the proposed model without both components, which is referred to as Ours\_SS. We also test the model trained with multi-step training but without early-stopping strategy, Ours\_MS, and our full model denoted by

Table 2: Comparisons of single model performance in the VQA test-standard. Asterisk (\*) denotes the concurrent submission with this paper.

	Open-Ended				Multiple-Choice			
	All	Y/N	Num	Others	All	Y/N	Num	Others
Human (Antol et al. 2015)	83.3	95.8	83.4	72.7	-	-	-	-
(Antol et al. 2015)	58.2	80.6	36.5	43.7	63.1	80.6	37.7	53.6
(Andreas et al. 2016a)	55.1	-	-	-	-	-	-	-
(Zhou et al. 2015)	55.9	76.8	35.0	42.6	62.0	76.9	37.3	54.6
(Shih, Singh, and Hoiem 2016)	-	-	-	-	62.4	77.2	33.5	56.1
(Noh, Seo, and Han 2016)	57.4	80.3	36.9	42.2	62.7	80.4	38.8	52.8
(Andreas et al. 2016b)	58.0	-	-	-	-	-	-	-
(Yang et al. 2016)	58.9	-	-	-	-	-	-	-
(Wu et al. 2016)	59.4	81.1	37.1	45.8	-	-	-	-
(Xiong, Merity, and Socher 2016)	60.4	80.4	36.8	48.3	-	-	-	-
(Kim et al. 2016) ResNet*	61.8	82.4	38.2	49.4	66.3	82.4	39.6	58.4
(Lu et al. 2016) ResNet*	62.1	-	-	-	66.1	-	-	-
Ours.ResNet	63.2	81.7	38.2	52.8	67.3	81.7	40.0	61.0

Ours\_FULL employ both strategies for training. We use the image features extracted from VGG-16 net (Simonyan and Zisserman 2015) for all the control experiments.

Table 1 illustrates the single model performance of various approaches trained without data augmentation. The proposed algorithm, Ours\_FULL outperforms the models based on multi-step attention (Kumar et al. 2016; Yang et al. 2016; Lu et al. 2016) in the VQA dataset using the same image encoder. The best result is achieved by the work based on the multimodal compact bilinear pooling (Fukui et al. 2016). We believe that the multimodal compact bilinear pooling can be integrated within the proposed model, but we leave it as a future work.

Note that the performance gain from Ours\_SS to Ours\_FULL is about 2.3, which are significant in VQA context. Both training schemes turn out to be useful for performance improvement. These results are interesting because multi-step training is effective even for a single-step prediction and early-stopping strategy improves accuracy as long as all active steps are free from overfitting. Since we schedule early-stopping based on the formula from simple empirical study, the accuracy may be further improved with more sophisticated validation.

Figure 4 illustrates training and validation accuracy of Ours\_FULL and Ours\_SS. The proposed training strategy denoted by Ours\_FULL outperforms Ours\_SS consistently even though they have the exactly same architecture for a single step prediction during testing. Figure 6 presents predicted answers with attended regions. Visualization of attention shows that Ours\_FULL tends to focus on critical regions while Ours\_SS are often distracted by irrelevant objects. We believe that our training strategy helps the model learn to answer questions in appropriate steps in training while it enables the answering unit to implicitly solve a series of subtasks given a question in testing.

## Comparison with Stacked Attention Network

The architecture of our single answering unit is similar to the stacked attention network (SAN) (Yang et al. 2016). The difference lies in how to handle questions that require multi-

ple steps to answer. As the number of required steps varies across questions and it is difficult to figure out the desirable number, SAN uses the same fixed number of steps for training and testing, which might result in overfitting observed in Figure 3. Instead, we provide multiple supervision in our network, but hope a single answering unit (the first one, in practice) to learn the best model for evaluation in testing. The comparison between the SAN and the proposed method in Table 1 clearly shows the effectiveness of our method.

## Experimental Results with ResNet

We can improve the accuracy of a VQA system by using a better image encoder. Hence, we train the model with image features extracted from ResNet-101 (He et al. 2016), and this model is denoted by Ours.ResNet. The model is trained with same hyper-parameters with Ours.FULL in Table 1, and evaluated both in VQA test-dev and test-standard. The results from Ours.ResNet is presented in Table 1 and 2, where we observe that ResNet improves performance substantially.

## Conclusion

We proposed a VQA algorithm based on a recurrent deep neural network, which is trained by minimizing the joint loss from all answering units. The answering units share model parameters while the outputs from the units in the later steps depend on the results from the ones in the earlier steps. To maximize performance, we introduce an early stopping training strategy, where individual answering units are disregarded in training as soon as they start to overfit to the training set. Only the first answering unit is employed for inference since it is the model learned from all the multiple answering units without overfitting. The proposed architecture illustrates the outstanding performance in the standard VQA dataset without data augmentation. We believe that our algorithm has great potential to be used as a general framework for VQA problems by replacing our answering unit with any other networks.

## References

- [Andreas et al. 2016a] Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016a. Deep compositional question answering with neural module networks. In *CVPR*.
- [Andreas et al. 2016b] Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016b. Learning to compose neural networks for question answering. In *NAACL*.
- [Antol et al. 2015] Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: visual question answering. In *ICCV*.
- [Fukui et al. 2016] Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.
- [He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- [Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- [Kim et al. 2016] Kim, J.-H.; Lee, S.-W.; Kwak, D.-H.; Heo, M.-O.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Multimodal residual learning for visual qa. *arXiv preprint arXiv:1606.01455*.
- [Kim, Lee, and Lee 2016] Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Deeply-recursive convolutional network for image super-resolution. In *CVPR*.
- [Kingma and Ba 2015] Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- [Kumar et al. 2016] Kumar, A.; Irsoy, O.; Su, J.; Bradbury, J.; English, R.; Pierce, B.; Ondruska, P.; Gulrajani, I.; and Socher, R. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*.
- [Lee et al. 2015] Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; and Tu, Z. 2015. Deeply-supervised nets. In *AISTATS*.
- [Lin et al. 2014] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: common objects in context. In *ECCV*.
- [Lu et al. 2016] Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*.
- [Ma, Lu, and Li 2016] Ma, L.; Lu, Z.; and Li, H. 2016. Learning to answer questions from image using convolutional neural network. In *AAAI*.
- [Malinowski and Fritz 2014] Malinowski, M., and Fritz, M. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*.
- [Malinowski, Rohrbach, and Fritz 2015] Malinowski, M.; Rohrbach, M.; and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*.
- [Neelakantan et al. 2016] Neelakantan, A.; Vilnis, L.; Le, Q. V.; Sutskever, I.; Kaiser, L.; Kurach, K.; and Martens, J. 2016. Adding gradient noise improves learning for very deep networks. In *ICLR Workshop*.
- [Noh, Seo, and Han 2016] Noh, H.; Seo, P. H.; and Han, B. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*.
- [Ren, Kiros, and Zemel 2015] Ren, M.; Kiros, R.; and Zemel, R. S. 2015. Exploring models and data for image question answering. In *NIPS*.
- [Shih, Singh, and Hoiem 2016] Shih, K. J.; Singh, S.; and Hoiem, D. 2016. Where to look: Focus regions for visual question answering. In *CVPR*.
- [Simonyan and Zisserman 2015] Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.
- [Sukhbaatar et al. 2015] Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *NIPS*.
- [Szegedy et al. 2015] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*.
- [Wu et al. 2016] Wu, Q.; Wang, P.; Shen, C.; Hengel, A. v. d.; and Dick, A. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*.
- [Xiong, Merity, and Socher 2016] Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *ICML*.
- [Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

[Yang et al. 2016] Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *CVPR*.

[Zhou et al. 2015] Zhou, B.; Tian, Y.; Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2015. Simple baseline for visual question answering. In *arXiv 1512:02167*.

# Training Recurrent Answering Units with Joint Loss Minimization for VQA

## Supplementary Document

This document provides our implementation details and additional results that could not be accommodated in the main paper due to space limitation.

### Implementation Details

In this section, we provide implementation details for each components of answering units described in Eq. (2) to (7) of the main paper. Without loss of generality, we describe the architecture of the  $k^{\text{th}}$  answering unit.

#### Subtask Module

Subtask module in Eq. (3) of the main paper generates sub-task feature  $\mathbf{f}_{\text{sub}} \in \mathbb{R}^S$  from a question feature  $\mathbf{f}_q \in \mathbb{R}^Q$  and the previous memory state  $\mathbf{h}_{\text{mem}}^{k-1} \in \mathbb{R}^M$ . This operation is given by

$$\mathbf{f}_{\text{sub}}^k = \tanh(\mathbf{W}_q \mathbf{f}_q + \mathbf{W}_h \mathbf{h}_{\text{mem}}^{k-1} + \mathbf{b}_{\text{sub}}), \quad (12)$$

where  $\mathbf{W}_q \in \mathbb{R}^{S \times Q}$ ,  $\mathbf{W}_h \in \mathbb{R}^{S \times M}$  and  $\mathbf{b}_{\text{sub}} \in \mathbb{R}^S$  are weight parameters.

#### Attention Module

Attention module in Eq. (4) of the main paper generates the attention probability map  $\mathbf{f}_{\text{loc}}^k \in \mathbb{R}^L$  and attended feature  $\mathbf{f}_{\text{att}}^k \in \mathbb{R}^S$  based on a soft attention mechanism. The computation of  $\mathbf{f}_{\text{loc}}^k$  and  $\mathbf{f}_{\text{att}}^k$  requires embedding of input image feature map  $\mathbf{f}_I \in \mathbb{R}^{P \times L}$  (with  $P$  channels and  $L$  locations), and the embedded feature map  $\mathbf{g}_I$  is given by

$$\mathbf{g}_I = \tanh(\mathbf{W}_I \mathbf{f}_I + \mathbf{b}_I), \quad (13)$$

where  $\mathbf{W}_I \in \mathbb{R}^{S \times P}$  and  $\mathbf{b}_I \in \mathbb{R}^S$  are weight parameters.

The procedure to compute attention probability map  $\mathbf{f}_{\text{loc}}^k \in \mathbb{R}^L$  is composed of the following three steps. First, we compute pre-attention score  $\alpha_k \in \mathbb{R}^L$  based on the embedded image feature map  $\mathbf{g}_I \in \mathbb{R}^{S \times L}$  and the subtask feature  $\mathbf{f}_{\text{sub}}^k$  by the following equation:

$$\alpha_k^T = \mathbf{W}_\alpha^1 \tanh(\mathbf{W}_\alpha^2 \mathbf{g}_I + \mathbf{W}_\alpha^3 \mathbf{f}_{\text{sub}}^k \mathbf{1}^T + \mathbf{b}_\alpha \mathbf{1}^T), \quad (14)$$

where  $\mathbf{W}_\alpha^1 \in \mathbb{R}^{L \times A}$ ,  $\mathbf{W}_\alpha^2 \in \mathbb{R}^{A \times S}$ ,  $\mathbf{W}_\alpha^3 \in \mathbb{R}^{A \times S}$  and  $\mathbf{b}_\alpha \in \mathbb{R}^A$  are weight parameters, and  $\mathbf{1} \in \mathbb{R}^L$  is one vector. Next, we compute attention score  $\beta_k \in \mathbb{R}^L$  by adding another pre-attention score extracted from the previous memory  $\mathbf{h}_{\text{mem}}^{k-1}$ , which is given by

$$\beta_k = \alpha_k + \mathbf{W}_\beta \mathbf{h}_{\text{mem}}^{k-1} + \mathbf{b}_\beta, \quad (15)$$

where  $\mathbf{W}_\beta \in \mathbb{R}^{L \times M}$  and  $\mathbf{b}_\beta \in \mathbb{R}^L$  are weight parameters. Last, attention probability map  $\mathbf{f}_{\text{loc}}^k \in \mathbb{R}^L$  is given by applying softmax function to the attention score  $\beta_k \in \mathbb{R}^L$  as described below:

$$\mathbf{f}_{\text{loc}}^k = \frac{\exp(\beta_k)}{\mathbf{1}^T \exp(\beta_k)}. \quad (16)$$

Once  $\mathbf{f}_{\text{loc}}^k$  is obtained, attended feature  $\mathbf{f}_{\text{att}}^k \in \mathbb{R}^S$  is computed using embedded image feature  $\mathbf{g}_I$  and attention probability map  $\mathbf{f}_{\text{loc}}^k$  as follows:

$$\mathbf{f}_{\text{att}}^k = \mathbf{g}_I \mathbf{f}_{\text{loc}}^k. \quad (17)$$

#### Prediction Module

Prediction module in Eq. (5) of the main paper has two sub-module: LSTM and softmax classification. As specified in the main paper, we use publicly available LSTM implementation. Single layer LSTM is constructed and the dimensionality of LSTM hidden state is equal to that of memory state  $\mathbf{h}_{\text{mem}}^{k-1} \in \mathbb{R}^S$ . Input to the LSTM, denoted by  $\mathbf{f}_{\text{join}}^k$ , is computed by

$$\mathbf{f}_{\text{join}}^k = \mathbf{f}_{\text{sub}}^k + \mathbf{f}_{\text{att}}^k + \mathbf{W}_{\text{join}} \mathbf{f}_{\text{loc}}^k + \mathbf{b}_{\text{join}}, \quad (18)$$

where  $\mathbf{W}_{\text{join}} \in \mathbb{R}^{S \times L}$  and  $\mathbf{b}_{\text{join}} \in \mathbb{R}^S$  are weight parameters. Answer probability  $\mathbf{a}^k \in \mathbb{R}^C$  is obtained by applying a softmax function, which is given by

$$\mathbf{a}^k = \frac{\exp(\mathbf{f}_{\text{join}}^k + \mathbf{W}_s \mathbf{h}_{\text{mem}}^k + \mathbf{b}_s)}{1^T \exp(\mathbf{f}_{\text{join}}^k + \mathbf{W}_s \mathbf{h}_{\text{mem}}^k + \mathbf{b}_s)}, \quad (19)$$

where  $\mathbf{W}_s \in \mathbb{R}^{C \times S}$  and  $\mathbf{b}_s \in \mathbb{R}^C$  are weight parameters.

#### Network Parameters

Network parameters used for the experiments are as follows:  $S = 512$ ,  $Q = 1024$ ,  $M = 512$ ,  $P = 512$ ,  $L = 196$ ,  $A = 256$ , and  $C = 1000$ .

### Additional Results

Figure 6 presents qualitative comparison between Ours\_Full and Ours\_SS on VQA validation dataset.

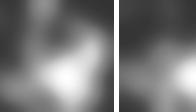
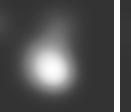
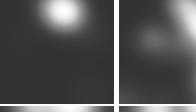
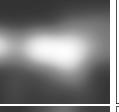
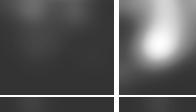
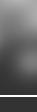
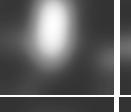
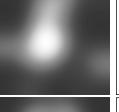
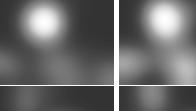
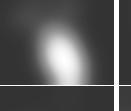
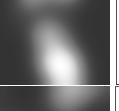
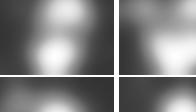
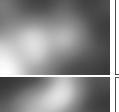
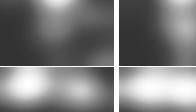
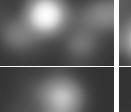
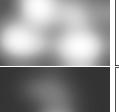
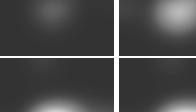
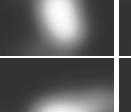
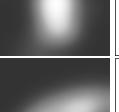
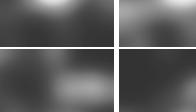
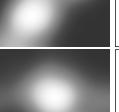
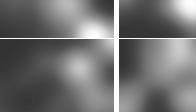
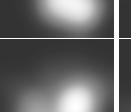
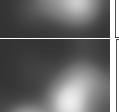
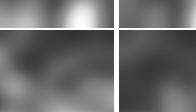
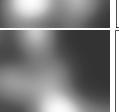
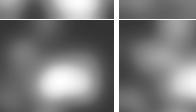
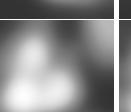
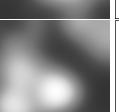
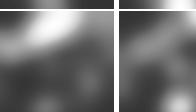
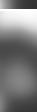
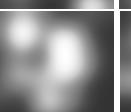
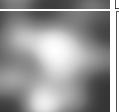
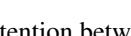
Image	Ours_Full	Ours_SS	Question / Answer	Image	Ours_Full	Ours_SS	Question / Answer
			<p><b>Q:</b> What color are the pillows on the couch?            Ours_Full: brown            Ours_SS: blue            Correct Answer: brown</p>				<p><b>Q:</b> What color is the boy's shirt?            Ours_Full: blue            Ours_SS: green            Correct Answer: blue</p>
			<p><b>Q:</b> What color is her hair?            Ours_Full: black            Ours_SS: blonde            Correct Answer: black</p>				<p><b>Q:</b> How many umbrellas are open?            Ours_Full: 1            Ours_SS: 3            Correct Answer: 1</p>
			<p><b>Q:</b> Is this a cloudy day?            Ours_Full: no            Ours_SS: yes            Correct Answer: no</p>				<p><b>Q:</b> What gender is the person standing up?            Ours_Full: female            Ours_SS: male            Correct Answer: female</p>
			<p><b>Q:</b> What is he waiting to do?            Ours_Full: hit ball            Ours_SS: skateboard            Correct Answer: hit tennis ball</p>				<p><b>Q:</b> What color is the woman's dress?            Ours_Full: white            Ours_SS: yellow            Correct Answer: white</p>
			<p><b>Q:</b> What language do locals speak?            Ours_Full: german            Ours_SS: french            Correct Answer: french</p>				<p><b>Q:</b> What is the girl walking with?            Ours_Full: dog            Ours_SS: suitcase            Correct Answer: dog</p>
			<p><b>Q:</b> How many sides does the red sign have?            Ours_Full: 8            Ours_SS: 4            Correct Answer: 8</p>				<p><b>Q:</b> Which sign is the man showing?            Ours_Full: peace            Ours_SS: stop            Correct Answer: peace</p>
			<p><b>Q:</b> Which color of the umbrella is the woman holding?            Ours_Full: purple            Ours_SS: pink            Correct Answer: pink</p>				<p><b>Q:</b> What video game system is the kid playing?            Ours_Full: wii            Ours_SS: wii            Correct Answer: wii</p>
			<p><b>Q:</b> What is the guy standing on?            Ours_Full: elephant            Ours_SS: surfboard            Correct Answer: elephant</p>				<p><b>Q:</b> What are the chopped white pieces?            Ours_Full: onions            Ours_SS: mustard            Correct Answer: onions</p>
			<p><b>Q:</b> What has fallen to the ground?            Ours_Full: leaves            Ours_SS: bench            Correct Answer: leaves</p>				<p><b>Q:</b> What in the picture can a person ride in?            Ours_Full: boat            Ours_SS: water            Correct Answer: canoe</p>
			<p><b>Q:</b> What season does this appear to be?            Ours_Full: fall            Ours_SS: summer            Correct Answer: fall</p>				<p><b>Q:</b> What is the yellow substance called?            Ours_Full: banana            Ours_SS: orange            Correct Answer: banana</p>
			<p><b>Q:</b> What it is, water or ice?            Ours_Full: water            Ours_SS: snow            Correct Answer: water</p>				<p><b>Q:</b> What shape is the bathtub?            Ours_Full: oval            Ours_SS: square            Correct Answer: oval</p>
			<p><b>Q:</b> What color are the letters on the sign?            Ours_Full: white            Ours_SS: red            Correct Answer: white</p>				<p><b>Q:</b> Are the person's veins visible?            Ours_Full: no            Ours_SS: yes            Correct Answer: yes</p>
			<p><b>Q:</b> What shape are the flags?            Ours_Full: triangle            Ours_SS: rectangle            Correct Answer: triangle</p>				<p><b>Q:</b> What model of Ford truck is this?            Ours_Full: ford            Ours_SS: ford            Correct Answer: f-150</p>

Figure 6: Qualitative comparisons of attention between Ours\_Full and Ours\_SS.