Assignment - 2
Design Review

Word embeddings have recently gained traction in the field of natural language processing. They have become so ubiquitous that it can be safely said that most natural language pipelines start with an embedding layer or pre-trained embeddings dictionaries. Mikolov et. al in their work " showed the semantic abilities of word embeddings trained on their skip-gram model (word2vec). Their experiments shows that words which have same linguistic meanings like Apple, Orange (fruits) are closely placed in the embedding space. Thus, words with similar meanings form clusters. They also shows that the vectors [Paris - France] and [Delhi - India] have very small cosine distance (same direction). This proves that the word embeddings pack in a lot of semantic information. Since, human visualization is limited to only three dimensions, the most commonly used approach used by NLP researchers, to scrutinize these embeddings is running a dimensionality reduction algorithm like PCA[1] or t-SNE.

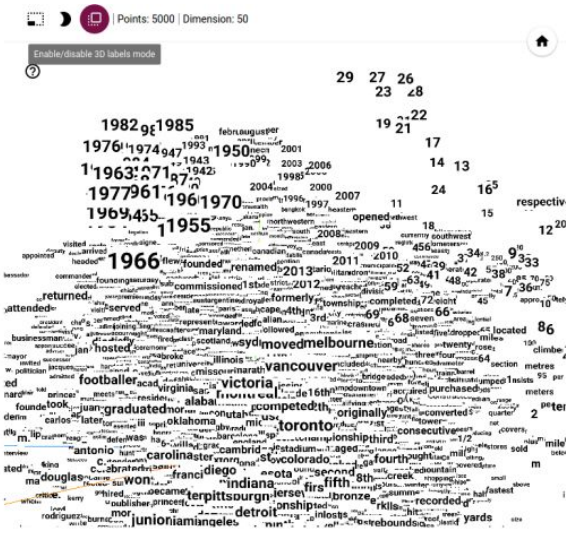

Figure 2: 3D labels view of word embeddings.

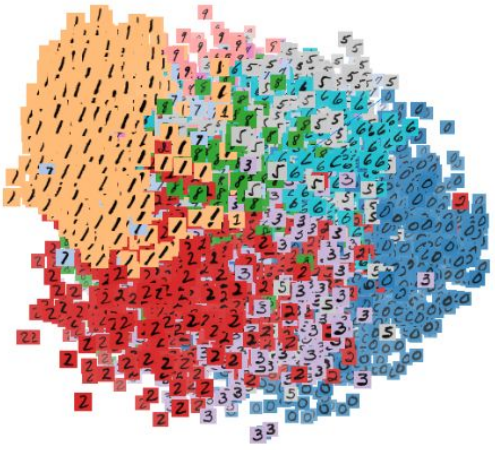

Figure 3: Image view of the MNIST dataset.

Figure A

Smilkov et al.[2] give a tool for generating visualizations for embeddings (not limited to words). The tools in integrated into tensorflow, a commonly used machine learning library. Figure one shows, two outputs generated by this tool. The first figure shows a 3D dimensional project of word-embeddings trained on a corpus (like Wikipedia).

[1] Principal Component Analysis

[2] "Embedding Projector: Interactive Visualization and Interpretation of ...." 16 Nov. 2016, https://arxiv.org/abs/1611.05469. Accessed 30 Jan. 2019.

This image is a gives a good preliminary analysis of embeddings the learnt by the model, but there is still room for improvements. The most noticeable elements are the clusters of words. Years, numbers and place names form three different clusters. This image does not however show the full potential of word embeddings, these embeddings have other non-intuitive properties like vector clusters (e.g. country-capital vectors). An important missing information is the scale. Even though this is a PCA output, the tool supports custom axes inputs as well. In such cases scale and ticks might be needed. The colors are not used which makes it difficult to analyze the image. A better visualization could have used different colors for different clusters. Some words like footballer and graduate are closely placed, this might be because of the context of text corpus, but these two words might not be so commonly used together in general. One improvement like adding dynamic utilities like highlighting clusters or filtering out unwanted data could further improve analysis in some cases. The second figure is 2D projection of MNIST data. Again clusters are the most important aspect of this visualization and since MNIST does not contain other semantic information, this is the most important aspect of this dataset. Different colors have been used for different number classes which helps in figuring out the clusters. A 3D project however, might be more effective. It can be seen that class 3 and class 2 are overlapping. Viewing the same dataset on a 3D plot might make more distinct clusters.

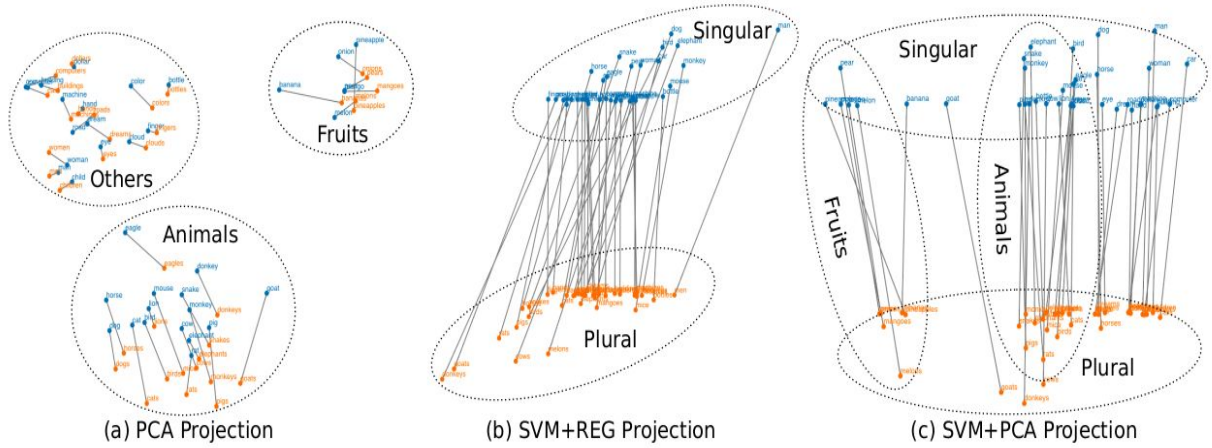

Figure B

Figure B shows another word embedding visualization technique taken from a study conducted by Liu et al.[3] The image shows two type of words, singular (blue) and plural (orange). The (a) is a simple PCA 2D-projection of the word embeddings. It

[3] "Visual Exploration of Semantic Relationships in Neural Word ...." http://www.sci.utah.edu/~beiwang/publications/Word_Embeddings_BeiWang_2017.pdf. Accessed 30 Jan. 2019.

can be seen that PCA clusters words into different clusters just like Figure A. This image has gives more information than Figure A. The image captures the relationship between singular and plural words. The use of different colors and lines between singular words and their plural counterparts makes this methods effective. (b) and (c) use a different technique (a). Instead of capturing the variance in the data, the new axes give more emphasis on the singular-plural relationship.