A Neural Probabilistic Language Model (Bengio et al., 2003)

Critique by Jingying Xu

September, 2025

Bengio et al. (2003) developed a neural network model that learns continuous word embeddings for next-word prediction. The model marked a significant shift from discrete n-gram models, yet several weaknesses remain evident:

1. *Limitation with long-distance dependencies*

   The model keeps a fixed n-gram window, which saves computational cost, but suffers from capturing hierarchical syntactic relations. Although the mode can detect certain local dependencies relations (e.g., *Det + Noun*: *the student, a cat*), it fails to capture long-distance syntactic dependencies, such as long-distance *wh*-movement as in (1), where the direct object of *eat* is moved to the sentence-initial position and interpreted non-locally.

   (1) What$_i$ did the smart girl say that the funny guy ate t_i?

2. *Uneven load in the parameter-parallel training.*

   The parameter-parallel strategy distributes output-layer computations across CPUs. However, it is not clear how the clusters would be built. If word clusters are randomly assigned, processors that happen to be assigned words with extremely high frequencies would have to handle much more updates than others. This would causes uneven computational loads: the busiest CPU slows the entire system.

3. *Evaluation via perplexity.*

   The model performance is evaluated with perplexity, which only checks whether the generated sentences match the observed local patterns in the data. It does not necessarily assess grammaticality or semantic acceptability. As a result, the model might assign high probability to locally frequent but ungrammatical constructions such as:

   (2) The keys to the cabinet **is** in the room.

References

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, *3*(null), 1137–1155.