T. Luke Banaszak – tbanas2

CS410 FA22 – Technology Review

How Word2Vec changed NLP and how BERT Fixed its Shortcomings

Word2Vec and BERT ("bidirectional encoder representations from transformers") are major milestones on the NLP ("natural language processing") timeline. They both use neural networks to produce advanced language models with data structures more advanced than indices and include semantics. Reviewing how the two techniques work, and their sequence on the timeline, suggests possible scenarios for the next critical milestone while providing a summary of key concepts in current NLP.

Understanding Word2Vec begins with understanding embedding. In math, embedding refers to when some mathematical structure is contained within another. Word2Vec implements this mathematical concept by embedding words as -- not surprisingly -- vectors of real numbers. Word embedding is used at many stages in NLP projects, so by improving such a fundamental NLP component's accuracy, Word2Vec launched the field forward.

Unlike a traditional n-gram model, where words are represented just as indices in a vocabulary, words represented as n-dimensional vectors maintain semantic information and meaning. While it was likely many modern computer enthusiasts' first introduction to word embedding, Word2Vec was not the first language model to embed words as vectors. Research papers on vector word embedding first appeared early in text information system research (Salton 1975) and continued leading up to Word2Vec. Word2Vec, though, used the tools available in 2013 to significantly improve the technique, and its opensource code allowed for limitless refinement and application.

Again, while Word2Vec notably used neural networks to embed words, it was not the first to do so (Y. Bengio 2003). Word2Vec did, however, improve the method's accuracy so greatly that it propelled the field forward. The method's success leveraged larger corpuses via more computing power, general neural network learning advancements, and a novel training method. This training method uses the "Continuous Bag of Words" and "Skipgram" concepts. In sum, these are dynamic language modeling methods where context, via surrounding words, is used during training.

Word2Vec's output are vectors that capture semantics. Because of their similar context words, semantically related words have vectors with a low cosine value. Famously, the vectors' semantic accuracy also allows for algebraic functions (e.g., king-man = queen). These new vectors' accuracy advanced nearly all applications of NLP: speech recognition, next word prediction, etc.

A noteworthy shortcoming of Word2Vec is that it maintains only a single vector for a word. This is OK when comparing different words with similar semantics, but becomes problematic when a word has multiple meanings. For example, "bank" could mean the verb to knock something off another, or the noun for a business dealing with finance. The single vector for "bank" would have used these two sets of context words for a single input. Current state of the art language models address this problem. On the timeline of NLP milestones, BERT is such an example, and it appears a few years after Word2Vec with its improved context handling.

Word2Vec uses a word's surrounding words as context, but this context is effectively a "bag of words". In contrast, BERT understands a word's context within syntactic structures; BERT understands what the dynamic entities "it", "he", "they", etc. refers to within each unique sentence it processes. BERT produces high scores on NLP evaluation tests in-part because of this understanding. Traditionally, language models disregard these words as irrelevant, focusing instead on keywords, but BERT uses them and achieves greater understanding.

BERT is also a language model trained using neural networks. In particular, BERT was made possible by research on “attention” and, subsequently, “transformers”. These are complex concepts in machine learning. Suffice to say: the methods allow for more iterative and dynamic processing of an input sentence that enables capabilities like updating understanding of an earlier-observed pronoun after interpreting later words.

BERT remains state of the art in 2022. Better language models made available since BERT’s release in 2018 are largely optimizations rather than new techniques. GPT-3 regularly appears in attempt’s to find BERT’s closest competitor. Rather than the minimal performance difference between the models, comparisons usually focus on the differing ways they are trained, BERT’s ability for domain-specific tuning vs GPT-3’s all-purpose approach, and BERT’s opensource licensing through Google vs. GPT-3’s commercial licensing through Meta.

Word2Vec and BERT are both critical milestones in NLP research. They are both language models using machine learning and neural networks, and they both have advanced the use of a word’s context to improve accuracy. BERT, the newer method, is now nearly 5 years old. Transformers, the networks used by BERT, though, remain at the forefront of NLP research. The next milestone of BERT’s magnitude could be a transformer optimization, or, like how Word2Vec came many years after the first word embedding, new tools and computing power could make a previously theoretical technique a reality and create an entirely new approach.

References

Alammar, Jay. n.d. *The Illustrated Word2vec.* https://jalammar.github.io/illustrated-word2vec/.

Chiusano, Fabio. 2022. *A Brief Timeline of NLP from Bag of Words to the Transformer Family.* February 12. https://medium.com/nlplanet/a-brief-timeline-of-nlp-from-bag-of-words-to-the-transformer-family-7caad8bbba56.

Karani, Dhruvil. 2018. *Introduction to Word Embedding and Word2Vec.* September 1. https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec.

Muller, Britney. 2022. *BERT 101 - State of the Art NLP Model Explained.* March 2. https://huggingface.co/blog/bert-101.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. "A vector space model for automatic indexing." *Communications of the ACM .*

Shekhar, Gaurav. 2020. *Understanding GPT-3: OpenAI's Latest Language Model.* September 1. https://medium.com/swlh/understanding-gpt-3-openais-latest-language-model-a3ef89cffac2.

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space."

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in neural information processing systems.*

Wikipedia. n.d. *Embedding.* https://en.wikipedia.org/wiki/Embedding.

—. n.d. *Word embedding.* https://en.wikipedia.org/wiki/Word_embedding.

Y. Bengio, R. Ducharme, P. Vincent. 2003. "A neural probabilistic language model." *Journal of Machine Learning Research.*