Doc2Img: A New Approach to Vectorization of Documents

ShreeRanjani SrirangamSridharan, Mudhakar Srivatsa, Raghu Ganti (IBM-US), and Christopher Simpkin (Cardiff University)

Challenge

Paragraph2Vec (also known as Doc2Vec) captures the word's context for a given document. However, the Doc2Vec approach does not capture the similarity between words across documents. We present our vector space embedding of documents approach, Doc2Img, which not only captures the context of the word in the document, but also the similarity between words across documents

Scenario

Workflows are made of micro-services and with increased availability of various forms of them, it is imperative to use machine learning to "understand" these micro-services to compose them. In this scenario, apps are described using text and sensor requirements for these apps are predicted using a multi-layer perceptron (neural network).

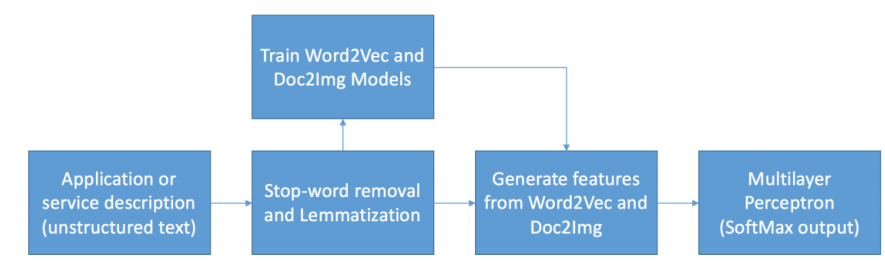

Fig. 7: End-to-end flow of building the classifier

Doc2Img Approach

Doc2Img computes word vector representations for each document and then "treats" all these word vectors as a multi-dimensional image. For example, if we learned n-dimensional word vectors and we let the image dimension of the document be q×q, each of the n-dimensional vector is inserted into a cell (pixel) of the q × q image. This multi-dimensional vector is then compressed using a convolutional auto-encoder to learn a compact representation for the document. Image representation captures the spatially the similarity of words in a document (different from another document).

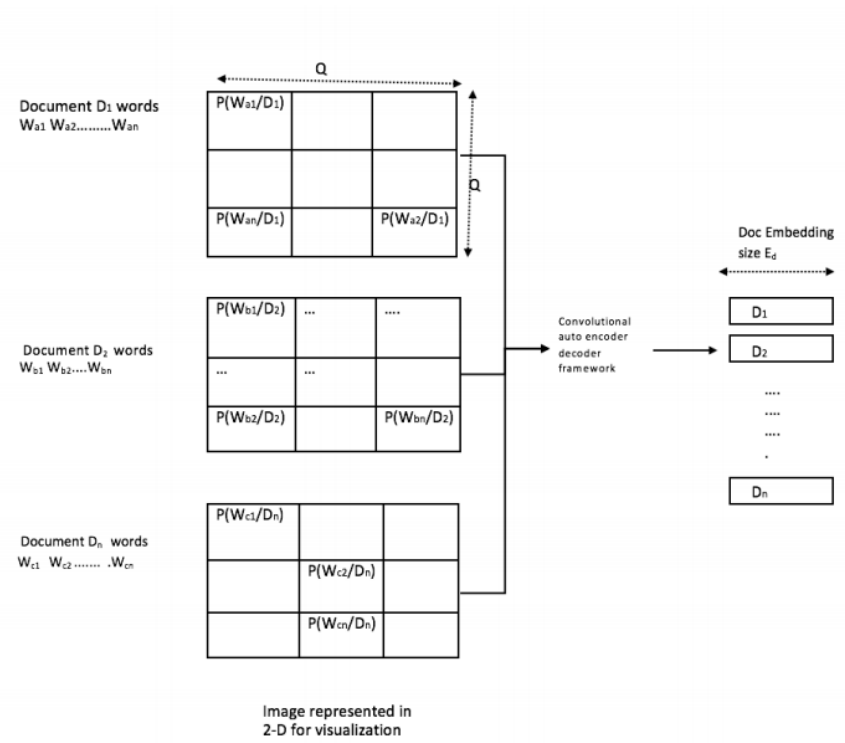

Fig. 6: Illustration of end-to-end document vector construction

Experiments

Three hundred applications and their text along with sensor requirements for each of these applications are collected; goal is to predict sensor needs based on vector embeddings.

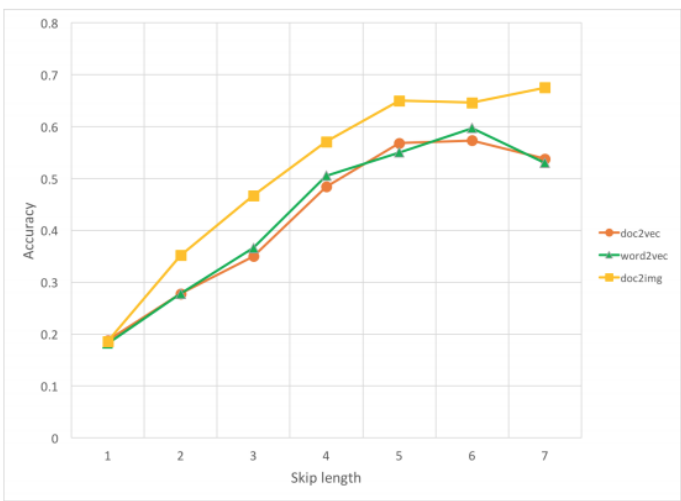

Fig. 12: Accuracy as skip length is increased for the proposed approach compared with the existing Doc2Vec approach

Doc2Img outperforms existing vector embedding approaches such as Doc2Vec and Word2Vec.