Transfer Learning for Machine Comprehension over Bi-Directional Attention Flow Networks

Carlos Castro
carlosscastro@berkeley.edu

Abstract

Machine comprehension to answer a query about a given context paragraph requires modeling complex interactions between the context and the query. In the recent years, with the release of the Stanford Question Answering dataset (SQuAD) (Rajpurkar et al., 2016) and the Microsoft MAchine Reading COmprehension Dataset (MS-MARCO) dataset (Nguyen et al., 2016), there were significant improvements in the state of the art, to the point where some neural network architectures are relatively close to achieving human level accuracy on the SQuAD dataset. Despite the great advances in the field of machine reading comprehension, training state of the art models requires humongous amounts of labeled data consisting of thousands of passages plus multiple question-answer pairs on each passage. This acts as a limiting factor in the applicatiblity of machine comprehension techniques in technology outside of research benchmarks. In this paper, we aim to provide a solution to this shortcoming, by studying transfer learning over a neural network trained using the SQuAD dataset to other corpora. Successful results would open the door to multiple direct applications and further research opportunities.

1 Introduction

Since the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) was released, rapid progress was made in the field of machine question answering. The original paperal-ready proposed a strong logistic regression model, and later other better performing approaches were published, based off match-LSTMs (Wang et al.) and bi-directional attention flow networks (Seo et al., 2017). Despite this recent progress in the field of machine text comprehension, SQuAD is a luxury dataset with a high amount of labeled data. This makes the applicability of these approaches to other corpora somewhat limited, since such humongous training data is not available.

In the field of computer vision, neural networks are rarely trained from scratch. Generally networks are trained with ImageNet (Deng et al.), a large-scale hierarchical image database, to obtain features. Transfer learning allows us to transfer this knowledge to other tasks. Analogously to ImageNet for computer vision, we could use weights from neural networks trained for SQuAD dataset as initial weights for different tasks in other corpora.

In this paper, we study training bi-directional attention flow networks (BiDAF) on the SQuAD dataset and then transferring that knowledge to perform on the Microsoft MAchine Reading COmprehension Dataset (MS-MARCO) dataset (Nguyen et al., 2016), which is based off real web queries and human written answers. Given that the SQuAD dataset is extracted from Wikipedia content (Rajpurkar et al., 2016), while the MS-MARCO dataset is obtained from web user queries and web content, the domains and quality of the text vary wildly. With MS-MARCO being quite representative of web content, it contains a wide variety of text and content quality, while the Wikipedia content from the SQuAD dataset is quite curated and edited by the community. Considering this, transfer learning from a model trained on the SQuAD dataset to the MS-MARCO dataset is a non-trivial task, and with relevant implications as we discuss in sub-section 1.1.

1.1 Motivation

The implications of successful transfer learning from models trained on the SQuAD dataset to other corpora with little or no labeled data are extremely relevant for most conversational interfaces, including chatbots, conversational agents and conversation-augmented applications. Below, we describe two brief scenarios that demonstrate potential applicability of this study.

First, consider question-answering chatbots. It might be useful to have chatbots that receive certain passages and can answer questions about those passages. A concrete example would be a car-embedded chatbot that ingests the car manual, and then the passengers can ask questions to the car, such as *what is the ideal tire pressure?*. Should transfer learning be proven feasible over BiDAF networks, this chatbot could be created with little or none labeled examples.

Another example could be augmenting web content with conversational agents. A possible implementation could be a web browser extension to which users can ask questions about the current web page. A user loads a web page and asks a question, then the text and the question are entered into a model trained over the SQuAD dataset, and the user gets an answer to their query without even reading the article.

Note that without the use of transfer learning, it would not be possible to enable the scenarios described above, because this is such a complex problem space that a rather large number of labeled documents is required to train a model. This is supported by the fact that machine comprehension state of the art skyrocketed after the release of the SQuAD dataset (Natural Language Computing Group) (Seo et al., 2017).

1.2 Paper organization

In section 2 we discuss bibliography and related work around machine comprehension datasets, transfer learning and machine comprehension neural network architectures. Section 3 describes the approach and experiments to study this problem space, and later we review the experiment results in section 4. We outline the next steps towards this study in section 5.

Passage: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. When asked where all the money had gone, Tesla responded by saying that he was affected by the Panic of 1901, which he (Morgan) had caused. Morgan was shocked by the reminder of his part in the stock market crash and by Teslas breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began. **Question:** On what did Tesla blame for the loss of the initial money? **Answer:** Panic of 1901

Table 1: An example from the SQuAD dataset.

2 Related work

2.1 SQuAD Dataset

The Stanford Question Answering Dataset (SQuAD) (Natural Language Computing Group), is a reading comprehension dataset consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage (Rajpurkar et al., 2016).

In the SQuAD dataset, the answer to every question is a segment of text, or *span*, from the corresponding reading passage. SQuAD contains 107,785 question-answer pairs on 536 Wikipedia articles.

2.2 MS-MARCO Dataset

The Microsoft MAchine Reading COmprehension (MS-MARCO) (Nguyen et al., 2016) dataset is aimed to overcome a number of well-known weaknesses of previous publicly available datasets for the same task of reading comprehension and question answering. In MS-MARCO, all questions are sampled from real anonymized user queries. The context passages, from which answers in the dataset are derived, are extracted from real web documents using the most advanced version of the Bing search engine. The answers to the queries are human generated, and contain words that may not be in the original paragraph. This last consideration is a crucial difference with the SQuAD dataset. MS-MARCO contains 100,000 queries

Passage: The goal you choose will determine your path. A clinical psychologist, for example, will have heavy training in both the theory and practice of psychology. Youll typically need a doctorate degree in a psychology-related field in order to build a career. A psychiatrist has even more rigorous demands; becoming one requires medical school training. In order to do this, you have to do three things: 1) Get work experience under the supervision of a licensed professional (usually for two years); 2) Pass the board exams; and 3) Depending on your state, present a valid case study to the board. After all that work, youll finally be able to call yourself a psychologist!
Question: do you have to do a phd to be a clinical psychologist
Answer: Yes

Table 2: An example from the MS-MARCO dataset.

with their corresponding answers.

2.3 Reading comprehension

For reading comprehension style question answering, a passage **P** and question **Q** are given, our task is to predict an answer **A** to question **Q** based on information found in **P**. The SQuAD dataset further constrains answer **A** to be a continuous sub-span of passage **P**. Answer **A** often includes non-entities and can be much longer phrases. This setup challenges us to understand and reason about both the question and passage in order to infer the answer. Table 1 shows a simple example from the SQuAD dataset. As for MS-MARCO dataset, several related passages **P** from Bing Index are provided for a question **Q**. Besides, the answer A in MS-MARCO is generated by human which can not be a continuous sub-span of the passage.

2.4 Bi-Directional Attention Flow

To study transfer learning, we choose one method out of the multiple state of the art techniques named Bi-Directional Attention Flow (BiDAF) network. BiDAF is a hierarchical, multi-stage architecture for modeling the representations of the context paragraph at different levels of granularity. BIDAF includes character-level, word-level, and contextual embeddings, and uses bi-directional attention flow to obtain a query-aware context representation. It would be interesting studying how other state of the art techniques, such as R-NET, respond to transfer learning.

2.5 Transfer learning

Here we provide useful definitions and conventions related to transfer learning. Let a domain **D** consist of a feature space **X** and a marfinal probability distribution $P(X)$. Given a source domain $\mathbf{D}_s$ and a learning task $\mathbf{T}_s$, a target domain $\mathbf{D}_t$ and a target task $\mathbf{T}_t$, *trasnfer learning* aims to improve the learning of the predictive function for $\mathbf{D}_t$ using the knowledge in $\mathbf{D}_s$ and $\mathbf{T}_s$. In the *inductive transfer learning* setting, the target task is different from the source task, no matter when the source and target domains are the same of not. Conversely, in *transductive transfer learning*, the source and target tasks are the same, while the domains are different. In this situation, little or no labeled data are available in the target domain, while lots of data are present for the source domain (Pan and Yang) (Conneau et al., 2017) (Peters et al.) (Conneau et al., 2017).

In computer vision, neural network models are rarely trained from scratch. In general, initial weights are the result of training with ImageNet (Deng et al.), a large-scale hierarchical image database, to obtain features. There is already relevant work studying transfer learning over text tasks, mostly around vector representations of words.

3 Methods

In this section, we describe our methods to study transfer learning over BiDAF networks.

First, we train a BiDAF network on the SQuAD dataset, being this our source task and domain, and we'll refer to the resulting model as the *source model*. Since we want to study how we can transfer the knowledge to other tasks and domains with none or limited labeled data, we perform a number of transfer experiments with varying amounts of labeled question-answer pairs from the MS-MARCO dataset, which will be our target task and domain for transfer learning.

Let n be the number of labeled samples in the target domain we want to experiment with, then an experiment is as follows: Select n random question-answer pairs from the MS-MARCO dataset, re-train the source model on the randomly selected pairs, and evaluate the resulting model on

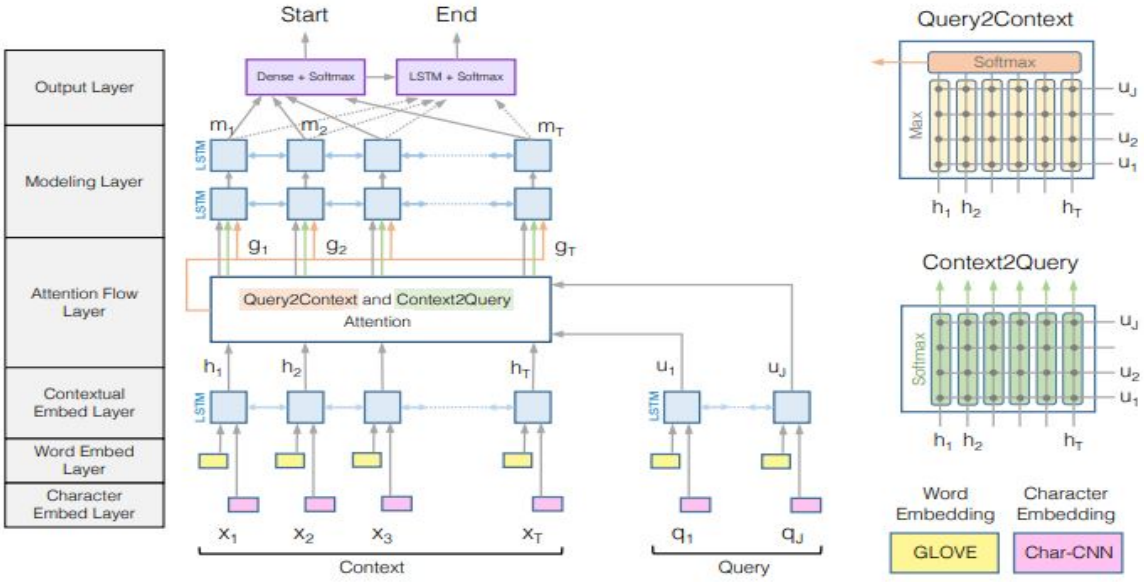

Figure 1: Bi-directional attention flow network architecture

the MS-MARCO evaluation set. We run this experiment multiple times, for different values of n, including $n = 0$, which is the case where the target domain has 0 labeled examples.

Finally, we also train another BiDAF network on the MS-MARCO dataset without transfer, as a goal for accuracy. The ideal result of this experiment is to achieve similar accuracy in our transfer experiments as in this non-transfer model, which would mean that we can perform this task successfully without fully training our model for this task. It is worth noting that BiDAF performance over MS-MARCO was never studied, but we will compare the results to those

4 Results

4.1 BiDAF over SQuAD dataset

The first step of our research is to train a BiDAF over the SQuAD dataset. We train for 16.4 hours using a single NVidia K80 GPU, and obtain results close to the current state of the art, with exact match percentage of 65.64 and F1 score of 75.68. The baseline logistic regression performance for this dataset is exact match percentage of 40.0 and F1 score of 51. Human performance for this task is EM at 82.3 and F1 score of 91.2 (Rajpurkar et al., 2016) (Natural Language Computing Group).

4.2 BiDAF over MS-MARCO dataset

We also consider the baseline models for MS-MARCO dataset. The baseline performance for this dataset using memory networks for passages is ROUGE-L at 0.119 and BLEU 0.340 (Nguyen et al., 2016). One of the top results for MS-MARCO, R-NET (Natural Language Computing Group), report ROUGE-L 0.4289 and BLEU 0.4222. Human performance was ascertained by having two judges answering the same question and measuring their responses, obtaining ROUGE-L 0.47 and BLEU 0.46.

5 Next steps

5.1 BiDAF over MS-MARCO

The next step to follow in the implementation is to train a BiDAF network over the MS-MARCO dataset. There is a major difference in that the MS-MARCO dataset has multiple text passages for each question, corresponding with the multiple results from the web query. R-NET has achieved the top results in the MS-MARCO leaderboard by concatenating all the candidate passages, which makes the data from MS-MARCO and SQuAD have the same shape. This training involves no transfer, it is used only for comparison with future models.

5.2 MS-MARCO Evaluation

While the passages provided in MS-MARCO generally contain useful information for the queries, the answer could contain words outside of the query. Because of the nature of the dataset, the authors of MS-MARCO chose to use ROUGE-L and

phrasing-aware evaluation framework to measure performance for long textual answers. We will implement ROUGE-L and BLEU metrics (Nguyen et al., 2016).

5.3 Transfer experiment

We will write the transfer experiment where we train a model with SQuAD and then re-train with a different numbers of samples from MS-MARCO (including zero examples). We will then evaluate the obtained accuracy metrics with our non-transfer MS-MARCO model.

References

A. Conneau et al. 2017. Supervised learning of universal sentence representations from natural language inference data.

Jia Deng et al. Imagenet: A large-scale hierarchical image database.

Microsoft Research Asia Natural Language Computing Group. R-net: Machine reading comprehension with self-matching networks.

Tri Nguyen et al. 2016. Ms marco: A human generated machine reading comprehension dataset.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning.

Matthew Peters et al. Deep contextualized word representations.

Pranav Rajpurkar et al. 2016. Squad: 100,000+ questions for machine comprehension of text.

Minjoon Seo et al. 2017. Bi-directional attention flow for machine comprehension.

Wang et al. Machine comprehension using match-lstm and answer pointer.