Affect encoding in word embeddings

Yuhan Zhang[1], Wenqi Chen[1], Ruihan Zhang[2], Xiajie Zhang[2]
[1] Harvard University, [2] Massachusetts Institute of Technology

An increasing trend in natural language processing has been investigating what syntactic and semantic knowledge can be learned by large neural networks (NN) in word embeddings (e.g., Ettinger, 2020; Linzen et al., 2016; Manning et al., 2020). **Here we ask whether word embeddings that are fed into NNs encode intricate lexical semantic meaning.** In particular, we focus on the affect meaning of words, which, according to Osgood et al. (1957), involves three dimensions -- "valence" represents the pleasantness of a word (*nightmare* vs. *love*); "arousal" represents the intensity of emotion invoked by the word (*napping* vs. *abduction*); "dominance" represents the level of control exerted by the word (*weak* vs. *powerful*). We adopted three different analytical methods–principal component analysis, cosine similarity analysis, and a supervised classifier probe–to investigate whether word embeddings encodes information along the three affect dimensions that resemble human judgments. A positive correlation will indicate that the affective meaning is well captured by word embeddings. The human judgments of words' affective values on three dimensions came from the VAD dataset–where 20k English words were annotated by raters based on their perceived values on the aforementioned affective dimensions (Mohammad, 2018). The tested word embeddings were GloVe (Pennington et al., 2014), vanilla BERT embeddings (Devlin et al., 2019), embeddings from BERT-based model fine-tuned on a GoEmotion dataset with 27 emotion categories (Demszky et al., 2020), and our BERT-based contextualized word embeddings derived from aggregating the context-specific word embeddings from a vanilla BERT with the IMDB movie review dataset (Maas et al., 2011).

Figure 1 shows the PCA results with the correlation coefficients between word vectors from human judgments and the two principal components of each word's embeddings. The significant correlation coefficients indicate that the affective meaning is captured by word embeddings. Interestingly, each type of word embedding encodes the affect meaning differently and with diverse strengths. This pattern also parallels with the correlation coefficients of pairwise cosine similarities for 80 strong affect words in Table 1. For our supervised probe, Figure 2 displays the pipeline: we added a linear classifier layer after word embeddings for binary classification tasks on each of the three affect dimensions. The validation and affect word sample test results indicate that BERT-based contextualized embedding performed the best and that the valence dimension was the easiest to predict. Besides, the attention-based model such as BERT is better at capturing the affect meaning of the words compared with unsupervised learning-based models such as GloVe.

Then we ask whether affect-enriched word embeddings improve the performance in downstream affect-related tasks. We compared the performance of (i) the vanilla BERT model and (ii) the BERT-based model fine-tuned on the human labeled VAD dataset, on the task of predicting positive/negative IMDB movie reviews. Figure 3 shows that the affect-enriched model performed better than the vanilla BERT in the 10 epochs under investigation. Noticeably, the performance of the affect-enriched model had been superior to the vanilla model from the very first epoch. As shown in Figure 4, the affect-enriched model improved rapidly as the training progressed. Thus, fine-tuning BERT on affect datasets enhanced the model's performance on downstream sentiment analysis tasks, especially in the small-data regime.

Above all, we provide positive evidence that word embeddings from statistical learning and large neural network models do capture the affect meaning of words, but in different ways which might result from their individual training algorithm. The classifier result indicates that the easiness to predict intricate dimensions of affect meaning differs by the word embedding type, which invites future investigation into neural networks' deep knowledge about meaning. We further show that affect-enriched word embeddings enhance the downstream sentiment-related tasks, which is informative and translatable to other NLP tasks.

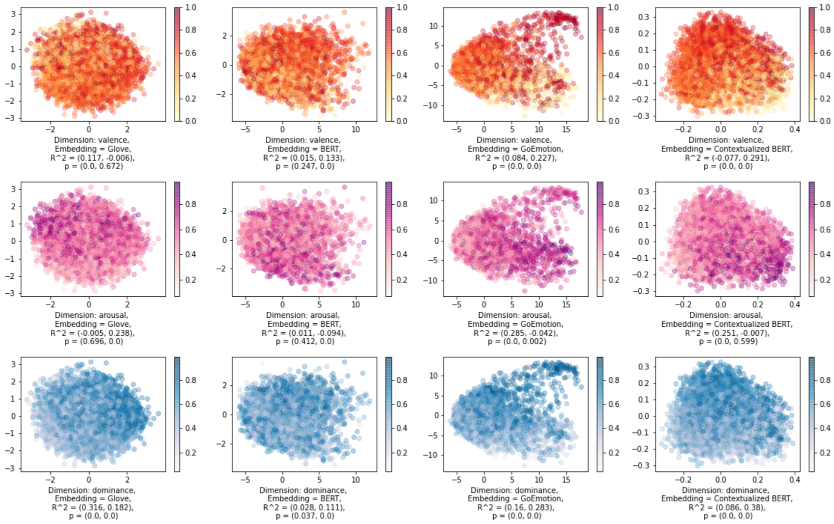

Fig.1 Two components PCA representations of Glove, vanilla BERT, BERT with GoEmotion, Contextualized BERT (Human ratings were color coded in a spectrum. Each dot represents a word. 5586 words are represented. The darker the dot, the more prominent the human rating in the respective dimension.)

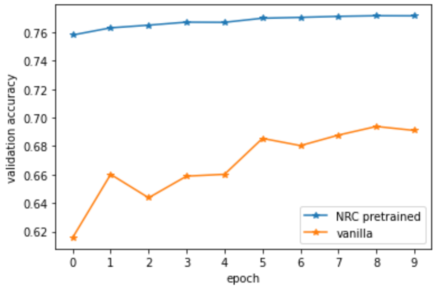

Fig.3 Validation accuracy of two models across 10 training epochs. Blue: the BERT model pretrained on human judgment dataset. Yellow: the vanilla BERT model.

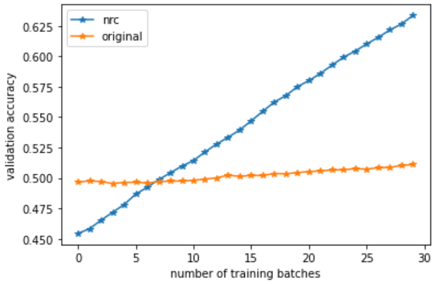

Fig.4 Performance of two models in the small-data regime. Each batch contains 32 training samples. Blue: the BERT model pretrained on human judgment dataset. Yellow: the vanilla BERT model.

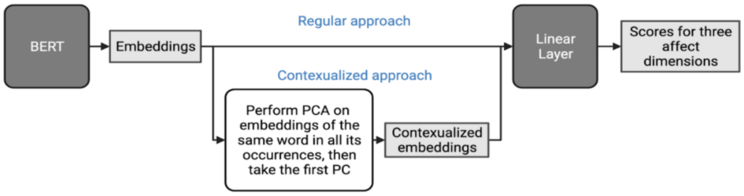
BERT
Embeddings
Regular approach
Contexualized approach
Perform PCA on embeddings of the same word in all its occurrences, then take the first PC
Contexualized embeddings
Linear Layer
Scores for three affect dimensions

Fig 2. Pipelines of two different approaches for extracting word embeddings. The regular approach is to extract the embedding from the last hidden layer in NLP models such as BERT. The contextualized approach is to perform PCA on embeddings of the same word in the IMDB dataset and take the first principal component of the multiple occurrences.

CORR (p)	VAD	Glove	BERT	GoEmotion BERT	Contextualized BERT
VAD	1.000 (.00)				
Glove	**0.272** (.00)	1.000 (.00)			
BERT	0.116 (.00)	0.148 (.00)	1.000 (.00)		
GoEmotion BERT	**0.252** (.00)	0.172 (.00)	0.013 (.47)	1.000 (.00)	
Contextualized BERT	**0.314** (.00)	**0.710** (.00)	**0.240** (.00)	**0.204** (.00)	1.000 (.00)

Table1. Spearman correlation coefficients (p value) of pairwise cosine similarities between each and the rest of word embedding types from human ratings(0-1) and four types of word embeddings. Correlation coefficients are in bold when larger than 0.2.

Pretrained Embeddings	Validation Accuracy			Affect Word Sample Accuracy		
	Valence	Arousal	Dominance	Valence	Arousal	Dominance
Glove	0.75	0.7	0.73	0.84	0.74	0.75
Context-free BERT	0.76	0.73	0.74	0.93	0.85	0.82
Contexualized BERT	**0.85**	**0.77**	**0.85**	**0.95**	**0.88**	**0.9**
BERT trained on GoEmotion	0.68	0.68	0.7	0.92	0.76	0.76

Table2. Performance of different word embeddings in predicting VAD Lexicon classification labels. This result comes from the linear classification probe model where the labels are the VAD binary classes. Validation accuracy is the prediction accuracy on validation set(2000 words) and affect word sample accuracy is the prediction accuracy on 130 affect words.