Alma Mater Studiorum · University of Bologna

Question answering on the SQuAD dataset

NLP course final project

Leonardo Calbi (leonardo.calbi@studio.unibo.it)
Lorenzo Cellini (lorenzo.cellini3@studio.unibo.it)
Alessio Falai (alessio.falai@studio.unibo.it)

January 22, 2021

Contents

1 Summary

The tasks of machine comprehension (MC) and question answering (QA) have gained significant popularity over the past few years within the natural language processing and computer vision communities. Systems trained end-to-end now achieve great results on a variety of tasks in the text and image domains.

In this work we address a question answering problem and, in particular, the Stanford Question Answer Dataset (SQuAD) problem: given a large collection of Wikipedia articles with associated questions, the goal is to identify the span of characters that contains the answer to the question.

This project focuses on the SQuAD v1.1 dataset, that contains more than 100,000 question-answer pairs on more than 500 articles. Here each question has at least one associated answer, whereas in the SQuAD v2.0 dataset there are also questions with no answers at all.

In this work we implement and compare three different models: a naïve LSTM encoder-decoder that will act as our baseline, the Bi-Directional Attention Flow (BIDAF) network (CITAZIONE) and a model that wraps a pretrained Bidirectional Encoder Representations from Transformers (BERT), as language model, coupled with a custom output layer on top of it.

Our experimental evaluations show that our models achieve .

2 Background

As already described above, the question answering task is based on the idea of identifying one possible answer to the given question as a subset of the given context. Since the input data is of textual form and the latest models for the task are all based on neural architectures, there is the need to encode such text into a numeric representation. As of today, there are two main approaches to embed words in numerical format: sparse embeddings, like TF-IDF [4] and PPMI [2], and the modern dense embeddings, such as Word2Vec [1] and GloVe [3]. In the latter case, the first step of the NLP pipeline is usually done with shallow encoders, that compute meaningful embeddings for each word in the input vocabulary.

Descriviamo la teoria dietro embedding, LSTM, attention, transformer, BERT e BIDAF?

3 System description

In this work we implement three distinct models: a naïve LSTM encoder-decoder that will act as our baseline, the Bi-Directional Attention Flow (BIDAF) network (CITAZIONE) and a model that wraps a pretrained Bidirectional Encoder Representations from Transformers (BERT), as language model, coupled with a custom output layer on top of it.

Each model is trained on the SQuAD v1.1 training set and evaluated on the SQuAD v1.1 dev set, available on the github repository of the SQuAD project (link).

The models evaluation is carried out on the same preprocessed data: indeed, we defined a preprocessing pipeline, applied it to the training and dev set once and the resulting dataset is fed to each model. (CONTROLLARE QUI mi pare che per bert abbiamo tokenizzato diversamente)

In order to perform an effective models comparison, each model shares the same input and output layer: what changes is the language model (embedding + attention layers) and the modelling layer.

The following is a detailed description of the models.

3.1 Baseline

.

3.2 BiDAF

Sintesi tra paper bidaf e medium

3.3 BERT

.

4 Experimental setup and results

5 Analysis of results

6 Discussion

Bibliography

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed Representations of Words and Phrases and their Compositionality". In: *NIPS*. Curran Associates, Inc., 2013, pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[2] Yoshiki Niwa and Yoshihiko Nitta. "Co-Occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries". In: *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*. COLING '94. Kyoto, Japan: Association for Computational Linguistics, 1994, pp. 304–309. DOI: 10.3115/991886.991938. URL: https://doi.org/10.3115/991886.991938.

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

[4] "TF–IDF". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 986–987. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_832. URL: https://doi.org/10.1007/978-0-387-30164-8_832.