DeepGuidance: Suggestions for Areas of Study

Anonymous ACL submission

1 Introduction

University students many enter as passionate students with undecided majors. As they matriculate, many will switch their major at least once before graduation. To guide the students' selection of a major, this paper uses a neural network model trained on a corpus with 3 billion running words to suggest an area of study. The model takes as input natural language terms representing a students interests or skills. After training, the model predict correct results with between 60% and 90% accuracy. Since there were some anomalies in training, future work involves smoothing the model to create more predictable accuracy and using a different data set for more accuracy.

The high ambiguity, subtlety, and variability embedded in the structure of human language makes deciphering natural language difficult for a computer (Goldberg, 2016). Thus, using a neural network model with distributed vector representations lets the computer help us to uncover hidden connections between related terms. This can assist in making better suggestions than a human might be able to, given seemingly unrelated interests. In this paper, we describe the our approach 2 and experiments 3 and discuss the results.

2 Architecture

Figure 1 shows the architecture of the neural network used in the area of study suggestion task. A matrix of 5 by 300 is fed in as input, flattened, and subsequently passed through 5 tanh-activated hidden layers, where each of these consists of decreasing numbers of individual nodes (512 down to 128). The final layer uses a softmax activation function to generate a normalized distribution over the 118 areas of study selected for the task.

Name	Activation	Deep	Dropout
NN01	ReLU	False	False
NN02	ReLU	False	True
NN03	ReLU	True	False
NN04	ReLU	True	True
NN05	tanh	False	False
NN06	tanh	False	True
NN07	tanh	True	False
NN08	tanh	True	True
NN09	sigmoid	False	False
NN10	sigmoid	False	True
NN11	sigmoid	True	False
NN12	sigmoid	True	True

Table 1: Tested Neural Networks' Properties

3 Experiments

This experiment consisted of two stages: choosing a neural network architecture and observing labeling trends. During the first phase, 12 different net architectures were compared on 3 variables: activation function used in hidden layers, depth of the net (2 or 5 layers), and whether a dropout layer was included to set a random half of inputs to zero during each sample training. Table 1 shows the varying properties for each neural net.

For the second experiment, samples were chosen from the pre-trained Google News embedding model. Each sample consists of five inputs, which are randomly chosen words from the embedding model (Mikolov et al., 2013). A true label is a one-hot vector indicating the correct output class, calculated by choosing the closest output class embedding to the average of a sample's inputs. This experiment uses cosine distance to measure closeness.

4 Results and Discussion

Somewhat surprisingly, NN07, the tanh-activated, deep net with no dropout, consistently performed the best on input data sets of varying sizes.

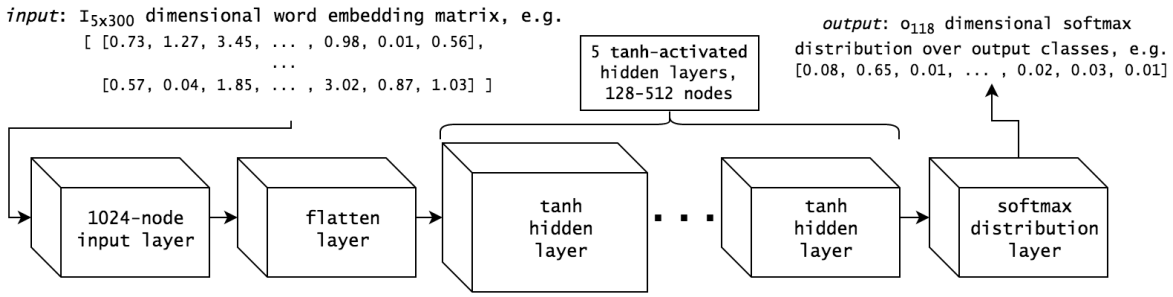

Figure 1: NN07, Tanh-Activated Deep Neural Network Architecture

Figure 2: Accuracies NN03 (middle), NN07 (top), NN11 (bottom) for a sample run of 5 million samples with batch size of 100.

For all but two nets tested, accuracy and loss both consistently improved over time. However, though the general trend showed learning progress, the different network architectures also showed consistent oscillation which did not decrease over time.

The two outlier architectures were deep sigmoid-activated networks, with and without dropout layers. No improvement or clear learning was shown, even over a very large set of samples.

In terms of labeling trends, using cosine distance yielded interesting results. With a run of 5 million random samples over 118 output classes, an even distribution of predictions would expect around 40,000 samples to be assigned each label. There were outliers, as can be seen in Figure 3. Computer Science as an area of study was assigned to only 654 samples, while Geography was assigned to 1,160,335.

Clearly there is some clustering occurring within the embedding model vector space. Since this specific embedding model was trained on co-occurrence within Google News articles, we can say that 'geography' cooccurred with many other terms. No matter which terms we choose, their average will likely be closest to 'geography.' To account for this, we can add an importance weighting to word embedding terms.

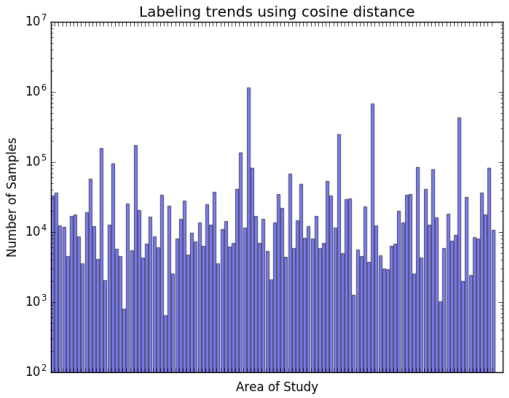

Figure 3: Distribution of labels among 5 million random samples.

5 Future Work

The accuracy of this model depends on the data on which it is trained. Therefore, rather than use the pre-trained Google News vectors, we will train vectors on corpora scraped from relevant websites. Samples will be created using random groups of words from university web pages or blogs of university faculty and students, labeled using subjects' self-declared area of study or specialization (Papoutsaki et al., 2015). Using this custom embedding model and more descriptive samples, we hope to refine the results to use the re-trained net for real application. In this vein, we plan to create an Alexa skill to implement this program as a useful application for a broader audience.

References

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57:345–420.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

Alexandra Papoutsaki, Hua Guo, Danae Metaxa-Kakavouli, Connor Gramazio, Jeff Rasley, Wenting Xie, Guan Wang, and Jeff Huang. 2015. Crowdsourcing from scratch: A pragmatic experiment in data collection by novice requesters. In *Third AAAI Conference on Human Computation and Crowdsourcing*.