University | School of
of Glasgow | Computing Science

Encodings Inspired from NLP

Dr. Fani Deligianni,

fani.deligianni@glasgow.ac.uk

Lecturer (Assistant Professor)

Lead of the Computing Technologies for Healthcare Theme

https://www.gla.ac.uk/schools/computing/staff/fanideligianni

THE AWARDS 2020

UNIVERSITY OF THE YEAR

WORLD CHANGING GLASGOW

NLP vs Processing of EHR

- Irregular sequence of words/EHR with relation to when they are generated

- Dependencies between words that are located close to each other

- Long-term dependencies are also present

- Representing the underline information in a consistent way to be processed via deep learning is challenging

NLP Approaches

- Co-occurrence models measure the frequency of words occurring alongside
  - Capture global dependencies and context
  - Suffer from the curse of dimensionality
- Neural **embeddings**
  - Predictive models that capture local dependencies
  - Learn dense word representations from unstructured text data

Neural Embeddings

300 dimensional continuous representation

100k dimensional (one hot encoded)
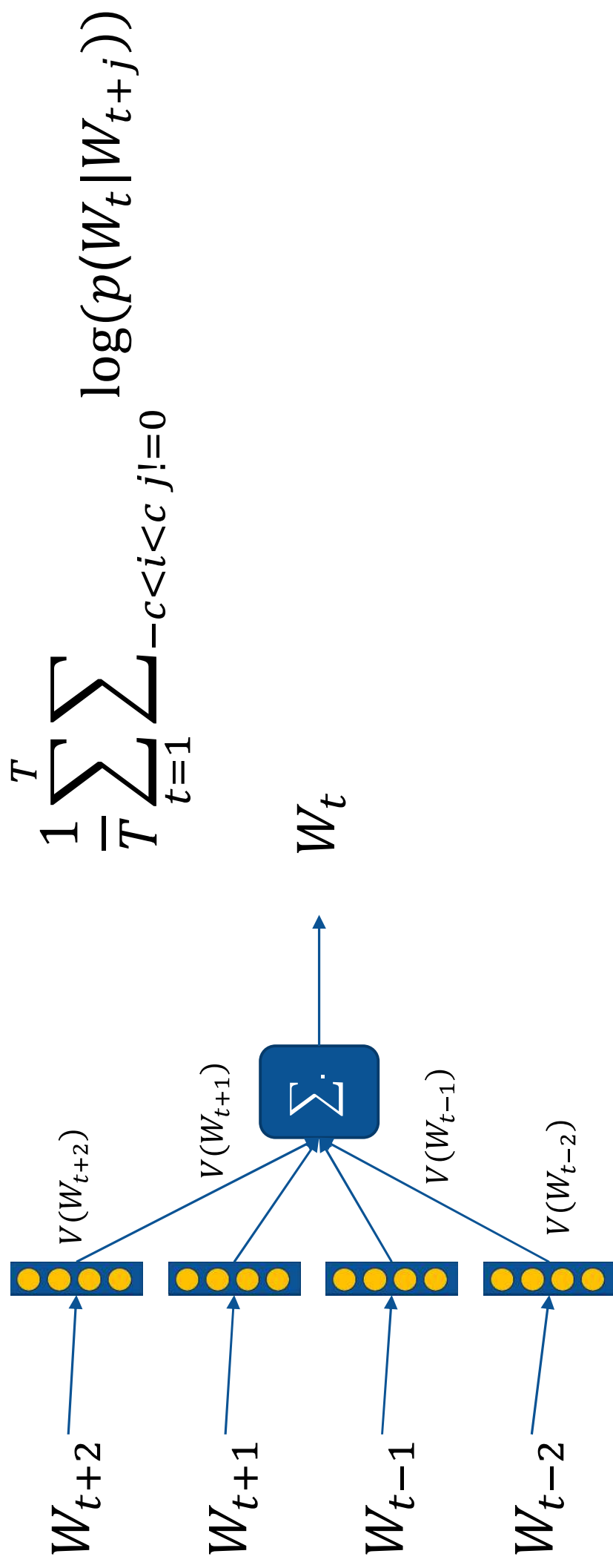
| 0 | 1 | 0 | 0 | 0 | 1 |
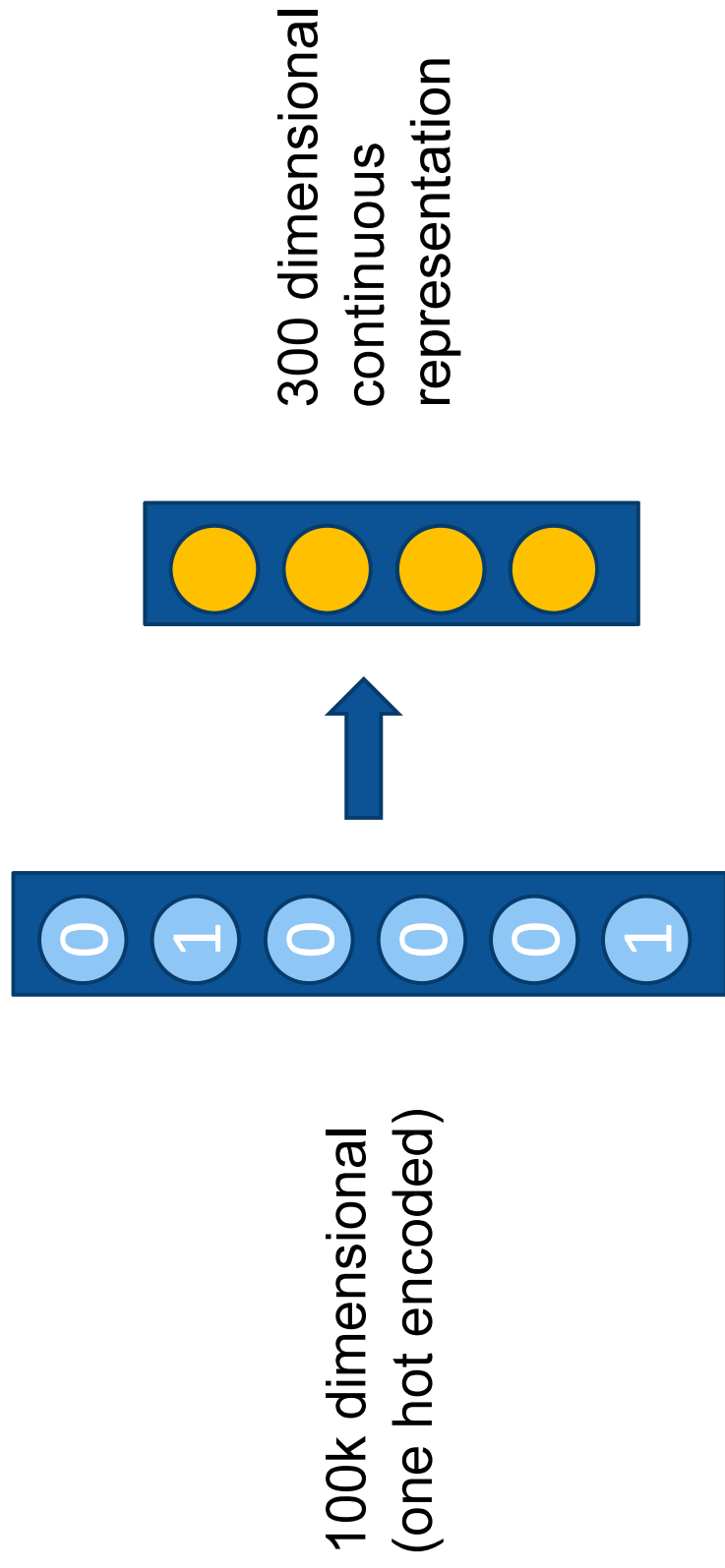
$$v(queen) = v(king) - v(man) + v(woman)$$

Common NLP Embeddings

- Word2vec models
  - Continuous bag-of-words
  - Skip-gram models
- Global and Vector (GloVe)

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<i<c\ j!=0} \log(p(W_t|W_{t+j}))$$

W_{t+2}

$V(W_{t+2})$

W_{t+1}

$V(W_{t+1})$

$\sum \cdot$

W_t

W_{t-1}

$V(W_{t-1})$

W_{t-2}

$V(W_{t-2})$

Continuous bag-of-words

Word2vec Embeddings

Continuous bag-of-words

W_{t+2} $V(W_{t+1})$

W_{t+1} $V(W_{t+2})$

W_{t-1} $V(W_{t-1})$

W_{t-2} $V(W_{t-2})$

$\sum \cdot$ W_t

Skip-Gram

$V(W_{t+1})$ W_{t+2}

$V(W_{t+2})$ W_{t+1}

$V(W_{t-1})$ W_{t-1}

$V(W_{t-2})$ W_{t-2}

$\sum \cdot$ W_t

Entity Encoding

Output layer

Dense layer

EE layer A

EE layer B

EE layer C

Hot one encoding layer A

Hot one encoding layer B

Hot one encoding layer C
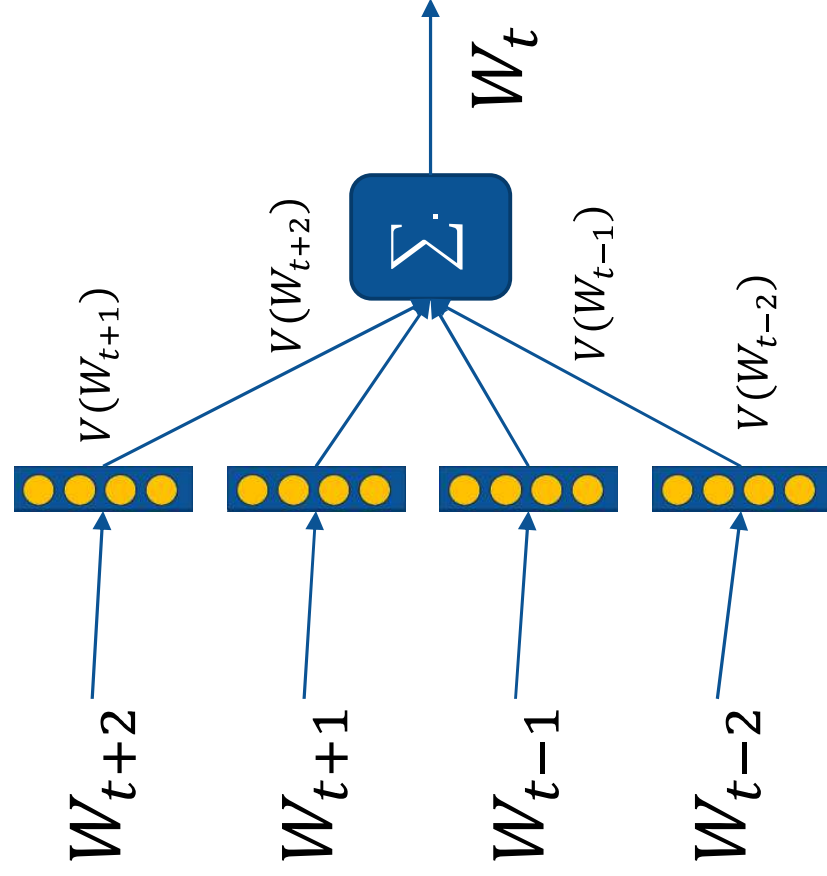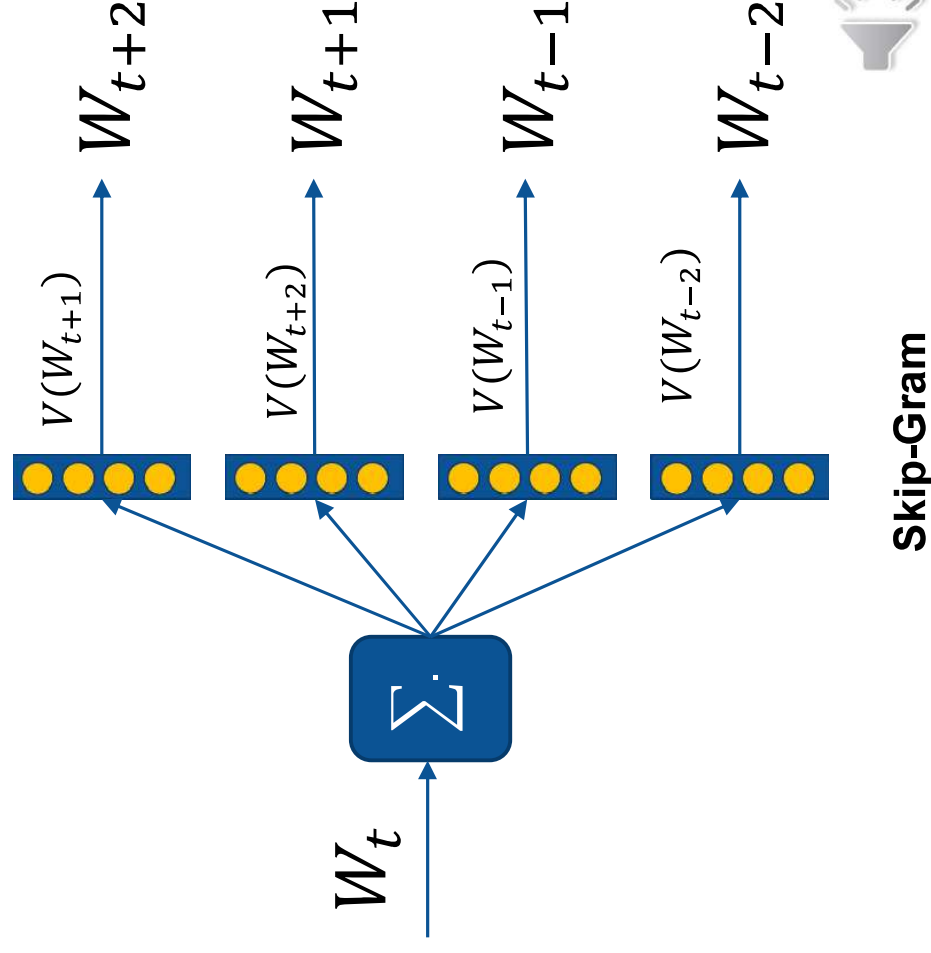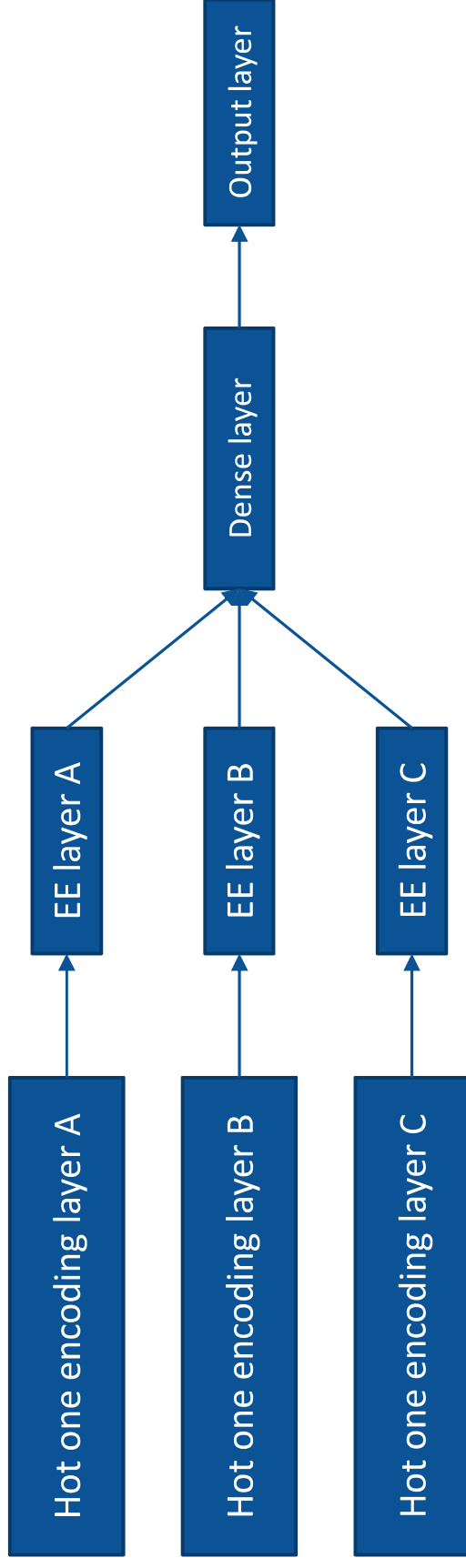
Entity Embeddings of Categorical Variables

Limitations of Word Embeddings

- Infrequent use of terms/words
- Antonymy
- Polysemy
- Biased Embeddings

Summary

- Neural embeddings can handle high dimensional data and improve performance

- Embeddings are powerful representations on capturing dependencies between words

- There is no systematic way on how these approach can be used in Electronic Health Records

References

- Kamath et al. 'Deep learning for NLP and Speech Recognition', Springer, 2019.
- Chen et al. 'Representation Learning for Electronic Health Records: A Survey', J. Phys Conf, 2020.