Multimodal Dual Attention Networks for 2019 DramaQA Challenge

Seri Lee
Seoul National University
sally20921@snu.ac.kr

Abstract

DramaQA challenge addresses the task of solving a natural language question answering (QA) that targets the story of a video clip. DramaQA dataset utilizes image frames, subtitles to answer the question. This paper proposes multimodal dual attention networks (MDAN), which exploits dual attention mechanism with with late fusion for each task. Experimental results on DramaQA dataset show that the proposed method performs significantly better than the baseline method that was introduced as a starter code.

1. Introduction

Question Answering (QA) on a video based on multimodal content input is an emerging topic in artificial intelligence. Video QA is more challenging than image QA for the following reasons: (1) Video QA involves multimodal content aligned on time-series. The model must learn the joint representations among at least two multimodal contents and given questions. (2) Video QA requires to extract high-level meaning form multimodal contents such as scene frames and captions which are highly complex and ambiguous information for the task. [2]

The recently introduced DramaQA aims to solve a natural language QA inquiring the story of a video clip. This paper proposes multimodal dual attention networks (MDAN) that was inspired by multimodal dual attention memory (MDAM) [2]. Although the architecture of the proposed model is very similar with the original MDAM model, important details such as preprocessing module, model parameters, hyper-parameters, loss function and optimizers vary in order to fit the DramaQA dataset. The learning process consists of four sub-modules, preprocessing, self-attention, attention by question, and multimodal fusion.

2. Background: Video QA Datasets

While several previous studies have suggested datasets for the Video QA task, they did not consider story-level understanding, resulting in a lack of variance in the degree of question difficulty and coherent story-centric description [9].

DramaQA is a large-scale video QA dataset based on a Korean popular TV show, “Another Miss Oh”. This dataset contains four levels of QA on difficulty and multi-level character-centered story descriptions. DramaQA dataset aims to evaluate human level video story understanding. Although multi-level character-centered story descriptions are provided, they are not used to train the proposed model. [9]

Instead, DramaQA utilizes image frames, subtitles of the video clip to answer the question. Two terms are introduced for image frames, shot and scene. A shot is formed by a series of continuous frames with consistent background setting, and a scene is a group of semantically related shots, which are coherent to certain subject or theme. The question is defined by four levels of difficulty in terms of two criteria: length of video clip and the number of logical reasoning step. [10]

3. Multimodal Dual Attention Networks

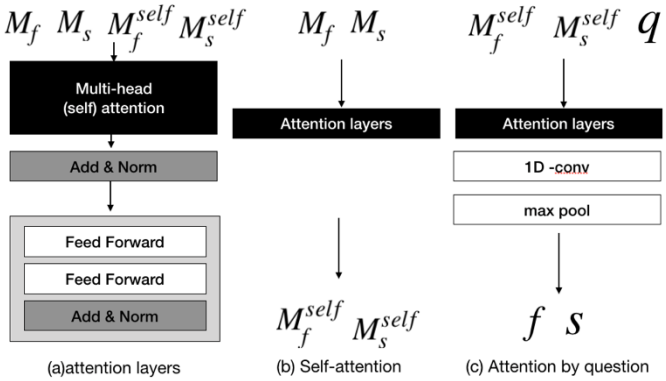
The goal of this paper is to build a video QA model that maximizes information needed for QA through attention mechanisms and fuses the multimodal information at a high-level of abstraction. This problem is solved by introducing two attention layers, which leverage the multi-head attention functions [5], followed by the residual learning of multimodal fusion.

3.1 Preprocessing

The input of the model is composed of a sequence of image frames, a sequence of subtitles, a question, and a set of five candidate answer sentences. The main purpose of the preprocessing module is to transform the raw input as tensor formats.

Linguistic Inputs Subtitles, question, and answers are first tokenized into words using *nltk.word_tokenize* [11]. Each word is further transformed into a vector using the *300-D GloVe word embeddings* [4] pre-trained on a large-scale corpus. The word embeddings are then passed through a 2-layer GRU network [15]. To obtain 256-D sentence-level tensor representations, shared 1-D convolution layers and max pooling operations are applied to the word-level tensor representations.

Visual Inputs Images are first resized to 224×224 to fit the ResNet50 [3] input size. The last layer of ResNet50 [3] is used to extract features which produces 512-D sized activation output that represents frame. Feature cache is stored after feature extraction for faster data-loading [9]. All images for a scene question are concatenated in a temporal order.



[Figure 1] Illustration of how attention module works.

3.2 Self-attention

This module imports the frame tensor M_f and subtitle tensor M_s as input. The output is the frame tensor M_f^{self} and subtitle tensor M_s^{self} that have latent values of the input by using attention layers [5]. The module provides separate attention to frames and subtitles. Attention layers consist of two sub layers: 1) multi-head self-attention networks and 2) point-wise fully connected feed forward networks [2]. There are a residual connection and layer normalization between each sub-layer.

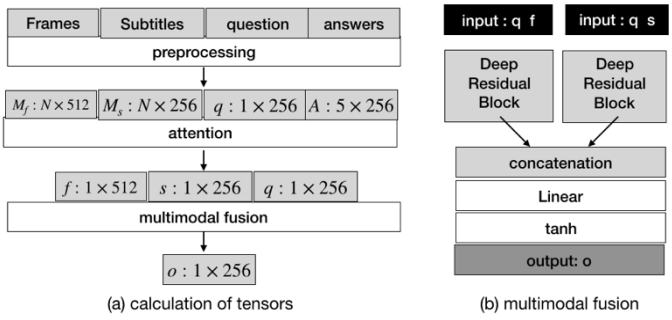
Multi-head Self-Attention Networks Multi-head attention introduced in [5] consists of h paralleled 'heads'. Each head corresponds to an independent scaled dot-product attention function. In this work, we use $h = 4$.

Feed Forward Networks Fully-connected feed forward networks apply two linear transformations and a ReLU activation function [12] separately and identically for every point of the input [2]. Where x is a point of the input, feed forward networks can be represented as the following equation:

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b2$$

3.3 Attention by Question

The module takes M_f^{self}, M_s^{self} as input and calculate the attention scores separately, this time using the question. Then using 1-D convolutional neural networks, output $f \in \mathbb{R}^{256}, s \in \mathbb{R}^{256}$ are produced [2]. Attention layers in this module differ from that of the self-attention module in that they have multi-head attention networks inside. The question tensor q is used as a pivot.



[Figure 2] Illustration of tensor size throughout the pipeline and multimodal fusion module.

3.4 Multimodal Fusion

Multimodal residual learning [6] is used for fusion. Taking $f \in \mathbb{R}^{256}$, $s \in \mathbb{R}^{256}$, $q \in \mathbb{R}^{256}$ as input, the final output is the concatenation of the two deep residual blocks denoted as $o \in \mathbb{R}^{256}$.

4. Experimental Results

4.1 Dataset

DramaQA Datasets consist of 9,649 video clips including 317 scenes and 9,332 shots. There are 15,760 question-answer pairs in total, each level containing 7991, 4116, 1833, 1821 questions starting from level 1. Random splits are used for train, test, and validation data (8:1:1) respectively.

4.2 Evaluation Criteria

Evaluation metrics are defined as follows: (1) Shot-level QA performance on the DramaQA dataset. The questions for this level are based on a video length less than about 10 seconds without change of

viewpoint. (2) Scene-level QA performance on the DramaQA dataset. The set of questions for this level is based on clips that are 1-3 minutes long without place change. (3) Overall performance on the entire DramaQA dataset. [10]

4.3 Experimental Setup

Hyper-parameters The batch size is 12 and the number of epochs is fixed to 30. Adagrad [14] is used for optimization and dropouts [13] are used for regularization.

Baselines To compare the performance of each component, experiments with the following models are conducted. 1) baseline model: 2-layer single-directional encoder-decoder GRU model with linear layer fusion. 2)variant model: 2-layer bi-directional encoder-decoder GRU model with residual block fusion.

4.4 Results

DramaQA Challenge The DramaQA Challenge provides a separate evaluation server for the test set so the participants can evaluate the performance of their models using the server. Table 1 shows the performance comparison with the baseline and variant model. The proposed MDAN model achieves 57% and shows the performance gain of 8% compared to the baseline model, which achieves 49%.

Method	Test
GRU + linear fusion	0.49
Bidirectional GRU + MLP	0.50
MDAN	0.57

[Table 1] Performance comparison with variant models using the DramaQA Challenge evaluation server.

5. Conclusions

This paper proposes MDAN for solving task in DramaQA. Multimodal dual attention networks provide dual attention structure that captures a high-level abstraction of the full video content [2]. The proposed method has been demonstrated to be valid by showing a significant increase in performance on DramaQA datasets compared to baseline or variant model.

References

[1] Yu, Zhou, et al. "Deep Modular Co-Attention Networks for Visual Question Answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[2] Kim, Kyung-Min, et al. "Multimodal dual attention memory for video story question answering." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[3] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[4] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

[5] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[6] Kim, Jin-Hwa, et al. "Multimodal residual learning for visual qa." *Advances in neural information processing systems*. 2016.

[7] Kim, Kyung-Min, et al. "Deepstory: Video story qa by deep embedded memory networks." *arXiv preprint arXiv:1707.00836*(2017).

[8] Na, Seil, et al. "A read-write memory network for movie story understanding." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[9] DramaQA, https://dramaqa.github.io, 2019.

[10] Heo, Yu-Jung, et al. "Constructing Hierarchical Q&A Datasets for Video Story Understanding." *arXiv preprint arXiv:1904.00623*(2019).

[11] Loper, Edward, and Steven Bird. "NLTK: the natural language toolkit." *arXiv preprint cs/0205028* (2002).

[12] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010.

[13] Hinton, Geoffrey E., et al. "Improving neural networks by preventing co-adaptation of feature detectors." *arXiv preprint arXiv:1207.0580* (2012).

[14] Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of Machine Learning Research* 12.Jul (2011): 2121-2159.

[15] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).