Similar Question Detection in *Community Question Answering*

Problem Statement

To detect similar questions in a given ***Community Question Answering Corpus*** by leveraging ***semantic similarity*** between ***Question-Answer*** and ***Question-Question*** pairs.

Architecture

Model Architecture

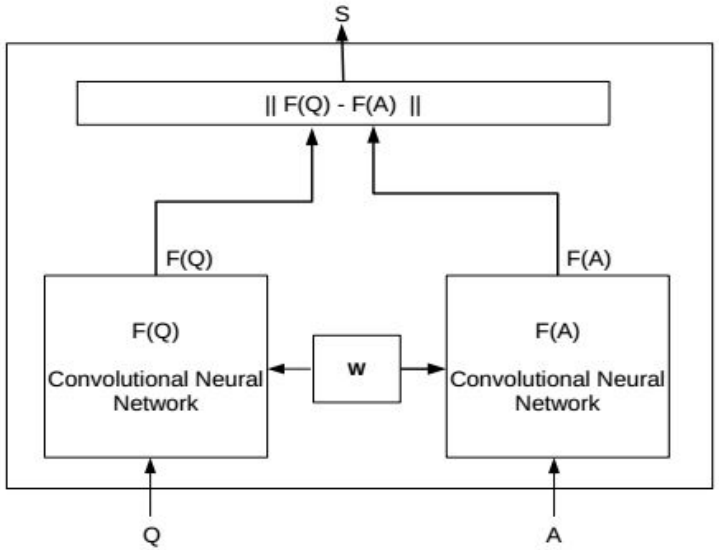

Figure 1: Architecture of Siamese network.

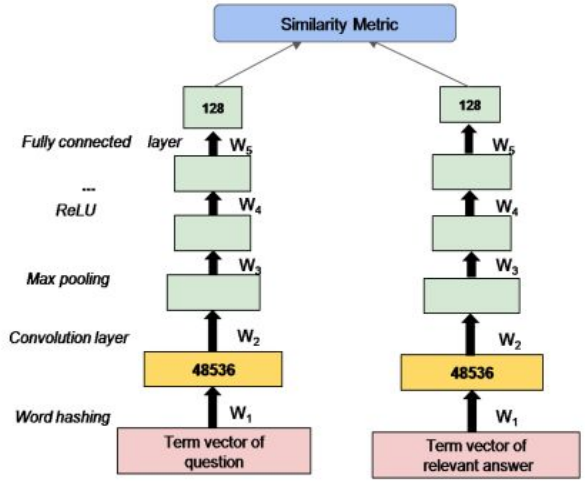

Figure 2: Architecture of SCQA. The network consists of repeating convolution, max pooling and ReLU layers and a fully connected layer. Also the weights W_1 to W_5 are shared between the sub-networks.

Siamese Convolutional Neural Network for Community Question Answering (SCQA)

- SCQA consists of a pair of ***deep convolutional neural networks (CNN)*** with convolution, max pooling and rectified linear (ReLU) layers and a fully connected layer at the top.
- CNN gives a ***non linear projection of the question and answer term vectors*** in the semantic space.
- Distance metric used in ***Normalised Cosine similarity***
- ***Contrastive loss function*** is used to compare both the representations and combines the distance measure and the label.

Continued..

- The gradient of the loss function with respect to the weights and biases shared by the sub-networks, is computed using back-propagation.
- ***Stochastic Gradient Descent*** method is used to update the parameters of the sub-networks.
- ***Sharing of weights*** occurs between the two CNN's, which helps the network learn semantic correlations between the phrases in the pairs

Datasets Used

- Quora Question - Answer Dataset
  - Dataset consists of over 400,000 lines of potential question duplicate pairs. Each line contains IDs for each question in the pair, the full text for each question, and a labels (determining similar or not).
- Labeled Data for QA Retrieval - Zhang
  - Standard dataset provided by Zhang (et. al) - around 24,000 question question pairs and binary labels.
- L5 - Yahoo! Answers Manner Questions
  - Dataset is a small subset of the questions, selected for their linguistic properties (for example, they all start with "how {to|do|did|does|can|would|could|should}"). Additionally, only questions and answers that have at least four words are kept , out of which at least one is a noun and at least one is a verb. The final set contains 142,627 questions and their answers along with a small amount of metadata.
- L6 - Yahoo! Answers Comprehensive Questions and Answers
  - Dataset contains 4,483,032 questions and their answers. In addition it contains a small amount of metadata, i.e., which answer was selected as the best answer, and the category and sub-category that was assigned to this question.

Data preprocessing

- Tokenization
- Case Folding
- Filtering
- Building Tri-character vocabulary
- Generating Question - Answer vectors
- Generating Negative samples
- Blocking and Packing

Observations and Results

Build - 1 (POC)

- Implemented SCQA in Keras
- Dataset used : L6 - Yahoo! Answers Comprehensive Questions and Answers
- Data size = 3120 (+) + 15595 (-) = 18714
- Vocabulary size = 56K
- Training - Test split = 0.75 : 0.25

Dataset	Data Size (Ques)	Learning Rate	Batch Size	f1 score
Yahoo!	3120	0.01	100	0.7741
			50	0.7986

Intermediate Builds

- **Build - 2**
- Training dataset : L6 - Yahoo! Answers Comprehensive Questions and Answers
  - Data size = 3119 (+) + 15595 (-) = 18714
- Test dataset : Labeled Data for QA Retrieval - Zhang
  - Data size = 1000 random pairs
- Accuracy = 0.48

- **Build - 3**
- Training dataset : L5 - Yahoo! Answers Manner Questions
  - Data size = 10000 (+) + 50000 (-) = 60000
- Test dataset : Labeled Data for QA Retrieval - Zhang
  - Data size = 1000 random pairs
- Vocabulary size = 12817
- Accuracy = 0.52

Build - 4 (Final)

- Training datasets : L5 - Yahoo! Answers Manner Questions and Quora Question - Answer Dataset
- QQ Data size = 5000
- QA Data size = 5000 (+) + 25000(-)
- Vocabulary = 13810

Expt 1 : Varying Epochs		
Runs	**Epochs**	**Accuracy**
1	10	0.62516
2	20	0.63348
3	50	0.63856

Expt 2 : Varying learning rate		
Runs	**LR**	**Accuracy**
1	0.01	0.6252
2	0.05	0.6404
3	0.08	0.6323
4	0.001	0.6316
5	0.000374	0.6292
6	0.0003	0.6264
7	0.00001	0.6220

References

- Arpita Das, Harish Yenala , Manoj Kumar Chinnakotla, Manish Shrivastava: Together we stand: Siamese Networks for Similar Question Retrieval. ACL (1) 2016
- Kai Zhang, Wei Wu, Fang Wang, Ming Zhou, Zhoujun Li: Learning Distributed Representations of Data in Community Question Answering for Question Retrieval. WSDM 2016: 533-542
- Arpita Das, Manish Shrivastava, Manoj Kumar Chinnakotla: Mirror on the Wall: Finding Similar Questions with Deep Structured Topic Modeling. PAKDD (2) 2016: 454-465

Thanks!

Team : 7

Members:

Adithya Avvaru (20162116)
Anupam Pandey (20162118)
Damodar Dukle (20162055)

Mentor :

Sriharsh Bhyravajjula