FastText

Enriching Word Vectors with Subword Information (Piotr Bojanowski, Edouard Grave et al., 2017)

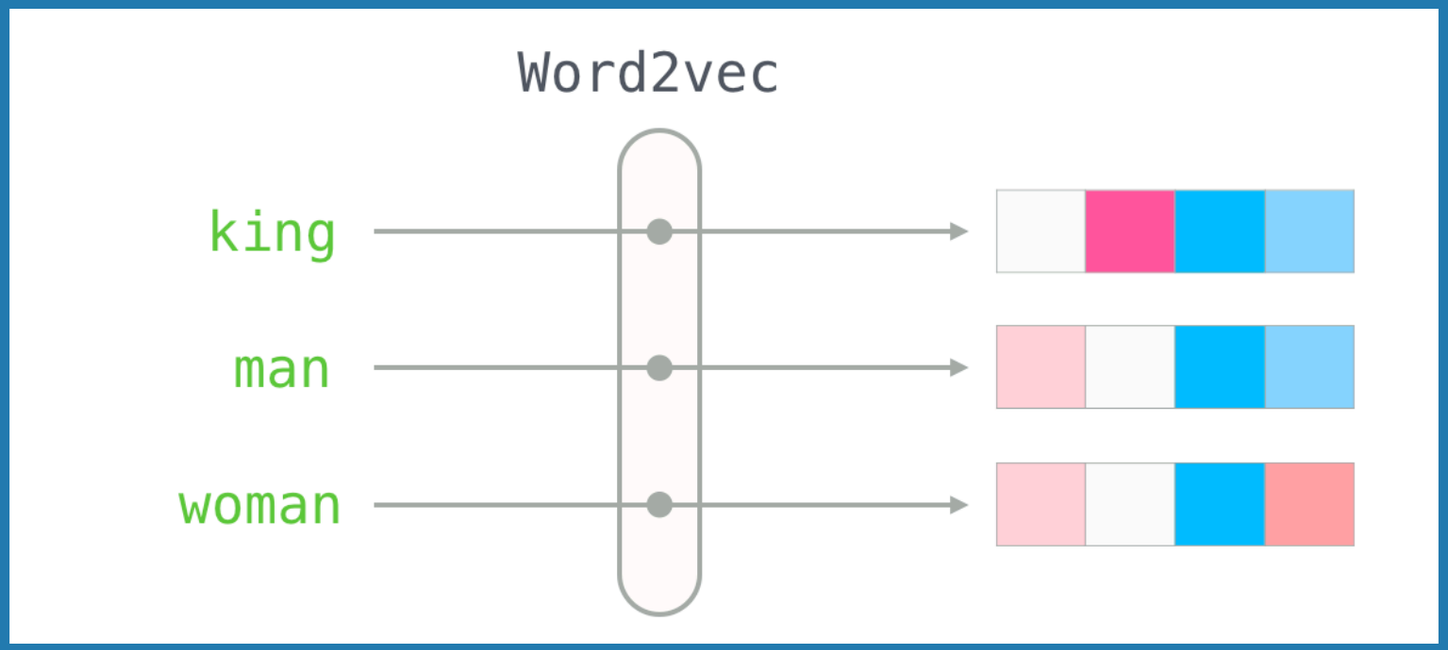
Hannah Do | Mar 5th, 2022

Library for efficient text classification and representation learning

Open-source, free, lightweight library that allows users to learn text representations and text classifiers.

- **Word embedding** : Representation for text where words that have same meaning have similar representation.
- **Subword-focused** : Can extract information of a word at a character level
  - 안녕 —> ㅇ, ㅏ, ㄴ, ㄴ, ㅕ, ㅇ
- **Faster** than CBoW or Skip-gram model in learning information, but high memory requirement

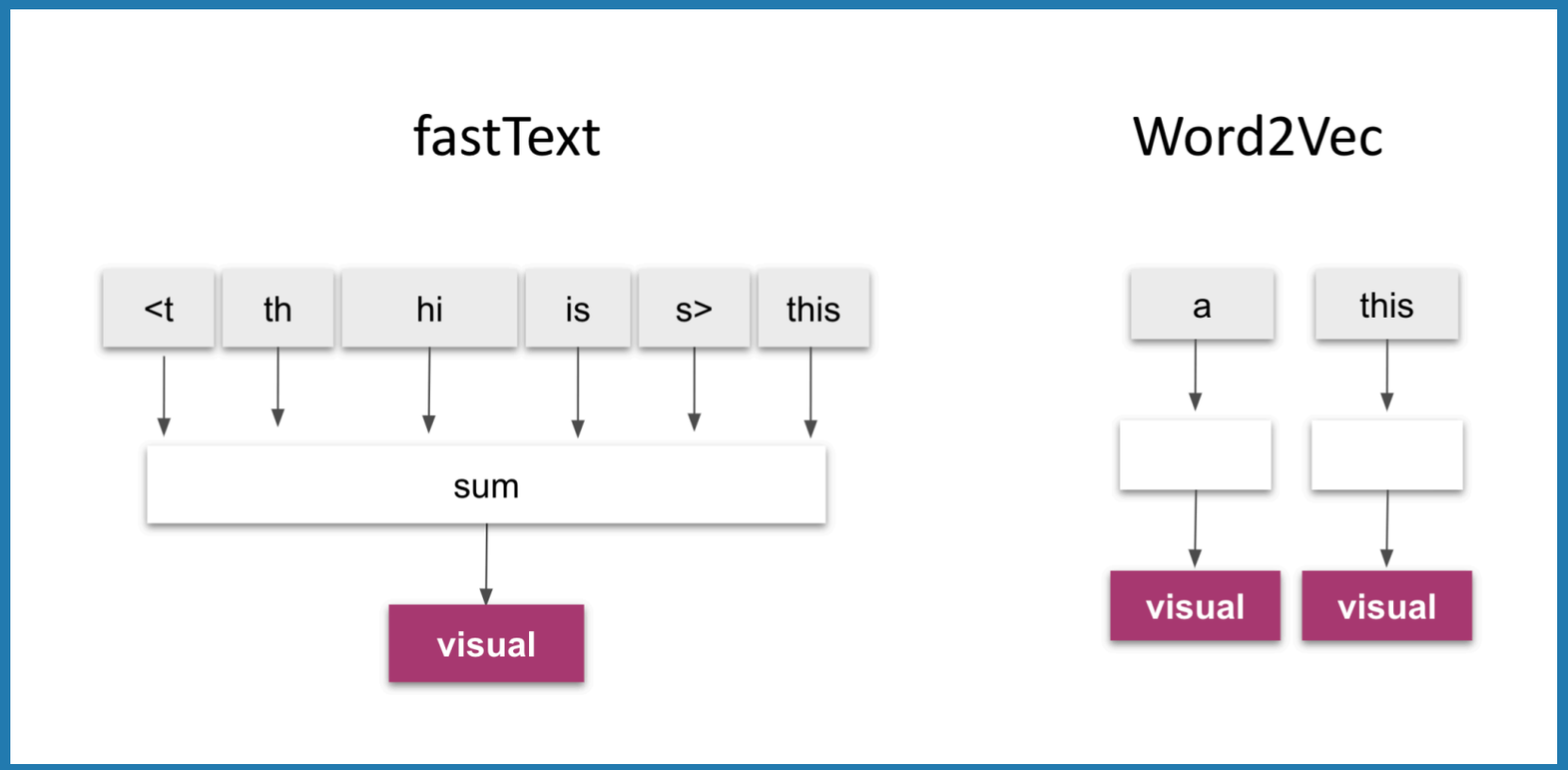
Previous Word Representation / Embedding method



https://jalammar.github.io/illustrated-word2vec/

Unable to represent words that do not appear in the training dataset.

FastText Word Representation/ Embedding method



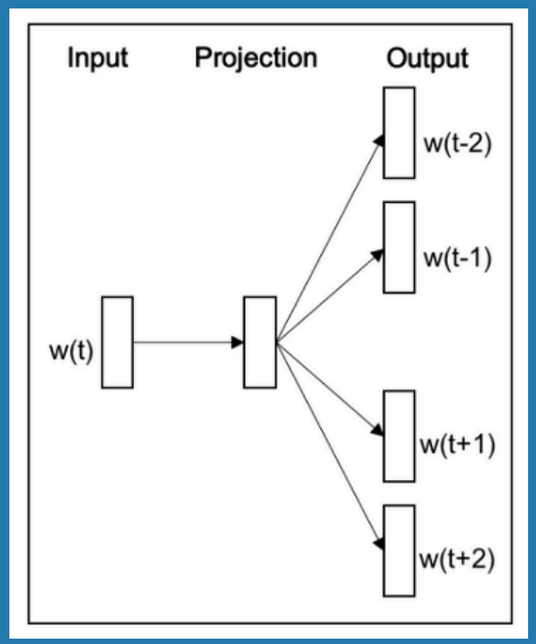
FastText operates at a character level, instead of a word level.

https://kavita-ganesan.com/fasttext-vs-word2vec/#.YiKhFBNBxYw

Model Morphology

Skip-gram + Negative Sampling

Predicting context words from the center word.



Updating the probability scores of the context words while the window slides over each center word in the sentence.

https://wooono.tistory.com/244

Maximizing the similarity of the words in the same context and minimizing it when occurring in different contexts.

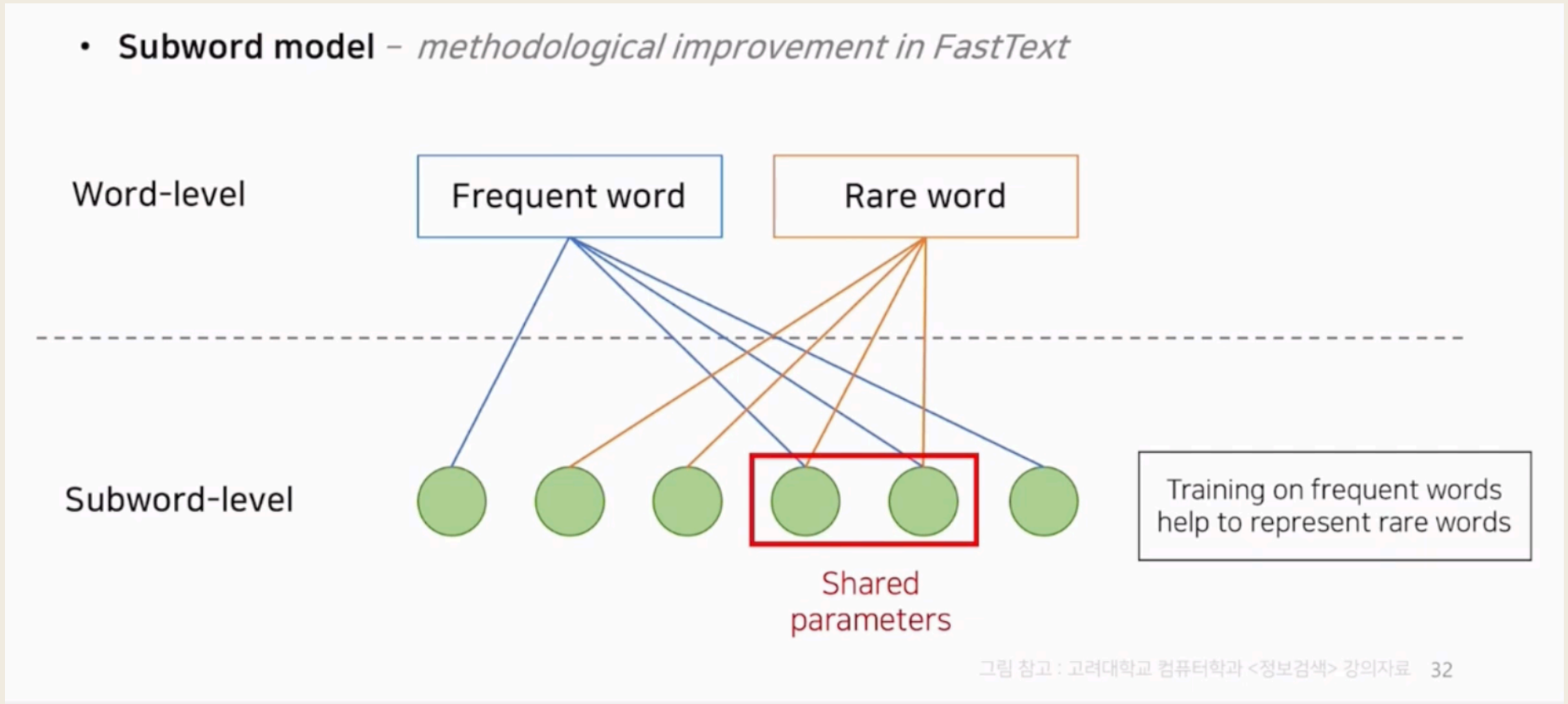
$$\sum_{t=1}^{T}\left[\sum_{c\in\mathcal{C}_t}\ell(s(w_t,\ w_c))+\sum_{n\in\mathcal{N}_{t,c}}\ell(-s(w_t,\ n))\right]$$

* $\ell : x \mapsto \log(1+e^{-x})$

1. Randomly sample handful of words (2 <= k <= 20)
2. Determine **positive example** (actual context word around target word) & **negative example** (context word chosen as a possibility, not from training data)
3. Binary classification on the different examples (actual context word as 1, opposite as 0) –> lower computational cost

https://www.youtube.com/watch?v=7UA21vg4kKE

Subword model



Since **FastText** provides subword model, it provides character or n-gram level embedding - allowing rare words to share semantic meaning with other words, along with bringing lower computational cost.

References

1. **Enriching Word Vectors with Subword Information** Piotr Bojanowski, Edouard Grave et al. (2017)

2. **FastText vs. Word2vec: A Quick Comparison** Kavita Ganesan https://kavita-ganesan.com/fasttext-vs-word2vec/#.YiK4xBNBxYx

3. **[Paper Review] FastText: Enriching Word Vectors with Subword Information** 고려대학교 산업경영공학부 DSBA 연구실 (김수빈) - https://www.youtube.com/watch?v=7UA21vg4kKE

4. **[DL] Word2Vec, CBOW, Skip-Gram, Negative Sampling** 우노 AI/Deep Learning - https://wooono.tistory.com/244