Multimodal Dual Attention Networks for 2019 DramaQA Challenge

이세리(서울대학교 컴퓨터공학부)
sally20921@snu.ac.kr

DramaQA dataset for Video Story Understanding

- four levels of questions in the degree of difficulty to consider story level understanding for Video QA task
- descriptions are not used to train the model
- utilize image frames, subtitles of the video clip to answer the question
- for more information, https://dramaqa.github.io/Dataset

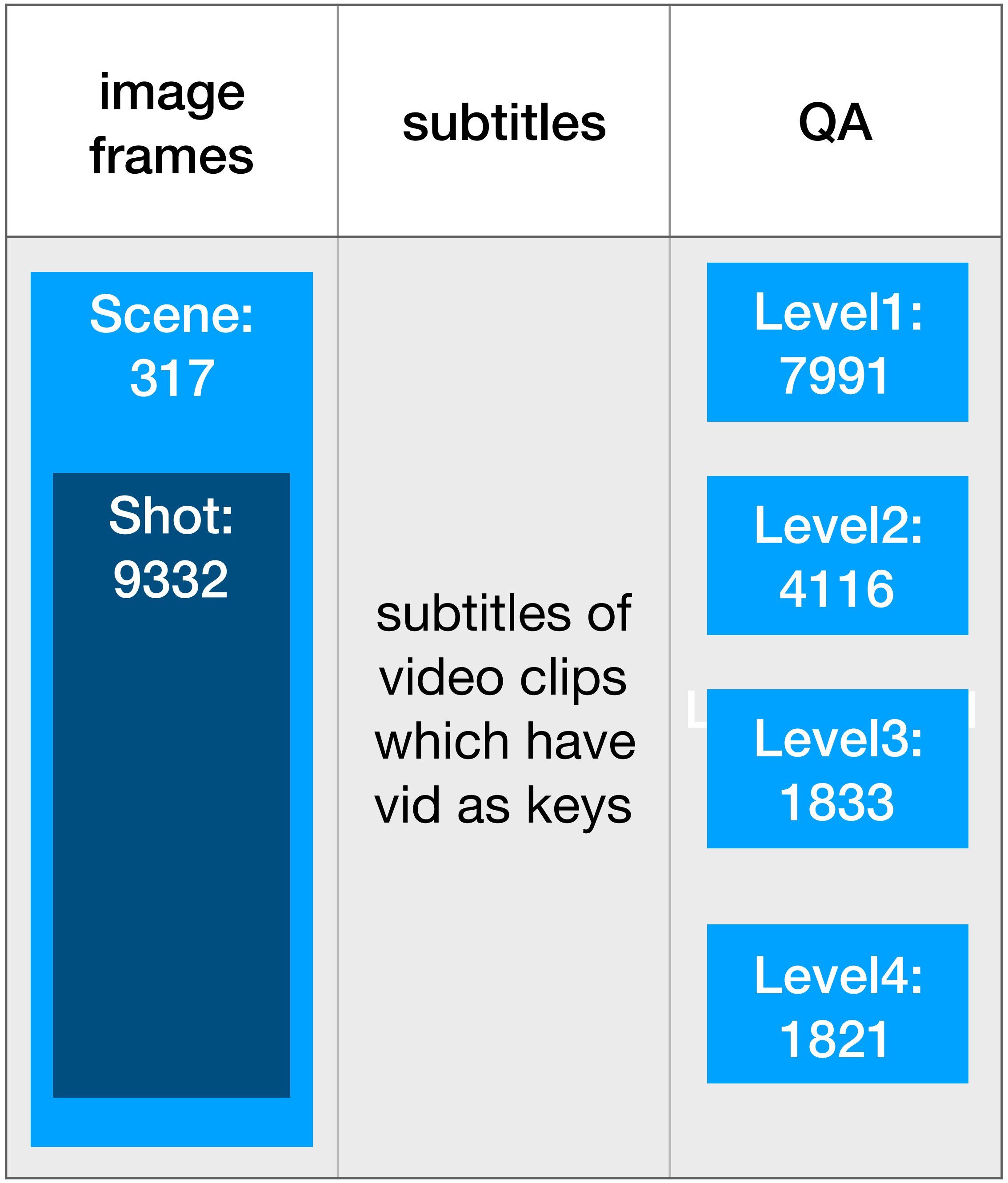
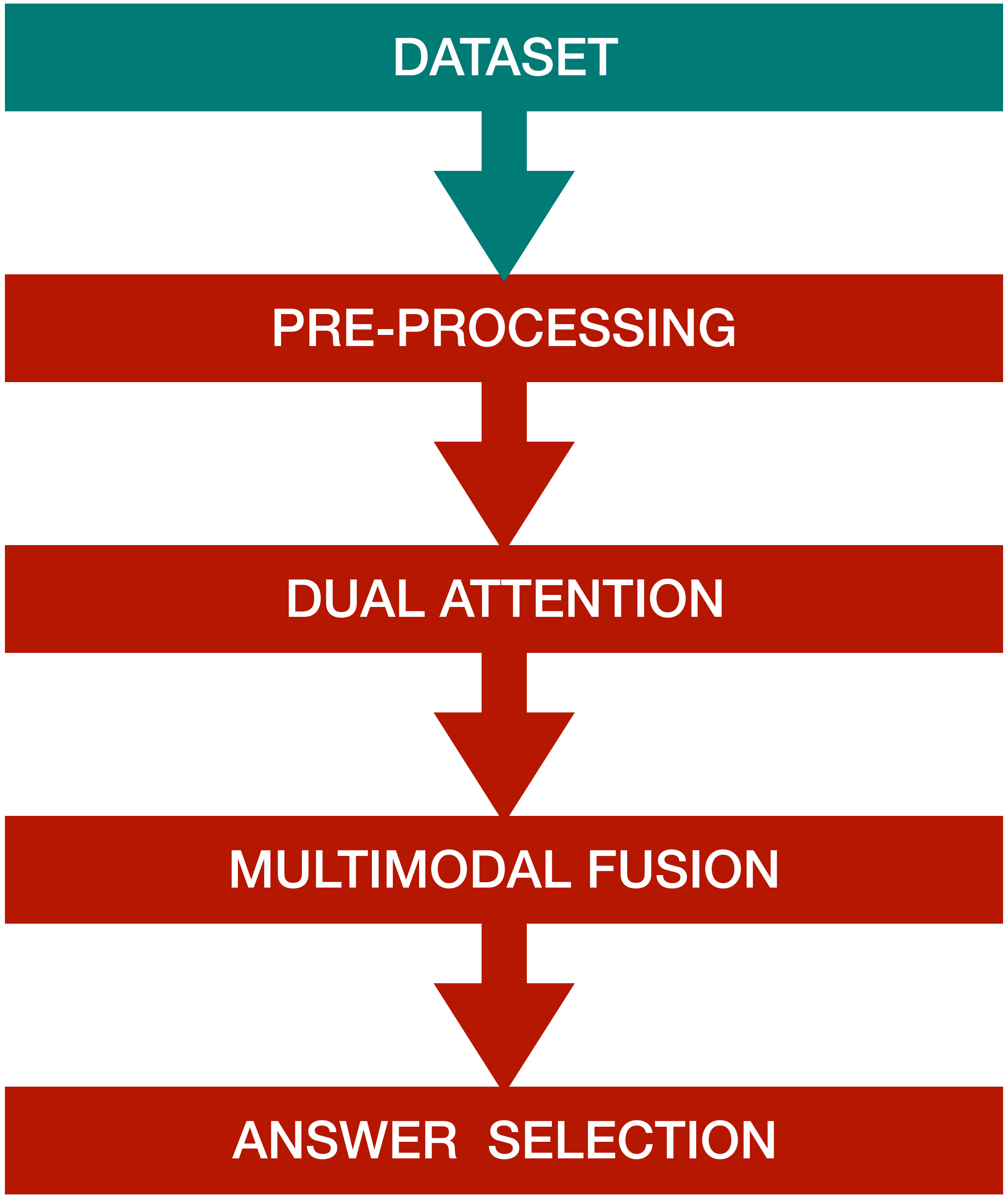


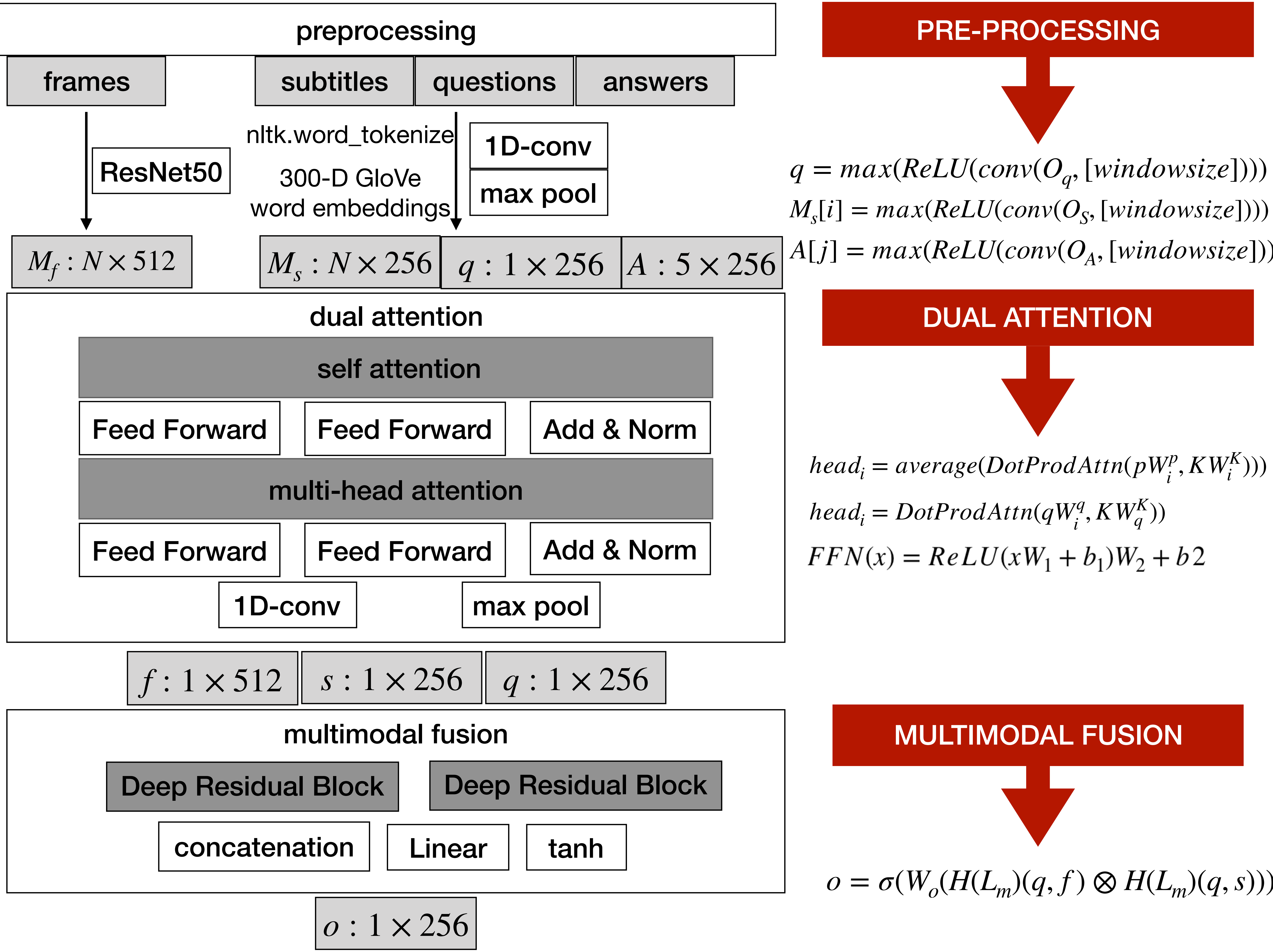
image frames	subtitles	QA
Scene: 317 Shot: 9332	subtitles of video clips which have vid as keys	Level1: 7991 Level2: 4116 Level3: 1833 Level4: 1821

Multimodal Dual Attention Networks

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

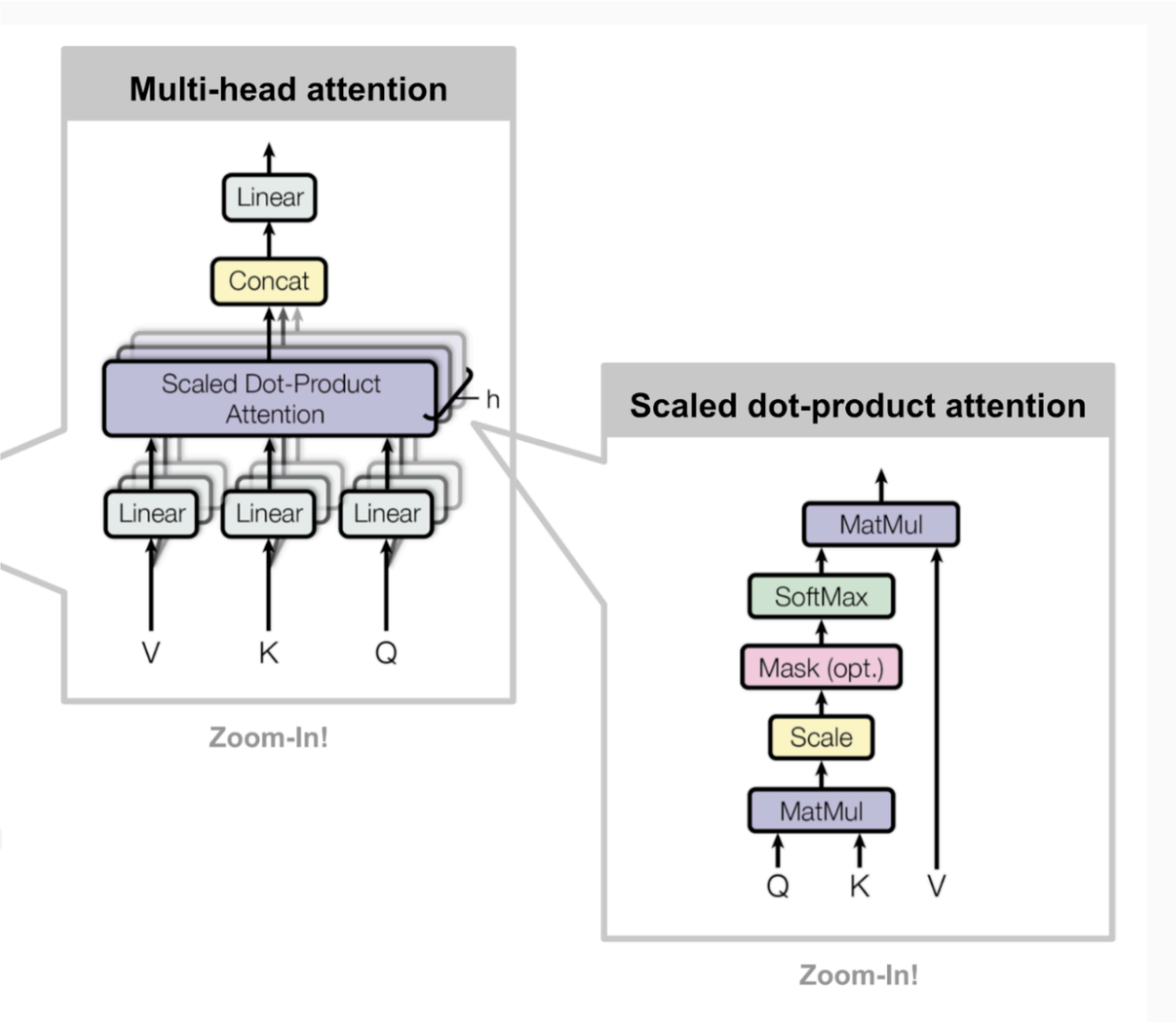
- reference code https://github.com/gicheonkang/DAN-VisDial




preprocessing
frames
subtitles
questions
answers
ResNet50
nltk.word_tokenize
300-D GloVe word embeddings
1D-conv
max pool
$M_f : N \times 512$
$M_s : N \times 256$
$q : 1 \times 256$
$A : 5 \times 256$
dual attention
self attention
Feed Forward
Feed Forward
Add & Norm
multi-head attention
Feed Forward
Feed Forward
Add & Norm
1D-conv
max pool
$f : 1 \times 512$
$s : 1 \times 256$
$q : 1 \times 256$
multimodal fusion
Deep Residual Block
Deep Residual Block
concatenation
Linear
tanh
$o : 1 \times 256$
PRE-PROCESSING
$q = max(ReLU(conv(O_q, [windowsize])))$
$M_s[i] = max(ReLU(conv(O_S, [windowsize])))$
$A[j] = max(ReLU(conv(O_A, [windowsize]))$
DUAL ATTENTION
$head_i = average(DotProdAttn(pW_i^p, KW_i^K)))$
$head_i = DotProdAttn(qW_i^q, KW_q^K))$
$FFN(x) = ReLU(xW_1 + b_1)W_2 + b2$
MULTIMODAL FUSION
$o = \sigma(W_o(H(L_m)(q, f) \otimes H(L_m)(q, s)))$

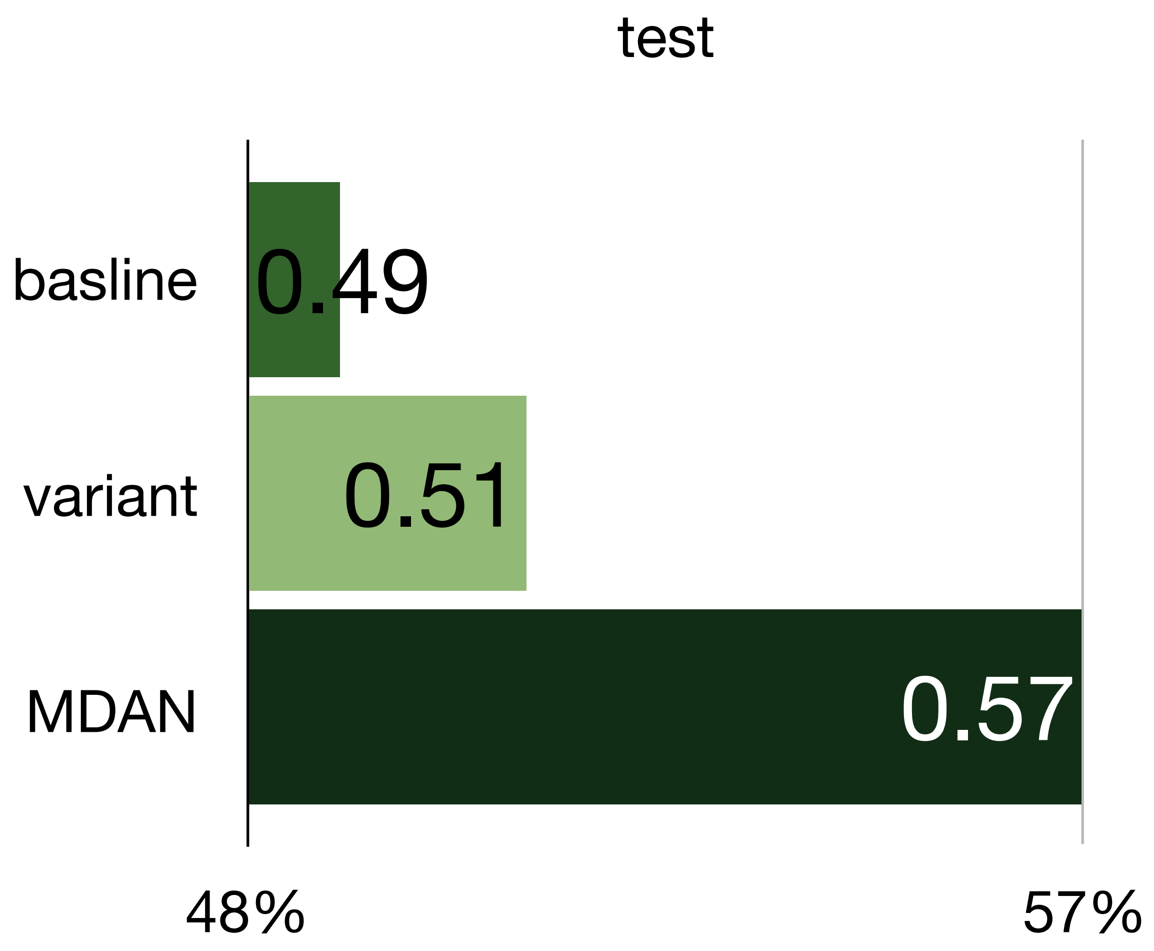
Dual Attention?

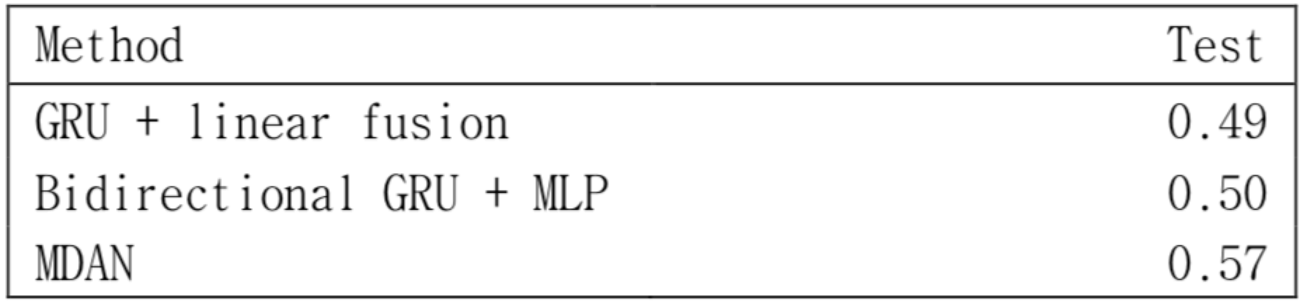
- self-attention module
- multi-head attention module
- key set(frames or subtitles) K, using pivot p, update to $\hat{K}$
- for self-attention, $p \in K$
- for multi-head attention module, $p = q(question)$
- from $head_1$ to $head_h$, $h = 4$ is used for implementation



Experiments

- batch size: 12
- number of epochs: 20
- optimization: Adagrad
- regularization: dropout 0.5
- 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
- 3) Multimodal Dual Attention Network





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

Any Questions?

for more information, please refer to
https://github.com/sally20921/MDANforDramaQA2019