UNIVERSITE ABDELMALEK ESSAADI

Master AI and Data Science

Atelier 2 : Word Embedding

Année Universitaire: 2023/2024

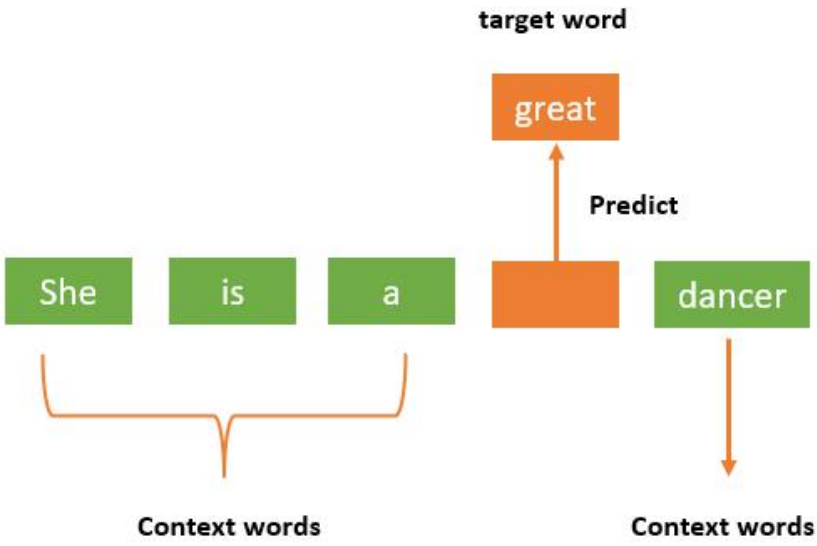
REALISE PAR :
BAKKALI AYOUB.

.

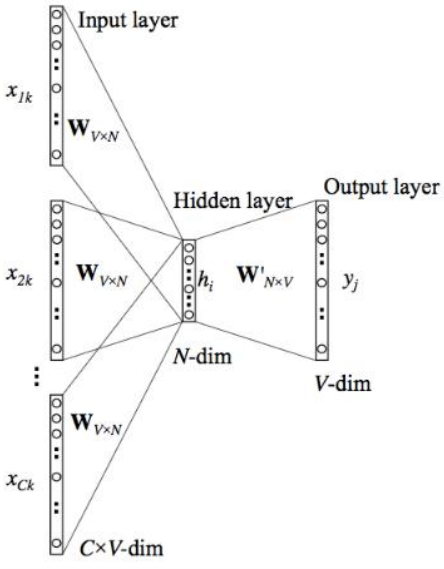
CBOW :

Continuous Bag of Words (CBOW) is a type of word embedding model used in natural language processing (NLP) tasks. It's one of the architectures used in Word2Vec, a popular method for learning word embeddings from large corpora of text.

The main idea behind CBOW is to predict the target word based on its surrounding context words within a fixed window size. Unlike Skip-gram, which predicts context words given a target word, CBOW predicts a target word given its context.



CBOW is trained by adjusting the weights of the neural network to minimize a loss function, such as cross-entropy loss, between the predicted probability distribution and the actual distribution of words.

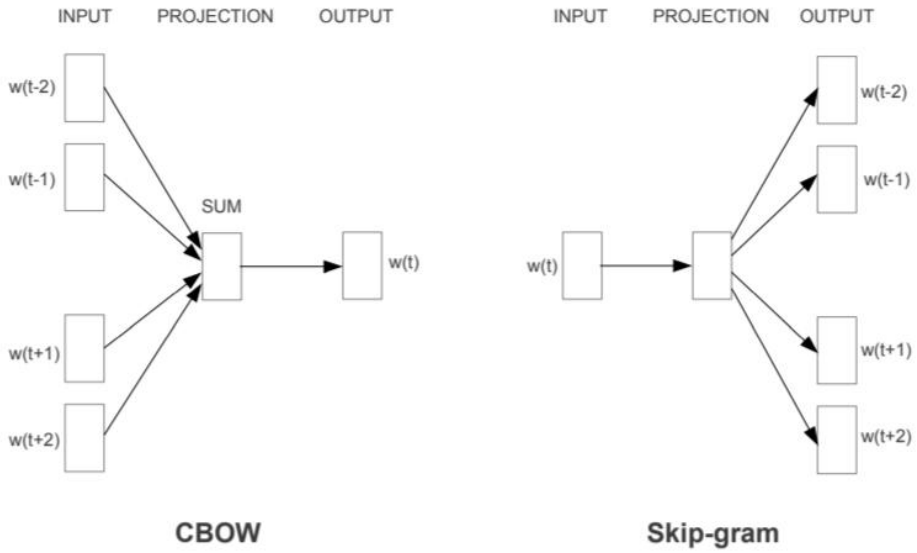


Once trained, the hidden layer weights, which represent the word embeddings, are used as the output of the model. These word embeddings capture semantic relationships between words, allowing similar words to have similar vector representations in the embedding space.

Skip-gram

The skip-gram model is a neural network architecture that aims to predict the context words (words surrounding a target word) given a center word. In this model, each word in the corpus is represented as a unique vector in a high-dimensional space. The skip-gram model takes a target word as input and tries to predict the probability of context words occurring around it. It does so by training a neural network to maximize the likelihood of predicting context words based on the target word.

The main difference between skip-gram and CBOW lies in their input-output relationships and the way they are trained. Skip-gram predicts context words given a target word, while CBOW predicts a target word given its context.

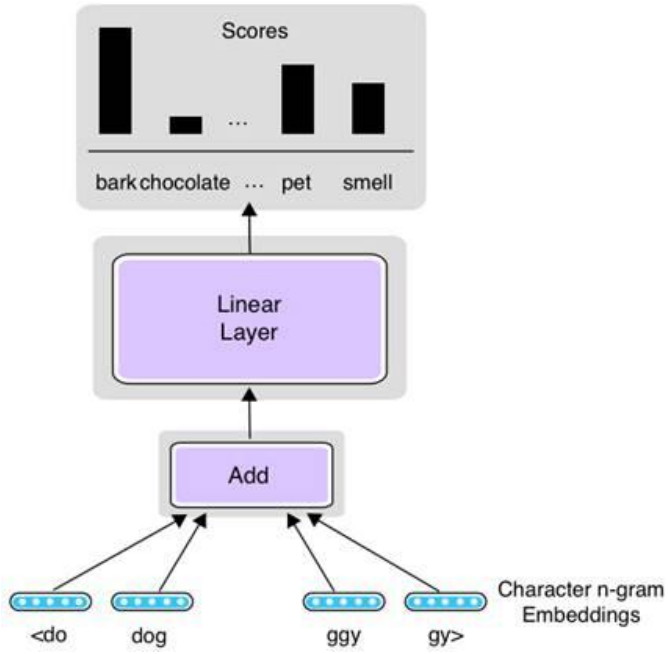


FastText

FastText is a word embedding and text classification library developed by Facebook's AI Research lab (FAIR). It is designed to efficiently train and generate word embeddings, which are dense vector representations of words in a continuous vector space. FastText offers several key features and innovations that distinguish it from traditional word embedding methods like Word2Vec

it breaks down words into smaller components, such as character n-grams, to capture morphological information and handle out-of-vocabulary words effectively.

For example, the word “fast” might be represented by the character n-gram “fa”, “as”, “st”, and “fast” itself.



Conclusion

Upon visualizing the embeddings of vocabulary derived from different models, such as FastText CBOW and skip-grams, using the t-distributed stochastic neighbor embedding (t-SNE) algorithm, it becomes evident that the resulting visualizations portray a striking similarity. Despite the variance in training methodologies—CBOW focusing on predicting the target word from its context, while skip-gram predicts the context from the target word—the clusters formed in the visualizations exhibit a notable resemblance. This alignment suggests that both models effectively capture semantic relationships within the vocabulary, as evidenced by the cohesive groupings of semantically related words. While slight discrepancies in cluster positioning may be observed, the overarching structure and inter-word associations remain consistent across the embeddings produced by the different models.