A summary on paper 1607.05368.pdf

In response to AI (MSc) Entry Questions from University of Leeds
Feng MIAO, lrrm523@gmail.com, 31/JAN/2023 (v2)

Introduced as an extension to word2vec, doc2vec can generate numerical representation from a document of arbitrary length. However, observed performances of doc2vec and the "better as standalone model" variant – dmpv have disproved the original paper of doc2vec. To resolve the uncertainty and discrepancy around the performance standing of doc2vec, this paper presents an arrangement of experiments to evaluate the performance of doc2vec and several simple or state-of-the-art document embedding methods.

The experiments are organized into two phases, the small corpora phase and the large corpora phase. Two task settings, Question Duplicate (D-DUP) and Semantic Textual Similarity (STS) are conducted throughout the experiments. In the small corpora phase, given the unsupervised learning assumption, doc2vec training dataset consists of development and test partitions, development partition is used for hyper-parameters optimization. To understand the impact of the document length, Q-DUP data is prepared from a forum dataset with longer documents, test partition is roughly 130 words in each document; whereas STS data is from a smaller dataset with shorter documents, test partition is only 13 words on average.

In the Q-DUP task, embedding vectors are generated from documents in test partition with the trained doc2vec model. As to word2vec, the embeddings are the mean of embedding vectors of the composing words. Document pairs' cosine similarity probabilities are predicted against vectors produced by the models and then put in a rank with descending order. While walking through the ordered likeness probabilities, TPR and FPR are calculated, and then the ROC curve emerges, which ultimately allows AUC calculation. Q-DUP evaluation in the small corpora phase involves dbow / dmpv (variants of doc2vec), skip-gram / cbow (variants of word2vec), and n-gram. STS task evaluation in the small corpora phase, however, additionally includes top-performing system in the competition as "DLS".

When it comes to large corpora phase, training data are changed to the full set of English Wikipedia and a subset of Associated Press English news. Considering the findings from the previous phase, dmpv and cbow are excluded in this phase. The test data are held out, pre-trained model is used in test inference, meanwhile, development partition is employed in tuning relevant hyper-parameters. Besides doc2vec, the evaluation also includes n-gram, pre-trained skip-gram model GL-NEWS, pre-trained skip-thought model BOOK-CORPUS, as well as pre-trained paragram-phrase (pp) model PPDB.

In conclusion, though not overtaking the performance of supervised benchmarking system "DLS", doc2vec performs better than baseline methods - word2vec and n-gram, especially in dealing with longer documents. Simpler doc2vec variant dbow outperformances the complex one – dmpv. With large external corpora, doc2vec demonstrates robustness by surpassing baseline methods and competing methods - skip-thought and paragram-phrase, also by showing narrow performance gaps between models trained from different source of external corpus. Pre-trained doc2vec can be used as an off-the-shelf model. The paper also explores options to improve doc2vec performance, e.g., with pre-trained word embeddings, doc2vec model training converges faster. A set of hyper-parameter settings for general-purpose applications are suggested. After all, source code and pre-trained models are provided for replication work.