Report on Similar Question Retrieval

BY:
Varun Mundale (20162011)
Saad Khan (20162043)

Introduction:

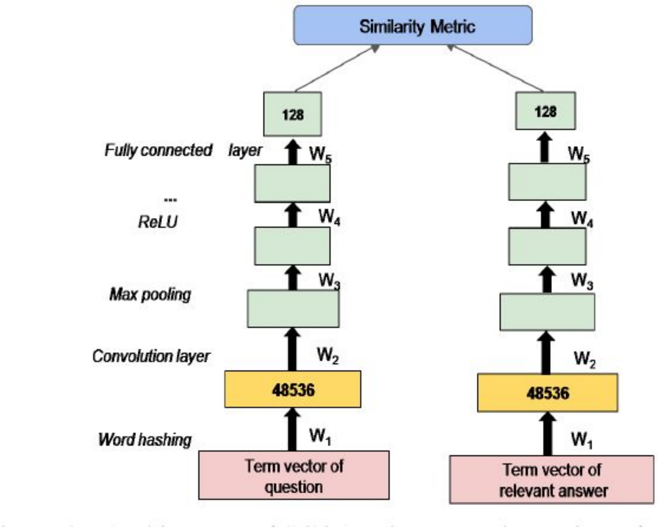
Community Question Answering services like Yahoo! Answers , Baidu Zhidao , Quora , StackOverflow etc. provide a platform for interaction with experts and help users to obtain precise and accurate answers to their questions. The time lag between the user posting a question and receiving its answer could be reduced by retrieving similar historic questions from the archives.

MODEL1: Siamese Convolutional Neural Network

(Reference: Together We Stand:Siamese Networks for Similar Question Retrieval)

Architecture

The semantic vectors yielded for question and answer respectively are connected to a layer that measures distance or similarity between them.The contrastive loss function combines the distance measure and the label. The gradient of the loss function with respect to the weights and biases shared by the sub-networks, is computed using back-propagation.



Feature Extraction

Word Hashing

Word hashing involves letter n-gram to reduce the dimensionality of term vectors. For a word, say, “table” represented as (#table#) where # is used as delimiter, letter 3-grams would be #ta, tab, abl, ble, le#.Thus word hashing is character level representation of documents which takes in SCQA question and relevant answer pairs are care of OOV words and words with minor spelling fed to train the network. It represents a query using a lower dimensional vector with dimension equal to number of unique letter trigrams in the training dataset.

MODEL2: Siamese Residual Neural Network

- **Architecture**

A 3 layer residual network is used as the base network. Batch Normalization is done per layer. The measure of instance similarity used is Euclidean distance (gives better results than cosine similarity) . Batch Normalization improves the performance considerably. The intuition behind using batch normalization is that the final feature vectors are normalized and Euclidean distance is able to perform better in the normalized space.

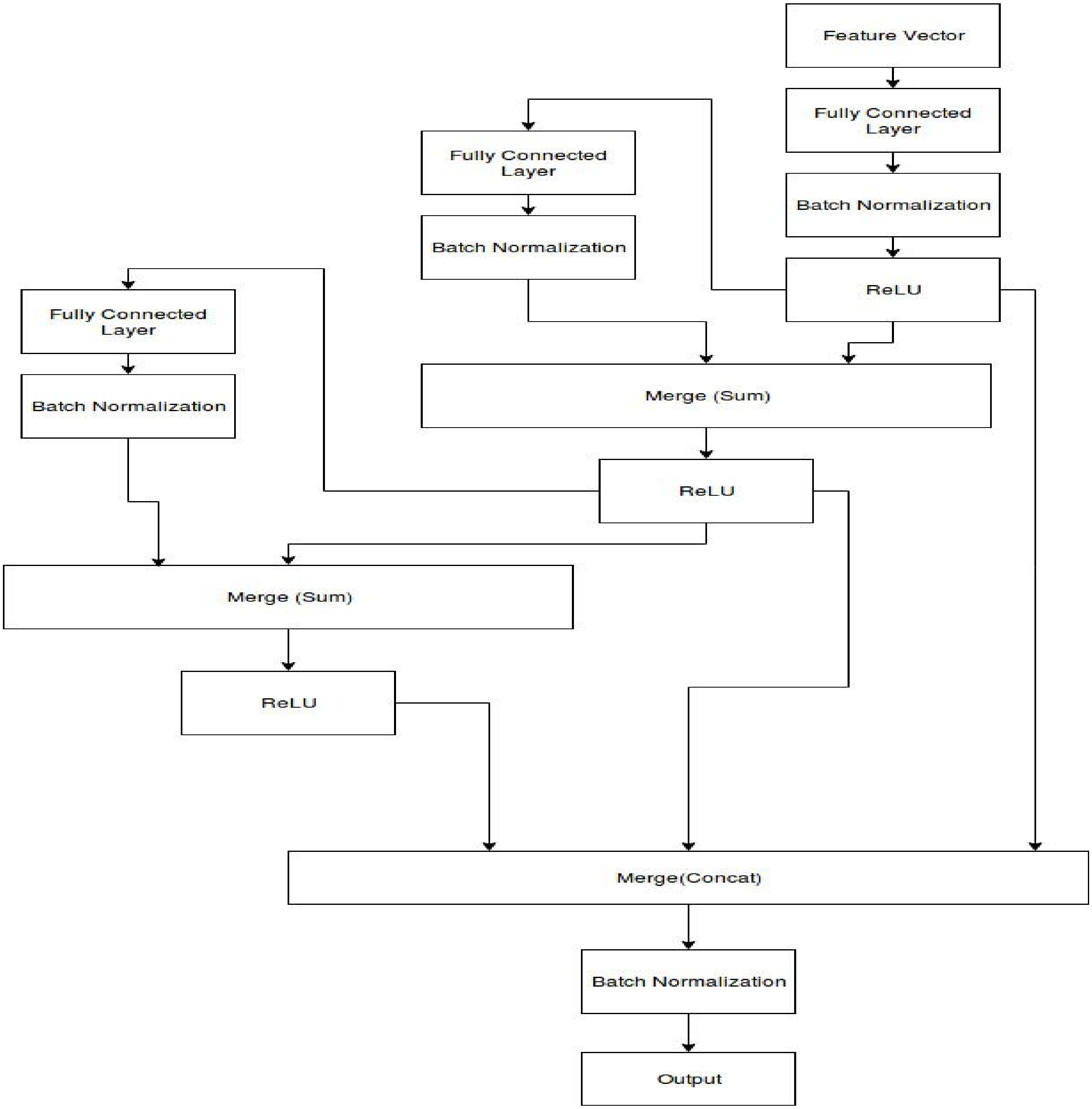


Fig: Base Residual Network

- ## Feature Extraction

GLOVE model was used for generating feature vectors.
It generates 300 dimensional vectors for words. Word vectors of all the words in a sentence are summed and average is taken to get a sentence vector.

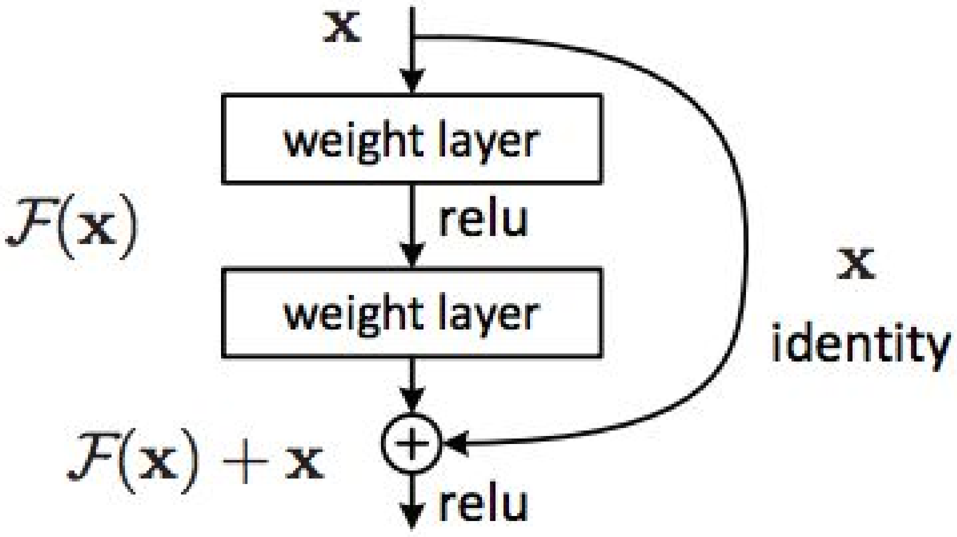
- ## Motivation

Network depth is of crucial importance in neural network architectures, but deeper networks are more difficult to train. The residual learning framework eases the training of these networks, and enables them to be substantially deeper — leading to improved performance in both visual and non-visual tasks. These residual networks are much deeper than their 'plain' counterparts, yet they require a similar number of parameters (weights).

The (degradation) problem:
With network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly. Unexpectedly, such degradation is not caused by overfitting, and adding more layers to a suitably deep model leads to higher training error.

The core insight:
Let us consider a shallower architecture and its deeper counterpart that adds more layers onto it. There exists a solution to the deeper model by construction: the layers are copied from the learned shallower model, and the added layers are identity mapping. The existence of this constructed solution indicates that a deeper model should produce no higher training error than its shallower counterpart.

The proposed solution:


x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Instead of hoping each stack of layers directly fits a desired underlying mapping, we explicitly let these layers fit a residual mapping. The original mapping is recast into F(x)+x. We hypothesize that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. To the extreme, if an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers.

We have reformulated the fundamental building block (figure above) of our network under the assumption that the optimal function a block is trying to model is closer to an identity mapping than to a zero mapping, and that it should be easier to find the perturbations with reference to an identity mapping than to a zero mapping. This simplifies the optimization of our network at almost no cost. Subsequent blocks in our network are thus responsible for fine-tuning the output of a previous block, instead of having to generate the desired output from scratch.

DATASETS USED

Training

- Yahoo! Answers Manner Questions (142627 Question - Answer pairs)
- Quora Question pairs (404301 Question pairs)
- Physics - StackExchange (46773 Question - Answer pairs)

Testing

- Annotated survey dataset by Zhang et al. (24642 Question pairs)

OBSERVATIONS & RESULTS

Addition of Quora dataset for training

We found that adding a small percentage of samples from the Quora dataset improved accuracy. Since the Quora data contains Q-Q data, it helps the network in learning textual similarity.

Siamese Convolutional Neural Network

- **Architecture** : Twin networks of 2 times convolution, max pooling and relu activation.
- **Input**: Word hashed term vectors of dimension = 42110
- **Distance metric** : Cosine distance
- **Loss**: Contrastive loss
- **Optimiser**: SGD
- Training with a Q-Q dataset and Q-A dataset gives better test accuracy than only training with Q-A dataset
- **Results**:
  - Quora data: 64.83
  - Physics Stack Exchange data: 67.17
  - Quora + Physics Stack Exchange data: 68.08
  - Yahoo webscope: 67.45
  - Quora + Yahoo Webscope : 68.18

Siamese Residual Neural Network

- **Architecture**: Twin networks of 3 layers of residual network with batch normalisation
- **Input**: Word2vec embeddings normalized by tf-idf weights of dimension = 300
- **Distance metric**: Euclidean distance
- **Loss**: Contrastive loss
- **Optimizer**: Adam
- Training with a Q-Q dataset and Q-A dataset gives better test accuracy than only training with Q-A dataset
- **Results (Accuracy)**:
  - Quora data : 66.25
  - Physics Stack Exchange data : 67.32
  - Quora + Physics Stack Exchange data: 68.16
  - Yahoo webscope : 66.13
  - Quora + Yahoo Webscope: 69.46

Analysis of Retrieval system

- **Siamese Convolutional Neural Network (with Quora data added)**

Query	**Top Results**
How to check your voicemail from another phone?	1. How can i check my voicemail from another phone? 2. How do u check your voicemail from another phone? 3. Is there a way to check my voicemail from another phone?
what files can i play on my xbox 360	1. Can my xbox 360 play avi files? 2. What movie file types can XBOX 360 play? 3. What type of video files can you play on a xbox 360?
how long should a synopsis be	1. When a producer asks for a synopsis, how long should it be? 2. How long is a brief synopsis? 3. How long should a screenplay synopsis be?
What GCSE and A Level grades do you need to become a doctor?	1. GCSE grades needed to become a doctor? 2. What GCSE's do I need to become a doctor? 3. What GCSE/A level grades do you need to become a Vet?

- **Siamese Residual Neural Network**

Query	**Top Results**
How to check your voicemail from another phone?	1. How can i check my voicemail from another phone? 2. How do you check cell phone voicemail from another phone? 3. Is there a way to check my voicemail from another phone?
what files can i play on my xbox 360	1. What type of video files does my xbox 360 play? 2. Can my xbox 360 play avi files? 3. Can you play a Xbox 360 game on an regular Xbox?
how long should a synopsis be	1. How long should a film review synopsis be? 2. How long is a brief synopsis? 3. How long should a screenplay synopsis be?
What GCSE and A Level grades do you need to become a doctor?	1. What GCSE grades do you need to be a doctor? 2. at GCSE Subjects do I need to do, to become a doctor? (UK) 3. GCSE grades needed to become a doctor?

SUGGESTIONS FOR FUTURE WORK

- The ratio of Questions and Answers used for training is currently 1:1 . This can be changed to include more than one answer per question i.e. a Q:A ratio of 1:2 or 1:3.
    - Doing this should help in capturing broader semantic relationship which in turn will help in matching of questions that have very little textual similarity.
    - On the other hand adding too many answers would reduce the network's textual matching capability.
- In this model we have, used <question,best answer> to form relevant pair in the training network. We could form multiple relevant pairs for same question to improve the training data (by using information like user votes, ratings, user reputation etc.)
- RNN and LSTM networks can be used for sentence embeddings instead of word2vec averaging of words.

GithubLink: https://github.com/varunmundale/Similar-Question-Retrieval