Efficient Estimation of Word Representations in Vector Space (Mikolov et al., 2013)

Critique by Jingying Xu

September, 2025

Mikolov et al. (2013) developed the Word2Vec model, which employs the Continuous Bag of Words (CBOW) and Skip-grams (SG) algorithms to represent word vectors. Despite its influential contribution, the paper has several key limitations:

1. *Single vector per word.*

   The model assigns a single static vector to each word, which would conflate different senses a word (e.g., *bank* as "financial institution" vs. "riverbank") and metaphorical uses (*drive a car* vs. *drive me mad*). A potential improvement would be to cluster contextual embeddings by sense $bank_1$, $bank_2$, and assign tokens to the nearest cluster based on the specific usage.

2. *Register effects.*

   The underlying assumption of the model is that words with similar meanings occur in similar contexts. While it works well with words that often occur in the same register (e.g., *spinach* vs. *bok choy*), it might fail with words that are semantically equivalent but always used in different registers (e.g., *pee* (informal) vs. *urinate* (formal)). These words almost never share overlapping environments, thus might be separated far away by the model.

3. *Limitations of the analogy test.*

   The main evaluation method adopted by the paper is the word analogy test. While the analogy test can detect categorical relations, such as A IS THE CAPITAL OF B and *A is the past tense of B*, it cannot well capture the graded or continuous semantic relations such as *warm* vs. *hot* vs. *scorching*. Thus the evaluation does not reflect the real performance of the model. Although the MSR Sentence Completion task partially addresses analogy's weaknesses, it's also limited by the multiple-choice format. More comprehensive, continuous evaluations should be adopted.

References

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781