Efficient Estimation of Word Representations in Vector Space

Introduction

Traditional NLP models treat words as independent tokens without capturing similarities between them. This paper introduces **efficient methods** for learning **continuous word vector representations** from large datasets, using simplified architectures that achieve **high-quality embeddings** with significantly **lower computational cost** compared to earlier neural language models.

Main Goals

- To learn **high-dimensional word embeddings** from datasets containing **billions of words** and vocabularies with **millions of words**.
- To develop **new architectures** (CBOW and Skip-gram) that scale better and preserve **semantic and syntactic regularities** in word relationships.
- To evaluate the embeddings using a **comprehensive test set** covering various semantic and syntactic analogy tasks.

Model Architectures

1. Feedforward NNLM

A traditional architecture using projection and hidden layers; computationally expensive due to dense operations and large output layers.

2. Recurrent NNLM (RNNLM)

Captures long-term dependencies, but expensive to train due to recurrent connections.

3. New Efficient Models

- **CBOW (Continuous Bag-of-Words):** Predicts a target word from its surrounding context using a shared projection layer without a hidden layer.
- **Skip-gram:** Predicts surrounding words given a target word; performs better on semantic tasks and captures long-range dependencies.

Both use **hierarchical softmax** for efficiency and are trained using **stochastic gradient descent**.

Results and Evaluation

- A new **semantic-syntactic analogy test set** was developed to evaluate word vectors using vector arithmetic (e.g., *king - man + woman ≈ queen*).
- Experiments show:
  - **Skip-gram** performs best on **semantic relationships**.
  - **CBOW** performs best on **syntactic tasks**.
  - Accuracy improves with **larger vector size and more training data**, but with diminishing returns.
- Compared to other models (RNNLMs, NNLMs), the proposed methods are **faster** and **more scalable**.

Applications

The embeddings can improve:

- Machine translation
- Sentiment analysis
- Question answering
- Knowledge base extension

An example task, the **Microsoft Research Sentence Completion Challenge**, shows that combining Skip-gram vectors with RNNLM scores achieves **state-of-the-art accuracy** (58.9%).

Conclusion

This paper demonstrates that **simple and efficient architectures (CBOW and Skip-gram)** can learn high-quality word embeddings from massive datasets. These vectors capture **linguistic regularities** and significantly outperform existing models in both quality and training speed. The models pave the way for applying word vectors in many **large-scale NLP tasks**.