Prediction-Based Text Vectorization

Indonesia AI

OBJECTIVE & OUTLINE

Prediction-Based Text Vectorization

Objektif: Memahami konsep dari Prediction-Based Text Vectorization dalam NLP

Outline:

1. Konsep Prediction-Based Text Vectorization
2. Word2Vec
3. CBOW
4. Skip-Gram

Apa itu Prediction-Based Text Vectorization?

PREDICTION-BASED TEXT VECTORIZATION

Metode dalam text vectorization yang menggunakan **model prediksi bahasa** untuk menghasilkan **representasi numerik dari teks.**

PREDICTION-BASED TEXT VECTORIZATION

Kelebihan Prediction-Based Text Vectorization

- Menangkap konteks dan makna
- Memperhitungkan hubungan semantik
- Mengatasi masalah urutan kata

PREDICTION-BASED TEXT VECTORIZATION

Kekurangan Prediction-Based Text Vectorization

- Membutuhkan jumlah data yang besar
- Memerlukan waktu dan sumber daya komputasi
- Keterbatasan dalam menangkap makna kata dengan banyak arti

PREDICTION-BASED TEXT VECTORIZATION

Macam Prediction-Based Text Vectorization

- **Word2Vec**
- GloVe (Global Vectors for Word Representation)
- FastText
- ELMo (Embeddings from Language Models)
- BERT (Bidirectional Encoder Representations from Transformers)
- GPT (Generative Pre-trained Transformer)

Any question guys ~

Word2Vec

Indonesia AI

WORD2VEC

Sebuah metode yang digunakan untuk menghasilkan **representasi vektor** kata dalam bentuk numerik berdasarkan **konteks kata** dalam sebuah teks

WORD2VEC

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

https://arxiv.org/pdf/1301.3781.pdf

Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

Greg Corrado
Google Inc.
Mountain View
gcorrado@google.com

Jeffrey Dean
Google Inc.
Mountain View
jeff@google.com

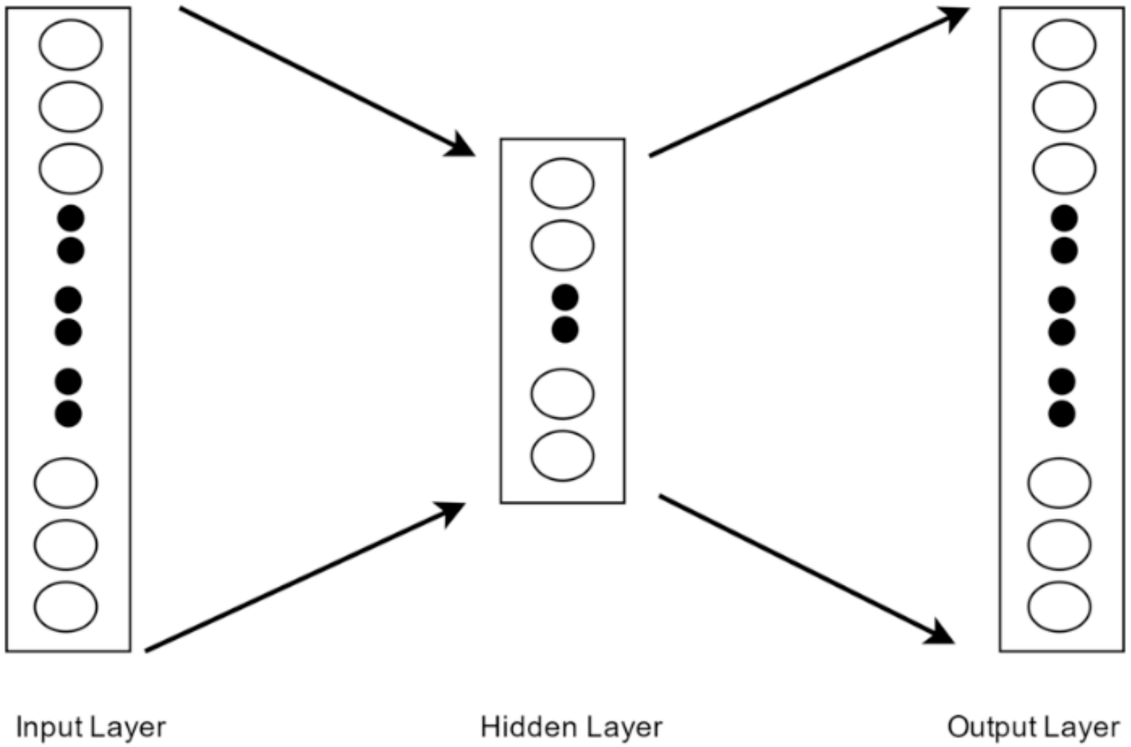
Abstract

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada" and "Air" cannot be easily combined to obtain "Air Canada". Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible.

https://arxiv.org/pdf/1310.4546.pdf

WORD2VEC



Word2Vec menggunakan **neural network** untuk **mendapatkan vektor** tersebut. **Arsitektur Word2vec** hanya terdiri dari **3 layer** yaitu **Input**, **Projection**(Hidden Layer), dan **Output**. Input pada Word2vec berbentuk one-hot encoded vector dengan panjang = jumlah kata unik pada data training.

PREDICTION-BASED TEXT VECTORIZATION

Word2Vec

- CBOW(Continuous Bag-of-Words)
- Skip-Gram

WORD2VEC

Kelebihan

- Representasi semantik
- Generalisasi

Kekurangan

- Kurang menangkap makna yang kompleks
- Sensitif terhadap frekuensi kata

Any question guys ~

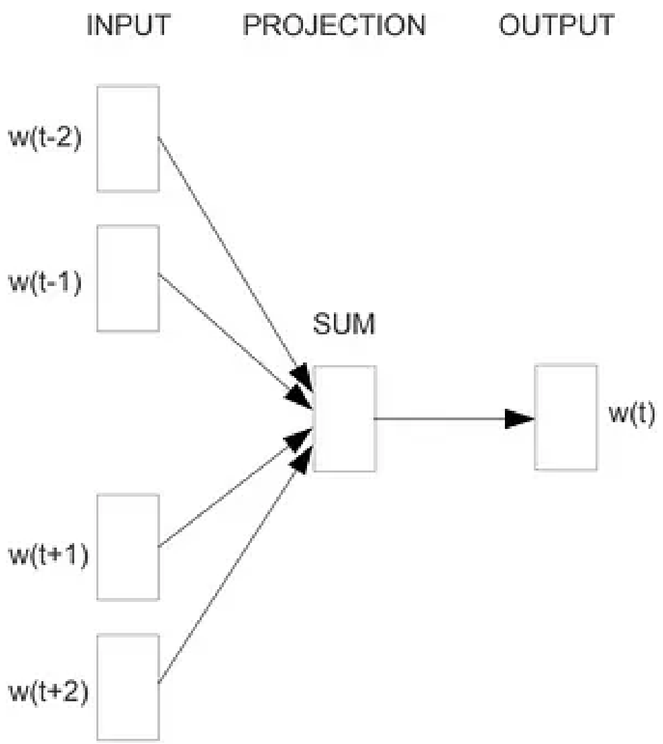
CBOW

Indonesia AI

CBOW

CBOW belajar **memprediksi kata target** dengan **memanfaatkan** semua **kata di sekitarnya**

CBOW



untuk memprediksi kata (output) ketika diberikan konteks disekitar kata tersebut (input).

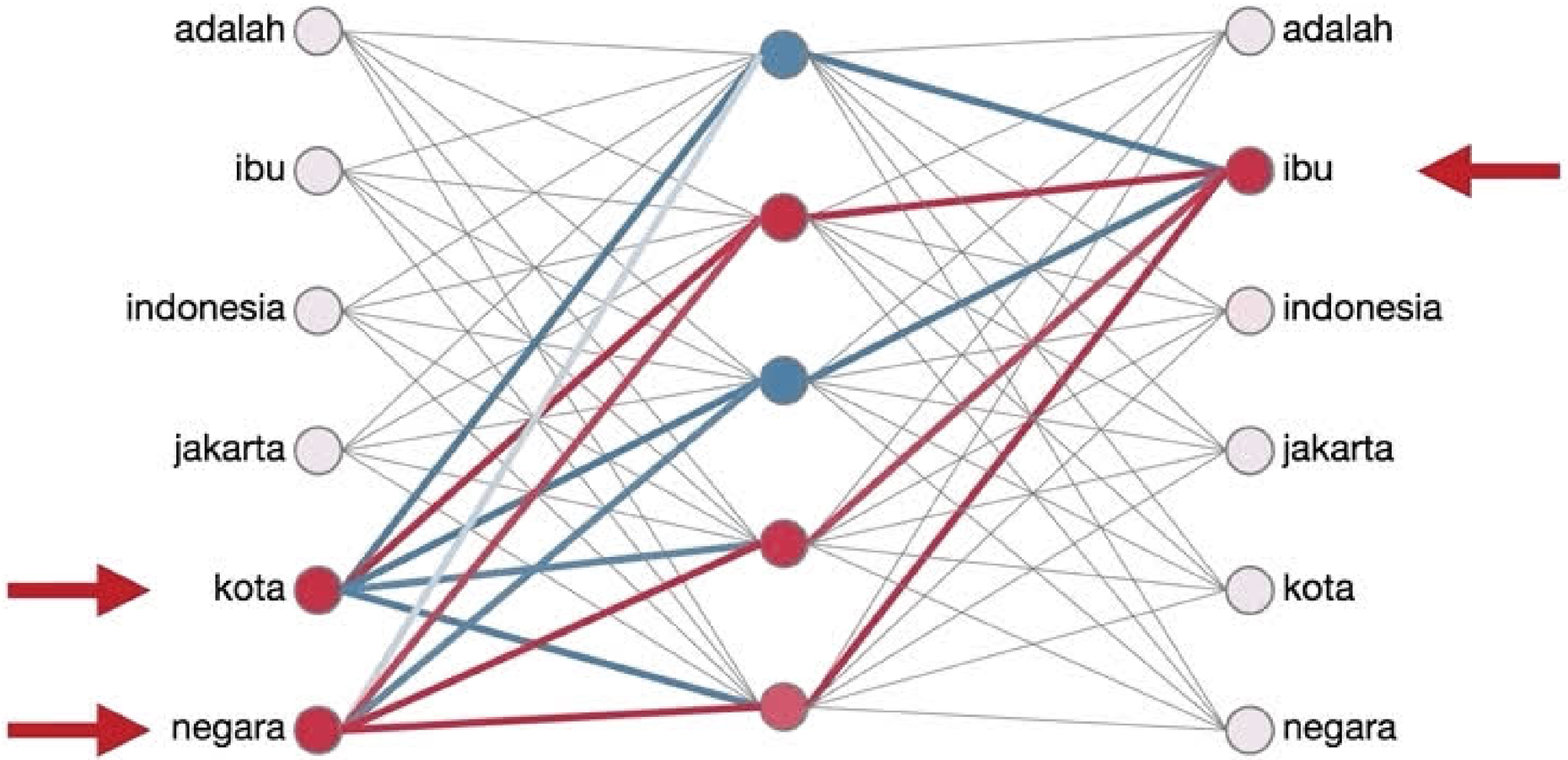
CBOW

"Ibu kota Negara Indonesia adalah Jakarta"

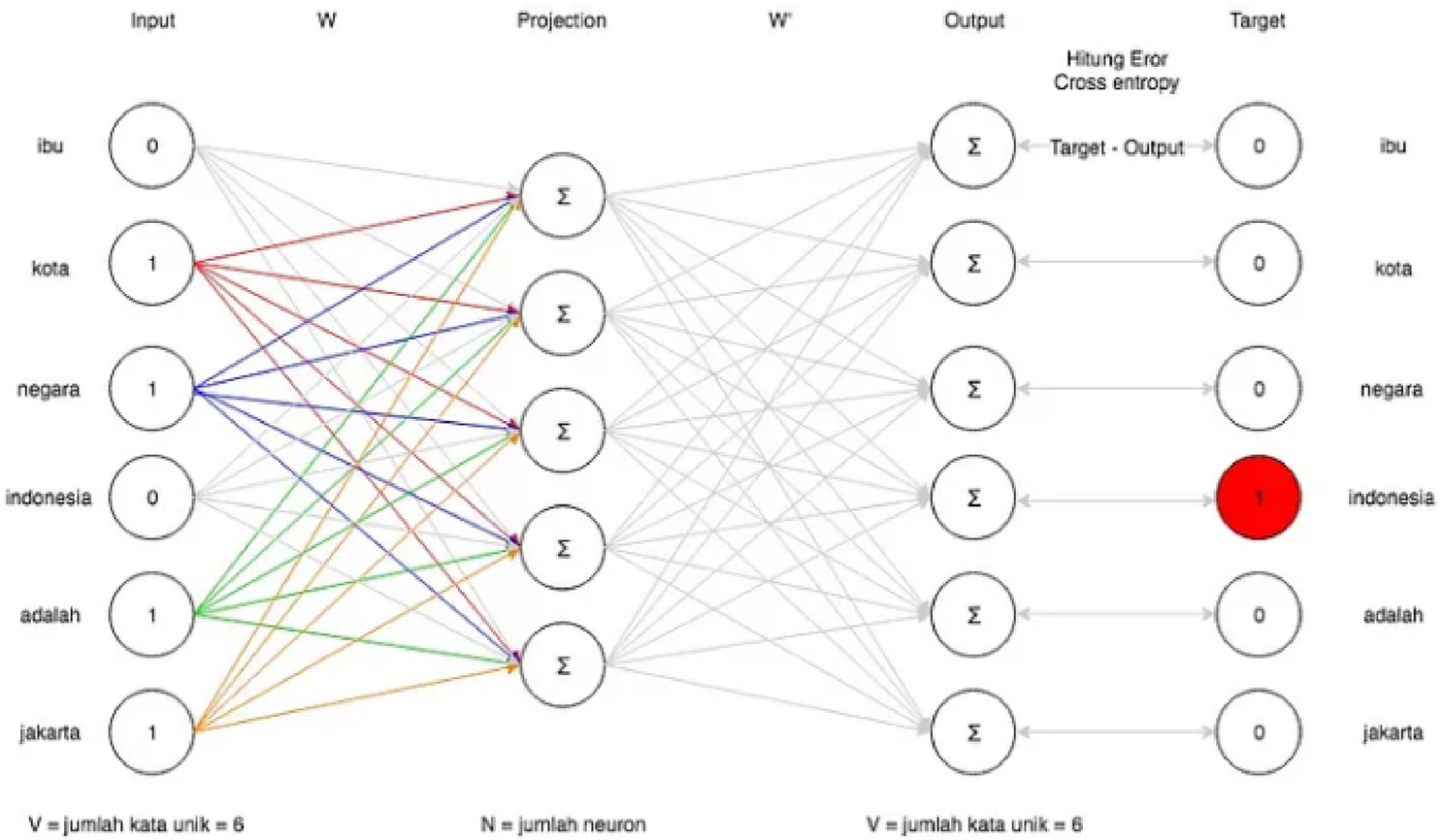
window = 2

CBOW

CBOW


adalah
ibu
indonesia
jakarta
kota
negara
adalah
ibu
indonesia
jakarta
kota
negara

CBOW


Input
W
Projection
W'
Output
Target
Hitung Eror
Cross entropy
ibu
kota
negara
indonesia
adalah
jakarta
0
1
1
0
1
1
Σ
Target - Output
V = jumlah kata unik = 6
N = jumlah neuron
V = jumlah kata unik = 6

CBOW

Kelebihan

- Lebih cepat dalam pelatihan
- Representasi Kata yang Umum

Kekurangan

- Tidak Memperhatikan Urutan Kata
- Tidak Efektif untuk Kata-Kata Langka

Any question guys ~

SKIP-GRAM

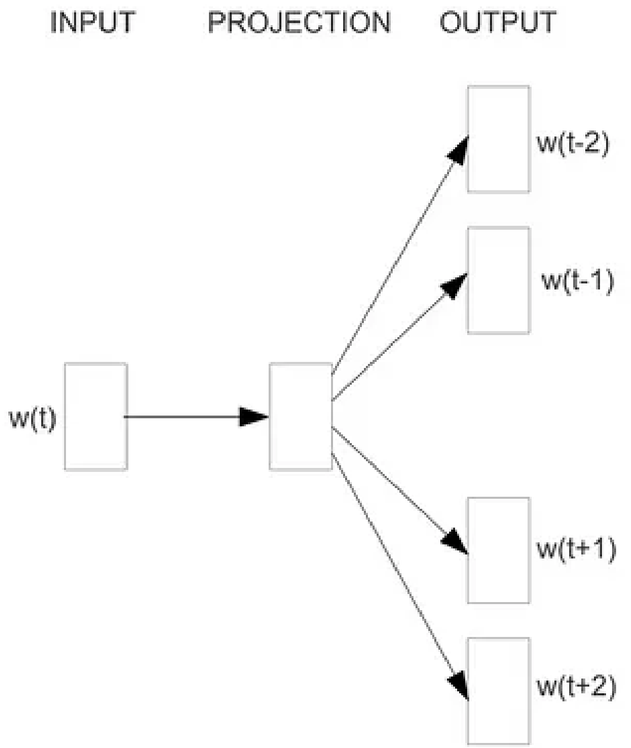
Indonesia AI

SKIP-GRAM

Skip-Gran memprediksi **kata-kata sekitar kata target** dan mempelajari bagaimana kata **target berhubungan dengan kata-kata di sekitarnya** untuk memperoleh representasi vektor kata yang kaya dengan **informasi kontekstual**

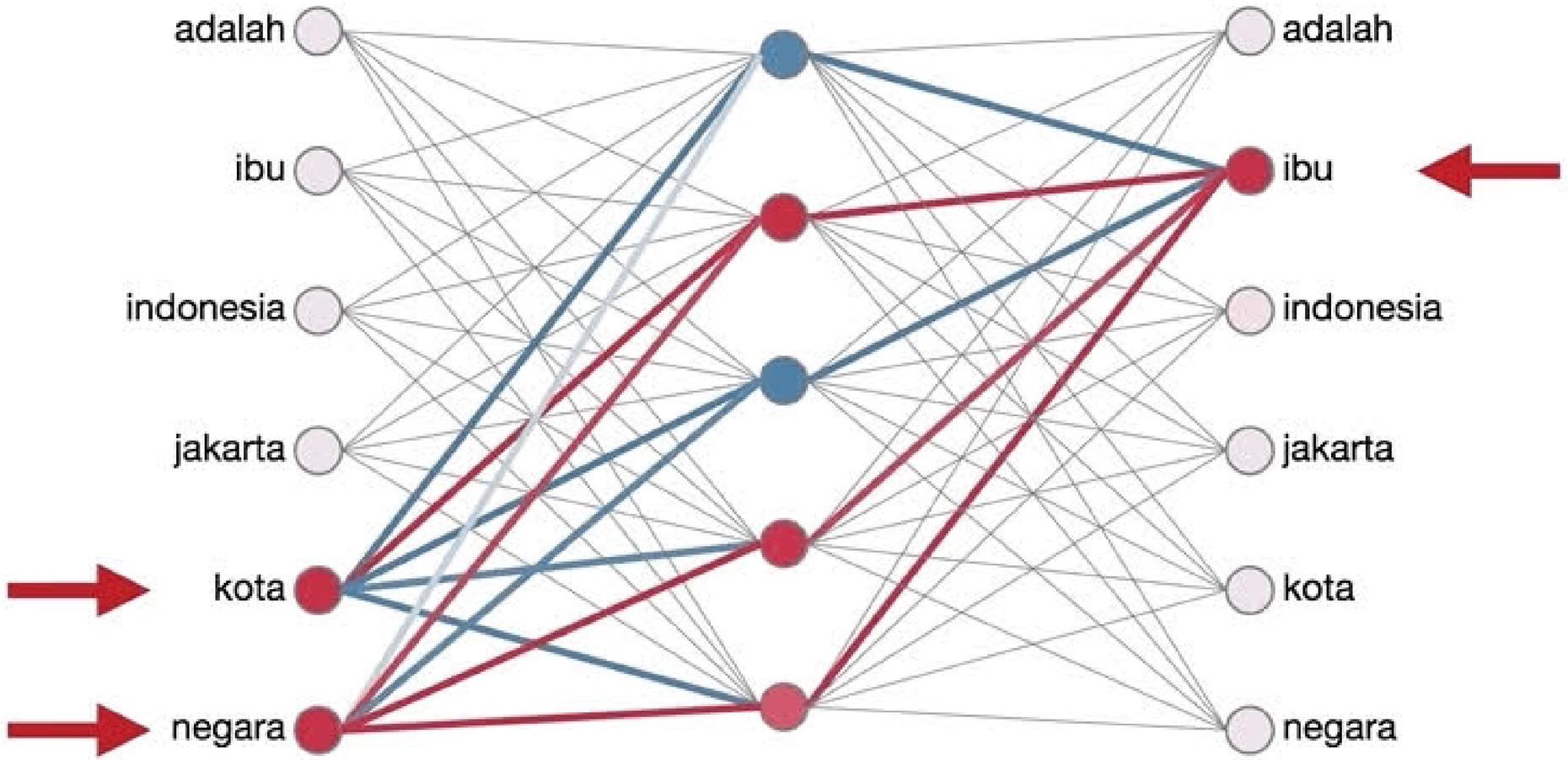
SKIP-GRAM

untuk memprediksi konteks (output) di sekitar current word (input)

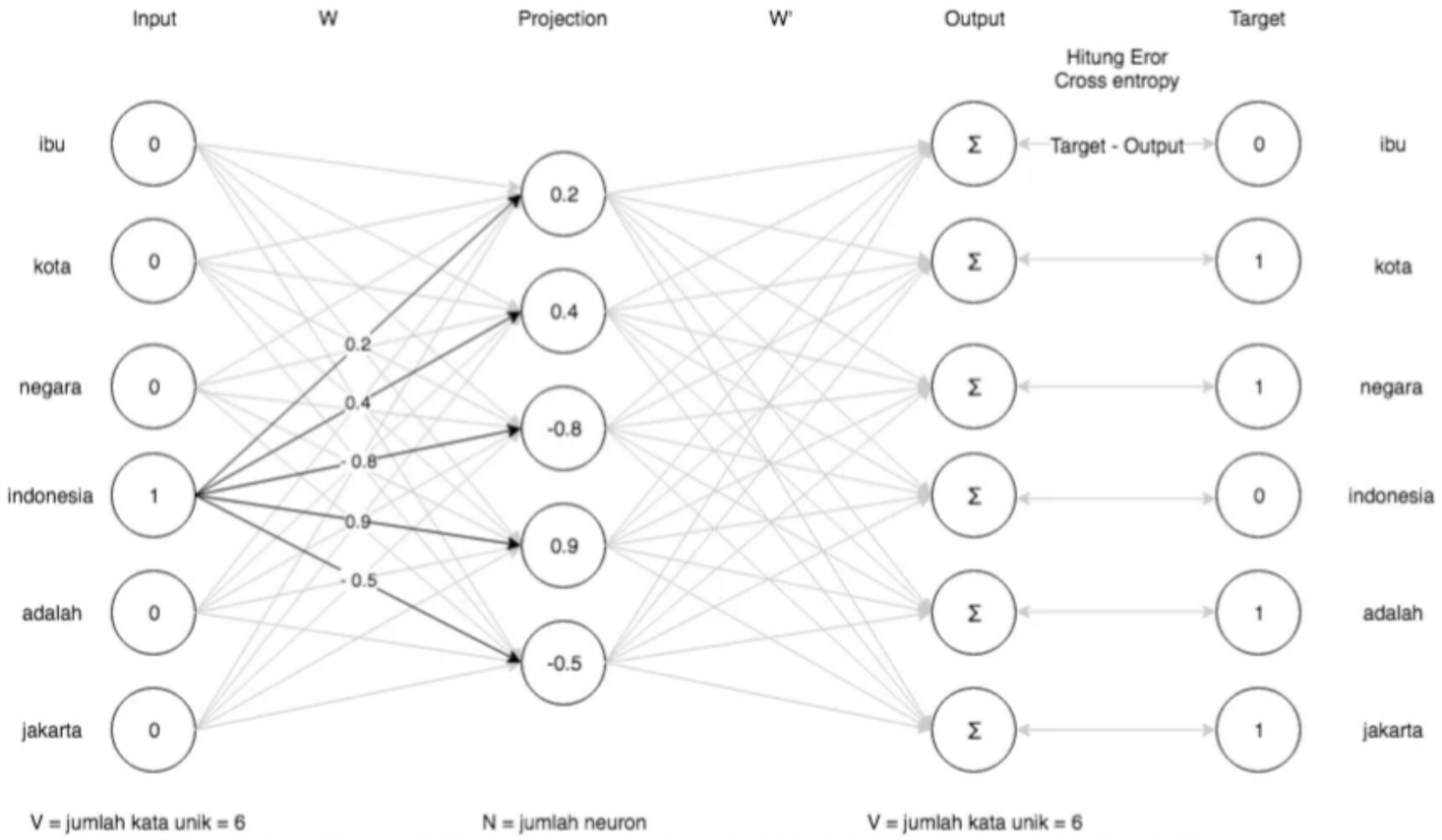


SKIP-GRAM

SKIP-GRAM



SKIP-GRAM


Input
W
Projection
W'
Output
Target
Hitung Eror
Cross entropy
ibu
kota
negara
indonesia
adalah
jakarta
0
0
0
1
0
0
0.2
0.4
-0.8
0.9
-0.5
Σ
Target - Output
0
1
1
0
1
1
ibu
kota
negara
indonesia
adalah
jakarta
0.2
0.4
- 0.8
0.9
- 0.5
V = jumlah kata unik = 6
N = jumlah neuron
V = jumlah kata unik = 6

SKIP-GRAM

Kelebihan

- Meningkatkan kemampuan dalam menangkap sinonim dan hubungan semantik antara kata-kata.
- Memperoleh representasi vektor kata yang lebih baik untuk kata-kata jarang muncul dalam dataset.

Kekurangan

- Memiliki waktu pelatihan yang lebih lama karena memproses setiap kata secara individual
- Rawan terhadap kata yang jarang muncul

Any question guys ~

Terima Kasih!