Multitask Learning and Extensions of Dynamic Coattention Network

Keven (Kedao) Wang
Computer Science Department
Stanford University
kvw@stanford.edu

Diana (Xinyuan) Huang
hxydiana@stanford.edu

Abstract

Dynamic Coattention Network (DCN) was introduced in late 2016 and achieved state-of-the-art performance on Stanford Question Answering Dataset (SQuAD). In this paper, we re-implement DCN and explore different extensions to DCN, including multi-task learning with Quora question pairs dataset, different loss function that account for distance from truth, variation of sentinel vectors, novel pre-processing trick, modification to coattention encoder architecture, as well as hyperparameter tuning. After joint training, we observe a 2% increase in f1 on Quora dataset. Our conclusion is that multi-task learning benefits the simpler task more than the more complicated task. On CodaLab leaderboard, we achieved Test f1 = 67.282, EM = 56.278.

1 Introduction

In order for a human to perform well on a question answering task, he/she needs to leverage knowledge from a wide range of NLP tasks, such as named entity recognition, co-reference resolution, semantic understanding, context understanding and possibly more. Our question is, whether these related skills can be learned via transfer-learning in a way that improves performance on question-answering tasks. In particular, we experiment with jointly training a network on both the task to detect duplicate questions (Quora) and the question answering task.

SQuAD[1] dataset contains 107,785 question-answer pairs, with each documents of up to 700 words in length, which can pose challenge for a neural network to attend to. The answers' exact texts are found within the document, which is not ideal, because the network might learn to "wing it" by simply learning the likely-correct answers such as named entities without understanding. Nonetheless, the SQuAD dataset contains some non-trivial questions, that would require lexical variation, syntactic variation, and reasoning.

Quora dataset[2] contains 404,302 question pairs, with median question length of 11.00 words, and standard deviation of 6.61 words. Each question pair correspond to a binary label, indicating whether the two questions are semantically the same.

Since both tasks require semantic understanding, our hope is that transfer learning can be constructive to both tasks. In particular, we hope the encoder weights learned represent common knowledge that can be shared between both tasks. Both tasks have two input text sequences, which fits well into Dynamic Coattention Network[3]'s architecture.

Further, we explore slight modifications on the original Dynamic Coattention Network, on both architecture and hyperparameters.

Dataset	# Data Entries	Max # Words	Median # Words	Std Dev # Words
SQuAD (documents)	107,785	766	126	57.00
SQuAD (questions)	107,785	60	11	3.74
Quora (questions)	404,432	272	11	6.61

Table 1: Dataset Statistics

2 Related Work

Dynamic Coattention Network has proven to be a really effective model when used to solve the Stanford question answering set. It is able to achieve 75.9 F1 score and 66.2 EM score on the official test set. In DCN, the coattention encoder attends to both question and document at the same time by fusing both attention contexts. Coattention has been shown to benefit question answering task by other researches as well. Zhang et al.[4] introduced a question-based filtering layer where they obtained weight distribution on the document for each question word. Wang et al.[5] developed a multi-perspective context matching layer to compare contextual embeddings with the question. In general, all these models attempt to capture the non-linear interaction between document and question, and form attention jointly on document and question.

Multi-task learning has shown to be beneficial across NLP tasks of chunking, dependency parsing, semantic relatedness, POS tagging, and textual entailment. Hashimoto's paper[6] shows that constructive transfer learning occurs when interleaving training data from different tasks.

3 Approach

We implement three models, SQuAD, Quora, and a joint model for interleaved training.

We follow DCN paper's description step-by-step for our squad model. The only thing that is unclear in the paper is the number of sentinel variables used, so we explore different possibilities.

We then implement our Quora model by taking advantage of most the DCN model. We keep both document and question encoder and treat the first question in a data entry as document and second question as question. During data pre-processing, for each original data entry, we create two processed entries by switching order of the two questions. That way, we make sure the ordering of questions doesn't factor into our model. Similarly for coattention encoder, we also keep the original DCN implementation. Finally, we replace the dynamic pointing decoder in DCN by a simple classifier as shown in Figure 1.

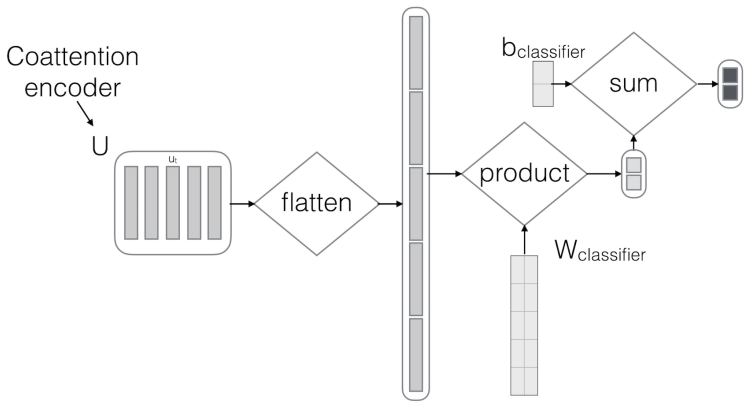

Figure 1: Quora classifier

We implement our joint model by combining the SQuAD model and the Quora model, as illustrated by Figure 2. The encoders' parameters are shared, while the decoders are separate for each dataset.

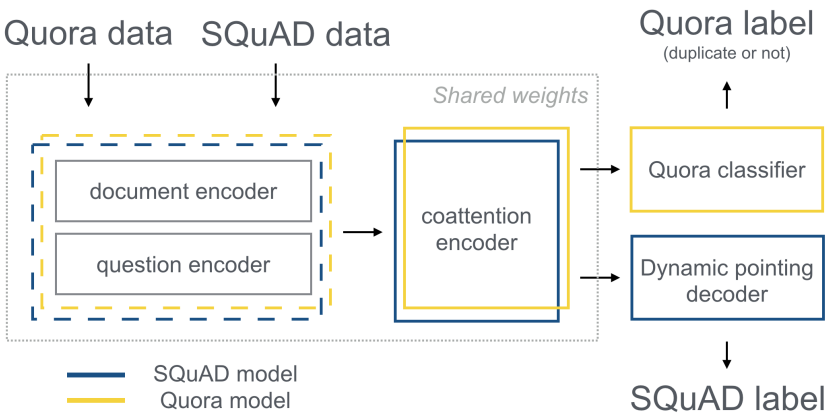

Figure 2: Multi-task learning flow diagram

3.1 Preprocessing

We use GLoVe[7] 840B pre-trained word vectors as word embeddings for our model. Pre-trained word vectors capture the semantic features of words, and reduce the training time it would otherwise take to acquire the semantic knowledge of the dataset. Both datasets are tokenized using nltk word tokenizer and further processed using different approaches. Here we use SQuAD as an example to describe our methods.

When loading word embeddings, we match SQuAD vocabulary against GloVe vocabulary. The out-of-vocabulary rate is a key indicator of the quality of embeddings. Therefore we use two tricks to lower the out-of-vocabulary rate of our embeddings.

3.1.1 Naive processing

We first use native word matching, by apply the following transformations to each word in SQuAD vocabulary: original, capitalized, lower-cased, upper-cased. The out-of-vocabulary rate is high after such processing. After inspection of out-of-vocabulary words, we notice most of them belong to the following categories:

Composite words: “co-inventor”, “higher-than-allowed”
Typo’s: “impressible”, “aeronatical”
Words of foreign language: “*kallī”
Numbers, rare-words, named-entities, words containing special characters

3.1.2 Simple heuristics

Since GLoVe vocabulary consists of English vocabulary, we stripped OOV words of non-ascii characters. In order to target the words that are composite words, we then apply the following transformations in order:

- Strip OOV words of ascii symbols: $/ : . \ <> - * |^{+}, =$, and attempt to match
- Split OOV words by above set of ascii symbols, and attempt to match each sub-word with GloVe vocabulary. Then take mean of the matched word vectors

3.1.3 Spell checker

Since a significant portion of the SQuAD vocabulary are typo’s, we use a simple spell checker by Peter Norvig[8][9] to attempt to correct these mis-spellings. Then we try to retrieve the corrected word embedding from GloVe.

3.2 Tasks

Transfer learning is an important topic among machine learning tasks. Transfer learning means that a higher level representation learned through one task can be applied to a neighboring task. Different NLP tasks are shown to benefit from transfer learning. We are particularly interested in the ability to transfer the knowledge learned from duplication question detection (Quora) to question answering (SQuAD).

3.2.1 Best single model

To establish baselines, we train two separate models, one model on Quora and the other on SQuAD. The Quora model takes two questions as input, and outputs a binary classification on whether the two questions are same or not. The SQuAD model takes a document and a question as input, and outputs two indices, start and end respectively, to specify the span of the question's answer in document.

3.2.2 Quora pre-trained

We pre-train the shared weights (between Quora and SQuAD tasks) on Quora task, and used these to initialize the weights for squad tasks. Weights in the text encoders (question and document) and coattention encoder are shared. The idea is that the encoders will learn representations that is common knowledge required to solve both tasks. One downside of this approach is that, when training squad tasks, we might overwrite the weights and hence lose the knowledge learned while training on quora task.

3.2.3 Jointly trained with Quora

To get around the overwriting weights problem, we also interleave Quora and SQuAD training data. This is done by alternating between Quora and SQuAD training data per minibatch, during which the two optimizers apply back-propagation in turns.

When jointly training, we adjust the batch size for Quora data so that the number of Quora batches are less than or equal to the number of SQuAD batches. This is because the ultimate goal is to train a model that optimizes for SQuAD task, rather than Quora. In an epoch, Quora batches and SQuAD batches are fed into the model alternately until there are only SQuAD batches left, if any. Then the rest of SQuAD batches are fed to the model.

3.3 Loss extensions

In the original DCN model, loss is calculated as the mean of softmax cross entropy of the start and end points across all iterations in the dynamic pointing decoder. As we investigate predictions from the DCN model, we see invalid spans where the predicted start index is greater than the end index. We also find examples where the predicted answer span has no overlap with the correct span at all, and moreover, the predicted span is quite far from the correct one. Therefore, we develop two different approaches to apply weights onto the vanilla softmax cross entropy. Note that in naive softmax cross entropy, the weights is all 1s.

3.3.1 Penalty on invalid spans

It is never right for the model to predict a start index that's bigger than the end index. Hence, we give such prediction a penalty. Specifically, define start to be the start labels vector for the batch and define end to be the end labels vector. weights for ith data entry in the batch of size n is calculated as follows:

$$w_i = \begin{cases} \frac{n}{2} + 1 & if\ start > end \\ 1 & otherwise \end{cases} \tag{1}$$

3.3.2 Penalty on further distance from true index

When the incorrect indices' predicted probability is greater than zero, softmax does not distinguish between different wrong indices. Given the nature of predicting an index, we should penalize more

if the predicted index is further away from true index. Therefore, we multiply the original loss by a weight defined as:

$$1 + \frac{log(|s - \hat{s}| + 1)}{C} \tag{2}$$

This weight ensures that correct prediction receives a weight of 1.0, with the weight increasing at a diminishing rate for "further" incorrect predictions. Constant C is chosen so that the mean of weighted loss is similar to the mean of unweighted loss.

3.3.3 Both

We also introduce a third loss extension, where both schemes above are used. The weights generated by the two extensions would be multiplied and then applied to the cross entropy loss.

4 Experiments & Results

4.1 Preprocessing

Applying the heuristics for word matching lowers OOV rate by more than 9%, and increases f1 by 3%. Applying spell-check (in addition to heuristics) further lowers OOV rate by 6%, and increases f1 by 0.5%. These results are run without dropout.

Preprocessing method	**Out-of-vocabulary rate**
Naive	19.52%
Heuristic	9.81%
Heuristic + spellchecker	3.37%

Table 2: OOV Rate across Preprocessing Methods

Preprocessing method	**Eval[1] EM**	**Eval[1] F1**
Naive	47.75	61.74
Heuristic	49.92	64.67
Heuristic + spellchecker	**51.02**	**65.09**

Table 3: Performance across Processing Methods

4.2 Tasks

We find that Quora task increases by 3 points from jointly training with SQuAD, while SQuAD performance suffered by 3 points. Pre-training on Quora hurts SQuAD performance by more than 13 points. It seems that Quora, the simpler task here, benefits from knowledge in solving the more complex question answering task.

[1]When training the model, we split out 5% of the official training set as the evalutiona/dev set. Thus we use the terminology Eval or Evalutation to refer to our custom made dev set, to differentiate from the official dev set published on https://rajpurkar.github.io/SQuAD-explorer/.

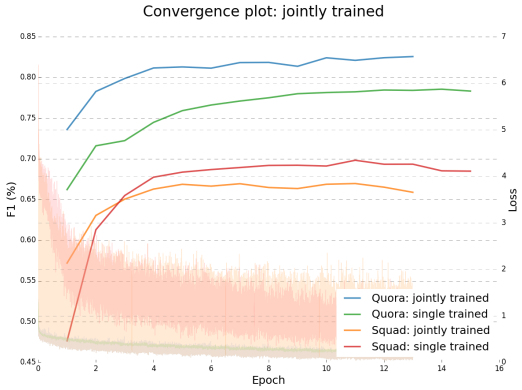

Figure 3: Joint training helps Quora, the similer task, more so than SQuAD, the more complex task.

	Single best model	**Jointly trained**	**Pre-trained w/ Quora**
SQuAD	**69.82**	66.96	57.52
Quora	78.45	**82.56**	78.32

Table 4: Eval F1 across Tasks

Although multi-tasking learning doesn't seem to benefit SQuAD as much on key metrics, we do see interesting patterns in predictions where model that has transfer-learned is better at certain skills.

DOCUMENT `Between 1975 and 2009 , Olympic Airways ( known after 2003 as Olympic Airlines ) was the country's state-owned flag carrier , but financial problems led to its privatization and relaunch as Olympic Air in 2009 ...`
QUESTION `What was Olympic Airways known as after 2003 ?`

Answer label: `(12, 13) Olympic Airlines`
Best single model: `(5, 6) Olympic Airways`
Pre-trained with Quora: `(12, 13) Olympic Airlines`

DOCUMENT `...This decrease in wages caused a period of compression and decreased inequality between skilled and unskilled workers. Education is very important for the growth of the economy, however educational inequality in gender also influence towards the economy. Lagerlof and Galor stated that gender inequality in education can result to low economic growth, and continued gender inequality in education, thus creating a poverty trap. It is suggested that a large gap in male and female education may indicate backwardness and so may be associated with lower economic growth, which can explain why there is economic inequality between countries.`
QUESTION `What impacts gender inequality in wages?`

Correct answers: `gender inequality in education` or `education`
Best single model: `skilled and unskilled workers`
Pre-trained with Quora: `gender inequality in education can result to low economic growth`
Jointly-trained model: `education`

In these two examples, the models learned from Quora tasks seem to better understand sentence structures and context. And in the second example, the jointly trained model is more precise and

to-the-point than the model pre-trained with Quora. That seems to inherit from SQuAD's strength on giving correct short answers.

4.3 Architecture

4.3.1 Number of sentinel variables

The original DCN paper suggests using sentinels to mark the end of a document or question; however, it does not clearly specify whether different sentinel variables or the same one is used for document versus question. So we explore both possibilities and compare them with the case in which no sentinel variable is used. When no sentinel variable is used, we simply put zero padding in the spot.

# Sentinel variables	Eval EM	Eval F1	Dev[2] EM	Dev[2] F1
0	**55.66**	69.34	54.87	66.25
1	55.57	**69.82**	56.58	66.85
2	55.10	69.21	**56.89**	**67.02**

Table 5: F1 across # Sentinels Used

Even though we find no significant difference on development set results among zero, one, and two sentinel variable(s) used, the model seems to generalize better on test set when sentinel variables are used.

4.3.2 Different modes of dropout

We explore two different approaches of dropout. In the first approach, we apply dropout to only the Bi-LSTM in the coattention encoder and the dynamic pointing decoder. In the second approach, we apply dropout to the text encoders as well. When applying full dropout (second approach), the model converges significantly more slowly. Yet, the full dropout approach overfits less than the partial dropout approach, as shown in plot.

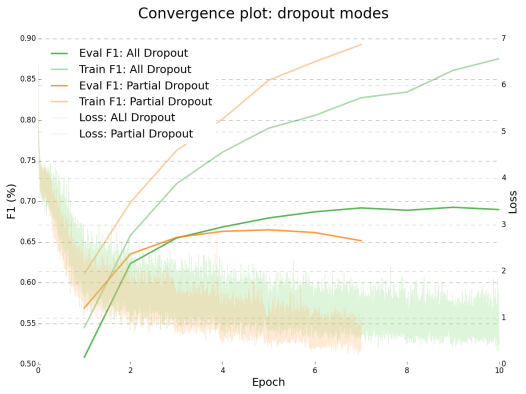

Figure 4: Full dropout, including dropout for coattention layer, results in less overfit and better performance.

4.3.3 Shared vs. separate encoding for document/question

We experiment with using separate encoding weights for question vs. document. The performance is 1% worse. This indicates that encoding encapsulates shared representation across document and questions.

[2]Dev here refers to the official dev set. Note that during training we don't use the official dev set to evaluate.

	Eval EM	Eval F1
Single best	**55.57**	**69.82**
Partial dropout	52.65	66.52
Separate encoding	55.15	68.84
Penalty: invalid start end	55.36	69.73
Penalty: further distance	55.41	69.59
Both Penalty	55.10	69.54

Table 6: Results across Architecture

4.3.4 Loss extensions

We further train our saved single best model with the loss extensions that are discussed in Approach.3.1, with a lower learning rate (0.0001). The loss extensions do not show meaningful improvement. This indicates that, the network is already fairly good at predicting a valid start/end, and the additional rule on validity of start/end does not provide new information.

However, we do still see cases where loss extensions have helped. So it is possible that the single best model has already reached a local optimum and continuing to train with the loss extension might not help much. Next we plan to start fresh training with the weighted cross entropy in place.

DOCUMENT `A party 's floor leader , ... He is kept constantly informed as to the status of legislative business and as to the sentiment of his party respecting particular legislation under consideration . Such information is derived in part from the floor leader 's contacts with his party 's members serving on House committees , and with the members of the party 's whip organization .`
QUESTION `How are floor leaders kept informed of legislative status ?`

Answer Label: `(117, 137) contacts with his party 's members serving on House`
Best single model: `(83, 91) constantly informed as to the status of legislative business`
Penalty - Further distance from true index: `(95, 104) the sentiment of his party respecting particular legislation under consideration`

DOCUMENT `... Kathmandu Metropolitan City has a population of 975,453 and measures 49.45 km2 ( 19.09 sq mi ) .`
QUESTION `How many square kilometers in size is Kathmandu ?`

Answer Label: `(132, 132) 49.45`
Best single model: `(135, 132) (N/A)`
Penalty - Invalid start end: `(132, 132) 49.45`

4.4 Hyper-parameter

4.4.1 Number of iterations in dynamic decoder

In this experiment, we use a model that applies dropout in only the Bi-LSTM in the coattention encoder and the dynamic pointing decoder. We train using different number of iterations, 4, 5 and 6, in the dynamic pointing decoder. The results show no significant difference between the different number of iterations, yet more iterations results in longer training time.

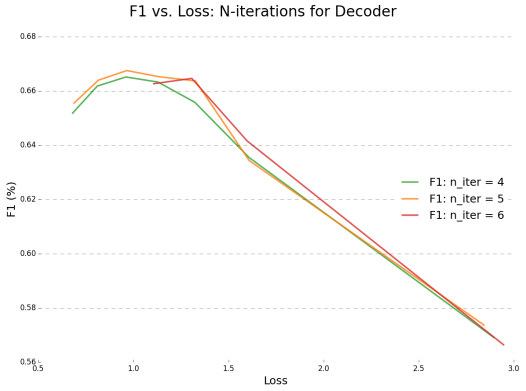

Figure 5: Evaluaton F1 vs. Loss for different iteration numbers for dynamic decoder. For the same training loss, the model with more iterations converge faster.

# iterations	Training EM	Training F1	Dev EM	Dev F1
4	63.00	76.06	52.70	66.59
5	73.73	85.06	52.81	66.75
6	63.74	76.30	52.86	66.46

Table 7: Training and evaluation results across # Decoder Iterations

5 Discussion

In this paper, we re-implement DCN and explore multi-task learning with Quora question pairs dataset. We find that careful pre-processing boosts performance by 4%. Other papers have introduced character embedding in addition to word vector embedding with performance improvement. In addition, upon inspecting error cases, we find our network is poor in understanding semantic structure. Additional features such as semantic parsing hierarchy could be beneficial.

With regards to network architecture, our addition of max-pooling and mean-pooling on top of questions attention boosts performance slightly. Among multiple papers, the attention/multi-perspective/co-attention module becomes increasingly complex. It would be ideal to have a generic architecture, such as LSTM, that can capture rich non-linear interaction between multiple input data schemes (document and question, in the case of SQuAD dataset).

ACKNOWLEDGEMENTS

We thank Richard Socher and Victor Zhong for their help and advice.

References

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[2] First quora dataset release: Question pairs. `https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs`. Accessed: 2017-03-21.

[3] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. *CoRR*, abs/1611.01604, 2016.

[4] Junbei Zhang, Xiaodan Zhu, Qian Chen, Lirong Dai, and Hui Jiang. Exploring question understanding and adaptation in neural-network-based question answering. *arXiv preprint arXiv:1703.04617*, 2017.

[5] Zhiguo Wang, Haitao Mi, Wael Hamza, and Radu Florian. Multi-perspective context matching for machine comprehension. *CoRR*, abs/1612.04211, 2016.

[6] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. A joint many-task model: Growing a neural network for multiple NLP tasks. *CoRR*, abs/1611.01587, 2016.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[8] How to write a spelling corrector. `http://norvig.com/spell-correct.html`. Accessed: 2017-03-21.

[9] blog/python-spell-checker.rst at master · mattalcock/blog. `https://github.com/mattalcock/blog/blob/master/2012/12/5/python-spell-checker.rst`. Accessed: 2017-03-21.

6 Appendix

6.1 Pre-processing OOV Samples

```
Naive on 840B (maintain cases):
Out of Vocab rate: 0.195166675316
Sample missing words:
        75:36
        Hoitzman
        Alloparenting
        Noorwood
        record-wettest
        Vartikas
        Mozni
        Hamovic
        Cctavianus
        parasitoid-host
        Aslations
        aeronatical
        'haut
        congitive
        un-vouched
        'dry
        Iranina
        Unittarians
        madrash
        intracellar
        'industrial
        riwaaydo
        Corixids
        Gamrents
        impressible
        higher-than-allowed
        familly-festival
        philosohy
        -selling
        'Why
        'instrumentalism
        LIangzhu

Smart heuristics on 840B:
Out of Vocab rate: 0.0981023059491
Sample missing words:
        lemuns
        Pietho
```

```
        Monrova
        couinty
        Aermenian
        Achaemind
        onearlier
        activitsts
        ratifing
        Gotamas
        Trukestan
        Anatoian
        sanskirt
        deveops
        authoriisations
        Confederacys
        pplace
        Callaeci
        cutlture
        larcrosse
        shoebirds
        realied
        sulfonamine
        repremend

Spell checker on 840B:
Out of Vocab rate: 0.0337156399744
Sample missing words:
        Extrapositioning
        then____
        Pseudopagurus
        empiresovereignty
        rosselliius
        Astadhyayi
        commercialand
        parchmentised
        problemsome
        haemophi
        gynephlia
        Cauberflote
        Neostokavian-Ijekavian
        protobacterium
        Ishmuhamedov
        Tajikfilm
        becomingUnitarian
        Zenlenyi
        premandolins
        Dahabhiil
        ethninticities
        bioremediates
        Paraneopter
        Whichgroup
```