Word Embeddings

Francis Liu

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
lvb.wiwi.hu-berlin.de

Motivation - Word Embeddings in Finance Research

1. Sentiment Analysis

Extent the Loughran & McDonald's Dictionary (Theil, Štajner, & Stuckenschmidt, 2018)

Polarity propagation of financial terms (Ito, T., Izumi, K., Tsubouchi, K., & Yamashita, T., 2016)

2. Predict Stock Price Movement

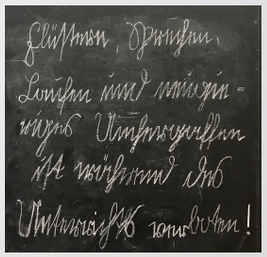
Predict stock returns subsequent to the disclosure of financial materials (Kraus, M., & Feuerriegel, S., 2017)

3. Sentiment based forecasting of stock returns using financial news

Ongoing research by Manuel Tonneau, M.sc. student of Prof. Dr. Stefan Lessmann and Prof. Dr. Wolfgang Karl Härdle

4. Turn things into vector

Japanese candle stick (Poženel Marko, & Lavbič Dejan.,2019)

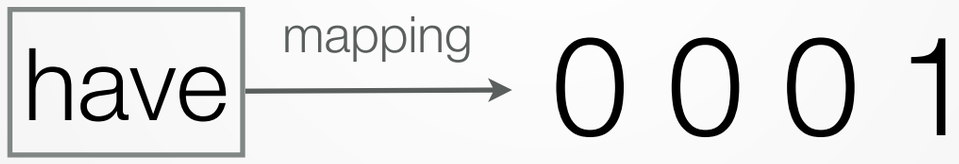


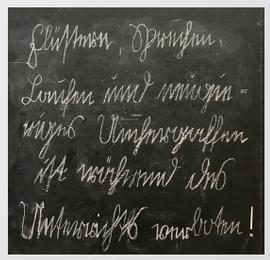
Motivation

One way to let computer knows a word: One-Hot Encoding

Corpus: have a good day

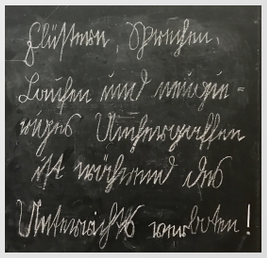
Dictionary: {a, day, good, have}





Motivation - Word Embeddings

- Problems of One-Hot Encoding
    1. Too sparse
    2. High dimension
    3. All words are independent of each other
- A learned representation for text from corpus (no human decision involved)
- Words with similar semantic meaning occupy similar space

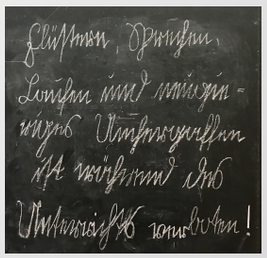


Outline

1. Motivation
2. Word Embeddings
3. Word2Vec
4. GloVe
5. BERT

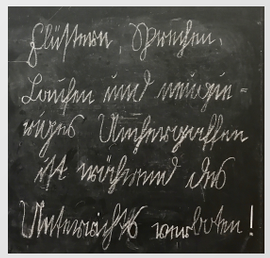
Different Techniques of Word Embeddings

- Word2Vec (W2V)
  - 2013 at Google
  - Neural Network with one hidden layer
- Global Vectors for Word Representation (GloVe)
  - 2014 at Stanford
  - By co-occurrence
- Bidirectional Encoder Representations from Transformers (BERT)
  - 2018 at Google
  - Transformer + Attention Mechanism



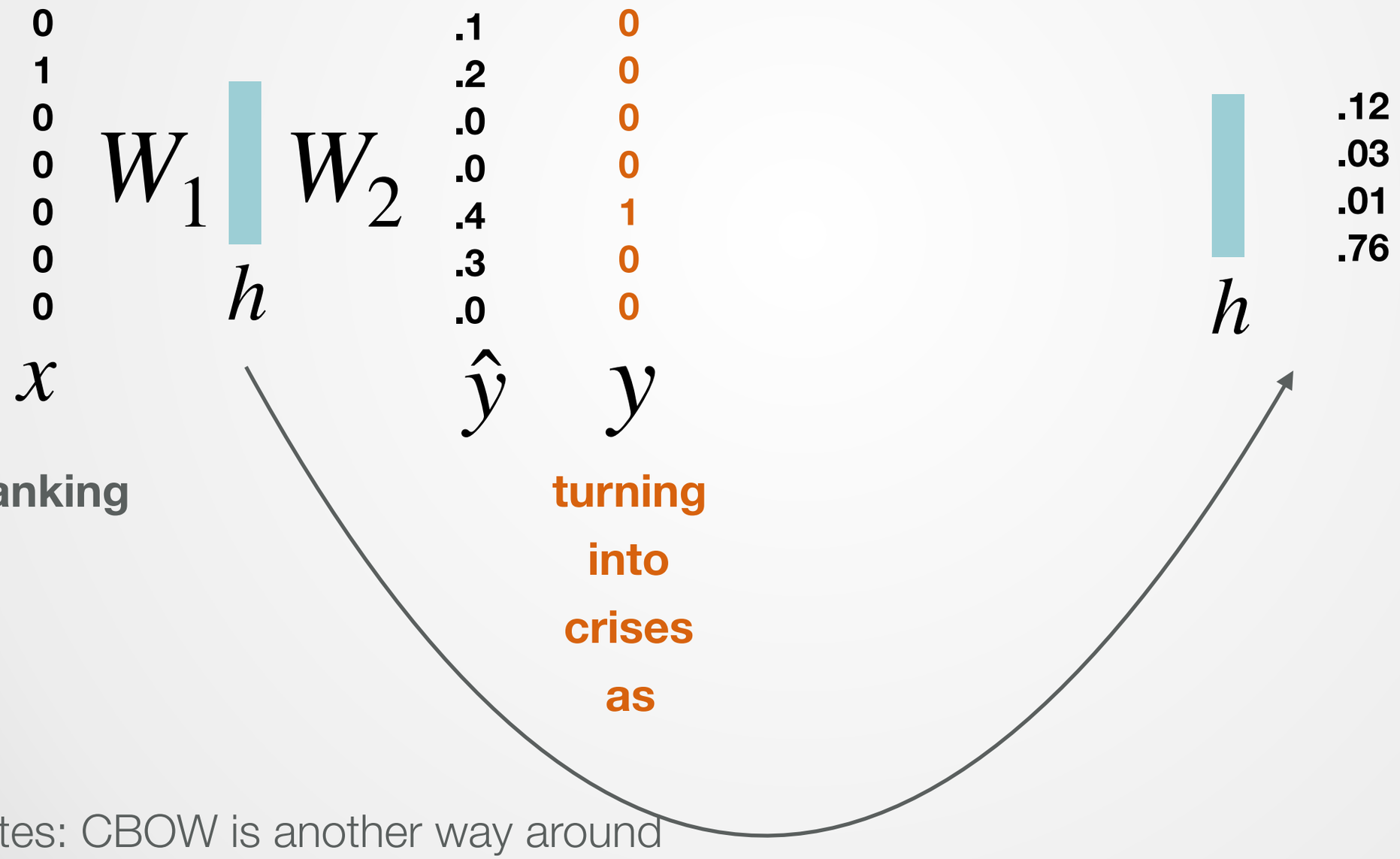
W2V: The Idea of Word2Vec

- Distributional Representation
- Words that are used and occur in the same contexts tend to purport similar meanings
- "You shall know a word by the company it keeps" (J. R. Firth 1957)
- Neural Network with one hidden layer; One-Hot Encoding input and output
- Sliding window to determine context

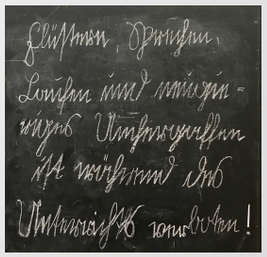


W2V: Neural Network Architecture of Word2Vec

One Hidden Layer Neural Network

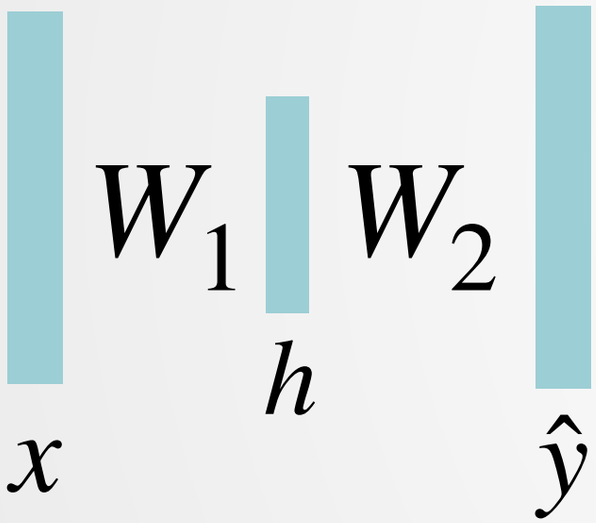


Notes: CBOW is another way around



W2V: Neural Network Architecture of Word2Vec

One Hidden Layer Neural Network



Feed forward:

$$h_{(1,emb)} = x^T_{(1,vocab)} W_{1(vocab,emb)} + b_{1(emb)}$$

$$\hat{y}_{(1,vocab)} = Softmax(h_{(1,emb)} W_{2(emb,vocab)} + b_{2(vocab)})$$

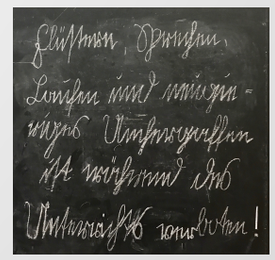
$$Softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)}$$

Cross entropy as loss function:

$$L = -\sum^{batchsize} p(y) log_2 q(\hat{y})$$

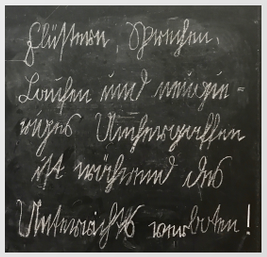
Back propagation by gradient decent

$$w_{new} := w_{old} - \alpha \frac{dL}{dw_{old}}$$

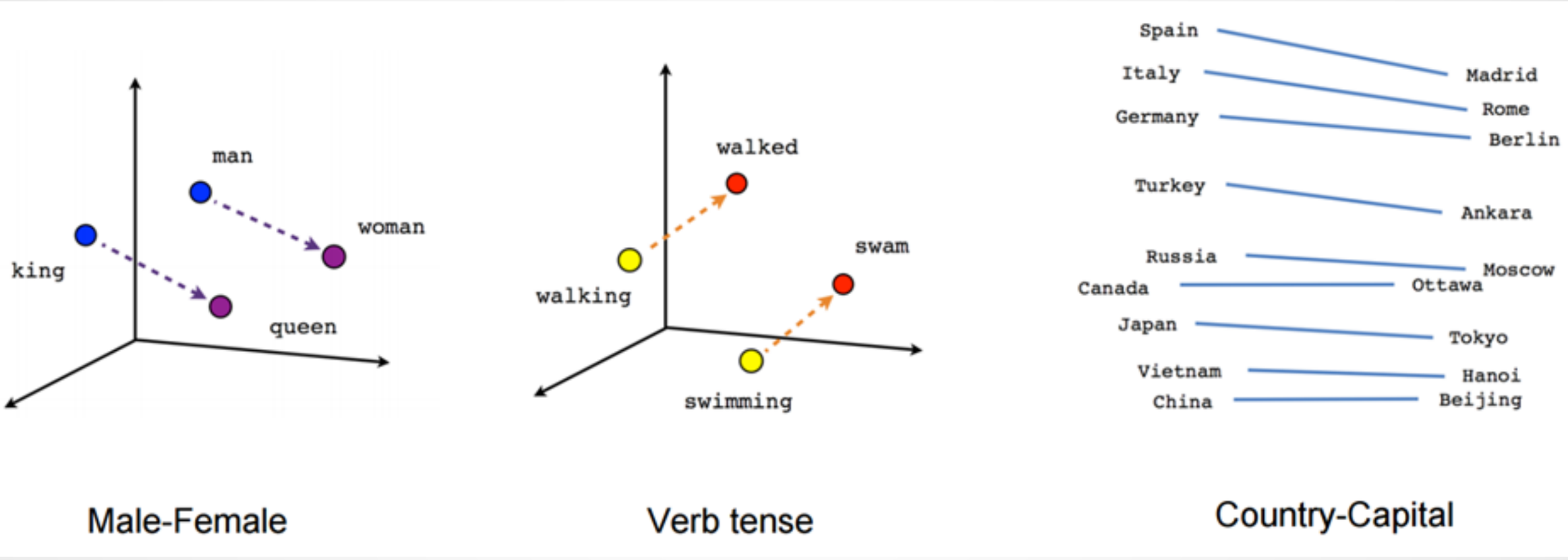


W2V: Extracting Word Vectors

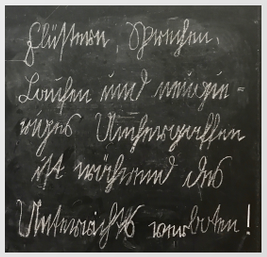
$$x^\top W_1 \qquad [0\ 1\ 0\ 0\ 0] \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 4 & 5 \\ 1 & 8 \\ 2 & 4 \end{bmatrix}$$



W2V: Outcomes



Source: https://www.tensorflow.org/tutorials/word2vec

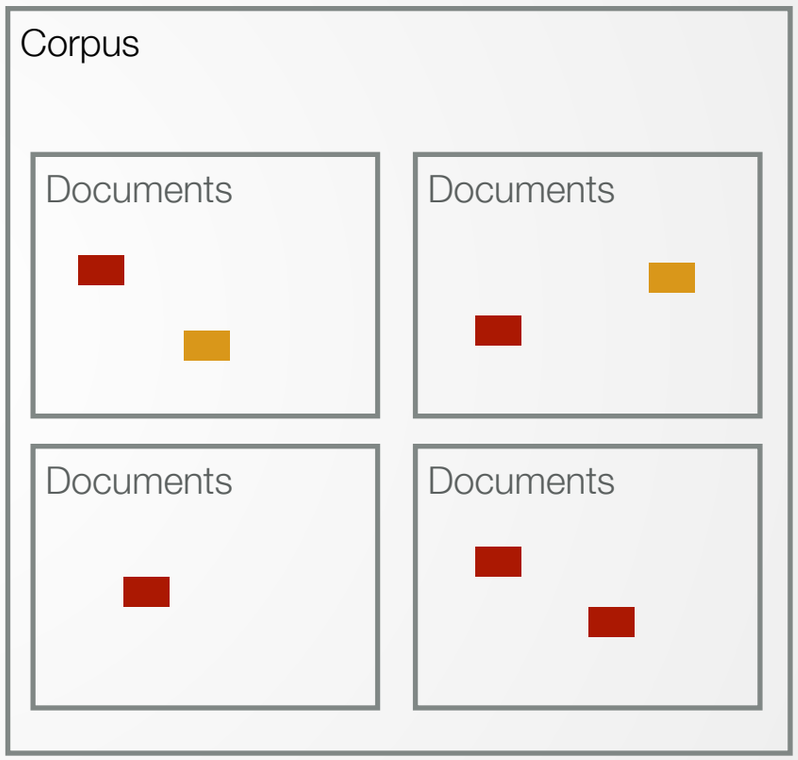


Global Vectors for Word Representation - GloVe

- A global cooccurrence thinking (Pennington, Socher, & Manning, 2014)

Word2Vec sliding window on sentence level

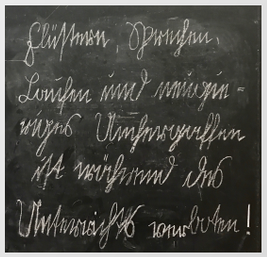
...needs unified **bank** regulation to ...



# of doc contains ▬ and ▬

# of doc contains ▬

Notes: Penalties on distance between words

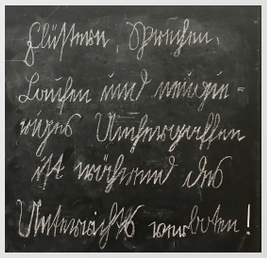


Global Vectors for Word Representation - GloVe

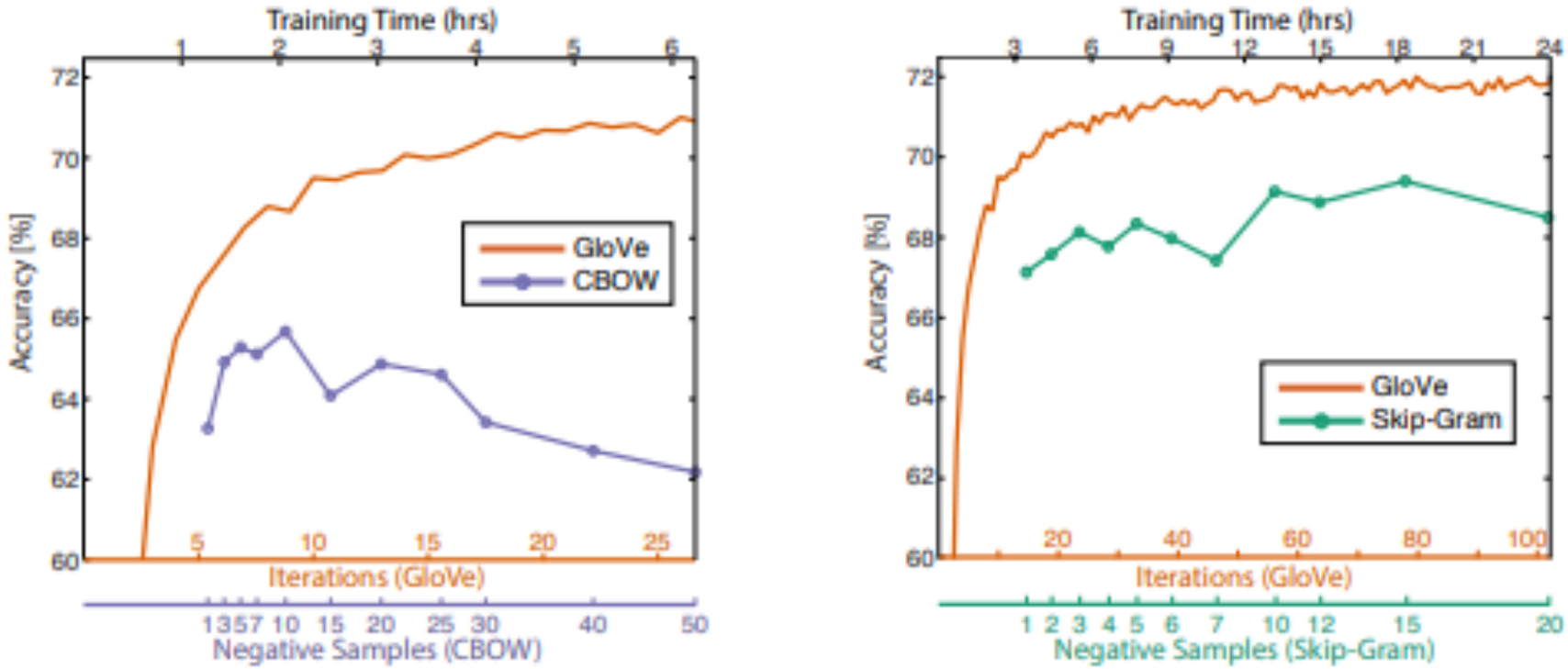
- Example

Probability and Ratio	k=solid	k=gas	k=fashion
P(k|ice)	1.9E-04	6.6E-05	1.7E-05
P(k|steam)	2.2E-05	7.8E-04	1.8E-05
P(k|ice)/P(k|steam)	8.9	8.5E-02	0.96

- Optimize weights by gradient decent so that $F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
- Researchers chose: $exp(w_i^T \tilde{w}_k + b_i + \tilde{b}_k) = X_{ik}$



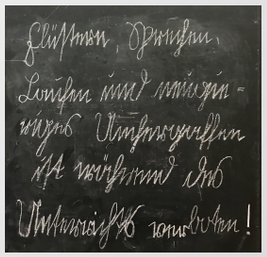
GloVe: The Competing W2V and GloVe



(a) GloVe vs CBOW

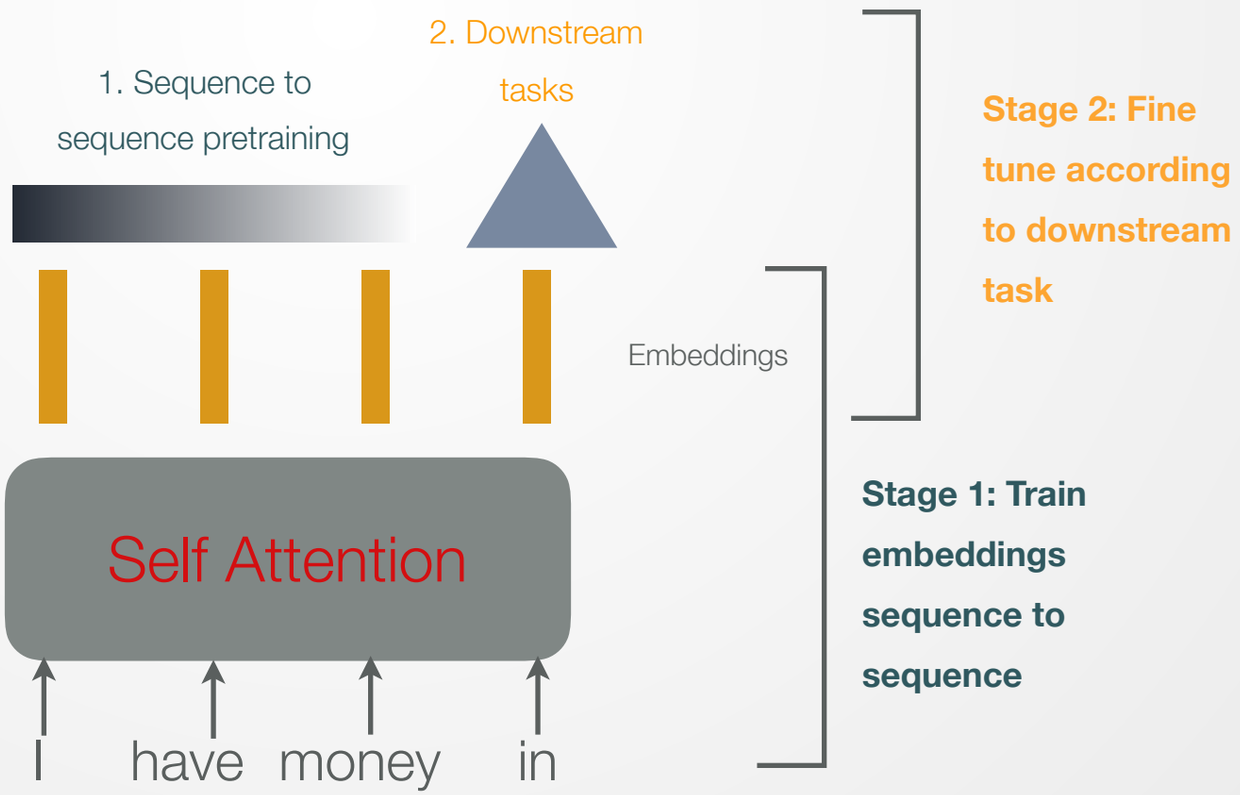
(b) GloVe vs Skip-Gram

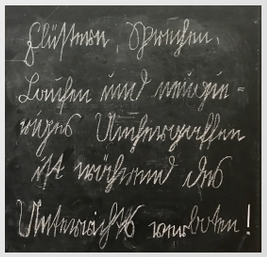
Figure 4: Overall accuracy on the word analogy task as a function of training time, which is governed by the number of iterations for GloVe and by the number of negative samples for CBOW (a) and skip-gram (b). In all cases, we train 300-dimensional vectors on the same 6B token corpus (Wikipedia 2014 + Gigaword 5) with the same 400,000 word vocabulary, and use a symmetric context window of size 10.



Bidirectional Encoder Representations from Transformers

- Main Idea: Give word different embeddings according to different context
- Encoder part of a Transformer, trained by a decoder
- Attention Neural Network (not CNN or RNN)
- Pretrain with large dataset and fine tune for downstream task
- Work extremely well in NLP tasks, e.g. Standford Question Answering Dataset,





BERT: Result Illustration

- Downstream Task: Text Classification
- Dataset: Reddit Comments from Jun 2016 and Dec 2017
- Data source:Google BigQuery Open Dataset
- 3 epochs, 75% accuracy on test set

Codes Available!

June –September 2016

600 – 630 USD

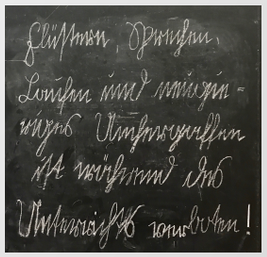
Decrease Price stabilized in the low 600 range.

December 2017

19,783.06 USD

Increase Price rose 5% in 24 hours, with its value being up 1,824% since 1 January 2017, to reach a new all-time high

from wiki



BERT: Bitcoin Example

June –September 2016

600 – 630 USD

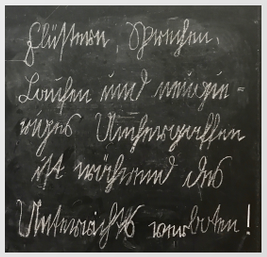
Decrease Price stabilized in the low 600 range.

December 2017

19,783.06 USD

Increase Price rose 5% in 24 hours, with its value being up 1,824% since 1 January 2017, to reach a new all-time high

from wiki

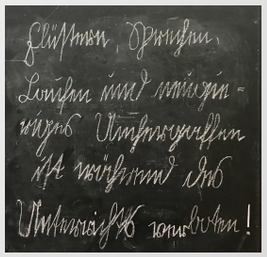


BERT: Reddit Comment (1)

When do you think the below Reddit post is posted?

Jun 2016 or Dec 2017?

"Actually price of electricity keeps bitcoin out of the hands of globalists. Cheap electricity is found in outlying, underdeveloped areas that lack significant industrial, commercial, residential, or agricultural demand for electricity proportionate to supply.\n\nhttps://en.m.wikipedia.org/wiki/Electricity_pricing#"

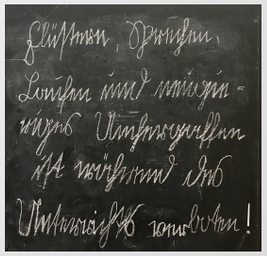


BERT: Machine Understanding

y=jun2016 (probability **0.933**, score **2.627**) top features

Contribution[?]	Feature
+2.015	Highlighted in text (sum)
+0.612	<BIAS>

'actually price of electricity keeps bitcoin out of the hands of globalists. cheap electricity is found in outlying, underdeveloped areas that lack significant industrial, commercial, residential, or agricultural demand for electricity proportionate to supply. https://en.m.wikipedia.org/wiki/electricity_pricing#'

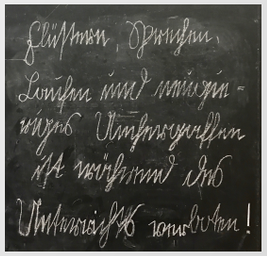


BERT: Reddit Comment (2)

When do you think the below Reddit post is posted?

Jul 2016 or Dec 2017?

"I have money in binance that I pretty much just leave as btc and if I see something taking off I buy in and then sell before bed. Has been working out pretty well so far. It's all most likely going to end up in Ripple and trx at some point tho. "



BERT: Machine Understanding

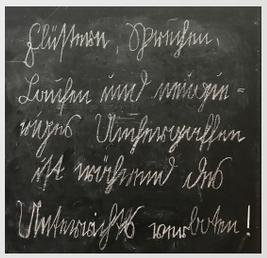
y=dec2017 (probability **0.997**, score **-5.673**) top features

Contribution[?]	Feature
+5.700	Highlighted in text (sum)
-0.027	<BIAS>

i have money in binance that i pretty much just leave as btc and if i see something taking off i buy in and then sell before bed. has been working out pretty well so far. it's all most likely going to end up in ripple and trx at some point tho.

"Binance" is a company founded 2017

"Ripple" was heavily discussed in Dec 2017 (Google Trend)



References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Proceedings of the 2019 Conference of the North. Presented at the Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1423

Ito, T., Izumi, K., Tsubouchi, K., & Yamashita, T. (2016). Polarity propagation of financial terms for market trend analyses using news articles. 2016 IEEE Congress on Evolutionary Computation (CEC). https://doi.org/10.1109/cec.2016.7744230

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems, 104, 38–48. https://doi.org/10.1016/j.dss.2017.10.001

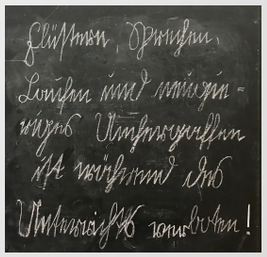
Luss, R., & D'Aspremont, A. (2012). Predicting abnormal returns from news using text classification. Quantitative Finance, 15(6), 999–1012. https://doi.org/10.1080/14697688.2012.672762

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. NIPS.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/d14-1162

Poženel Marko, & Lavbič Dejan. (2019). Discovering Language of the Stocks. Frontiers in Artificial Intelligence and Applications, 315(Databases and Information Systems X), 243–258. https://doi.org/10.3233/978-1-61499-941-6-243

Theil, C. K., Štajner, S., & Stuckenschmidt, H. (2018). Word Embeddings-Based Uncertainty Detection in Financial Disclosures. Proceedings of the First Workshop on Economics and Natural Language Processing, 32–37. https://doi.org/10.18653/v1/W18-3104



Appendix: Innovation Timeline

1986: Learning Representations by Back-Propagating Errors

2003: Neural Language Models

2013: Word2Vec

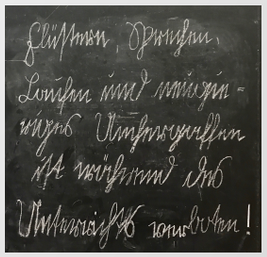
2014: GloVe

2015: Skip-Thought Vectors

2017: Fasttext, Transformers

2018: ELMo, BERT

2019: T5



Appendix - Loughran & McDonald's Dictionary

Identify sentiment lexicon list

Count the positive and negative words to determine sentiment

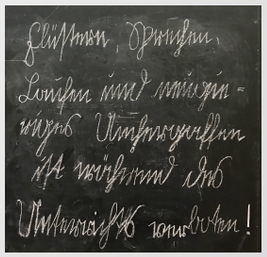
Easy to understand

Cannot handle negation

Regress 10-Ks and 10-K405 documents from www.sec.gov (1994 - 2008) on Stock returns relative to the 10-K filing date

Fin Neg: LOSS, IMPAIRMENT, ADVERSE, AGAINST

For more detail: https://www.uts.edu.au/sites/default/files/ADG_Cons2015_Loughran%20McDonald%20JE%202011.pdf



Appendix - Self Attention

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{k=1}^{T} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T} exp(e_{ij})}$$

$$e_{ij} = \alpha(s_{s-1}, h_j)$$

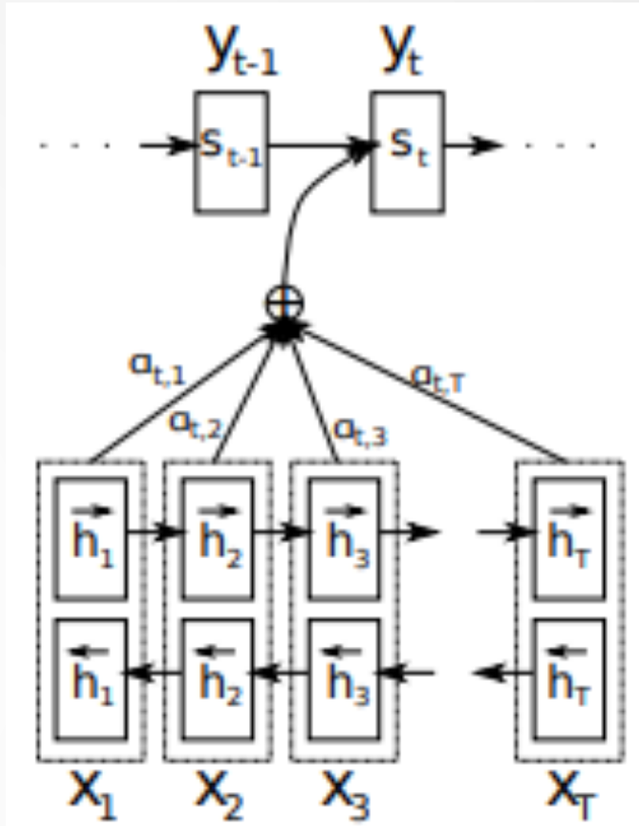
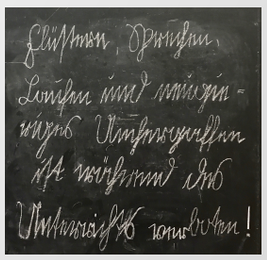


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Paper: Attention Is All You Need (link)



Appendix - Word2Vec

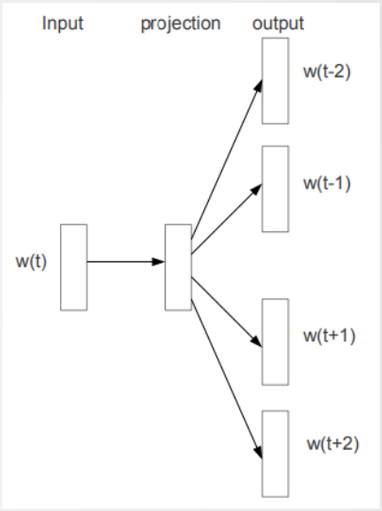
Two ways to construct the dataset for the neural network

Skipgram and Continuous Bag of Words (CBOW)

example:

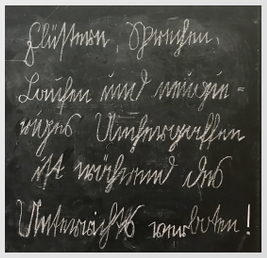
Sentence 1: have a good day

Sentence 2: have a great day



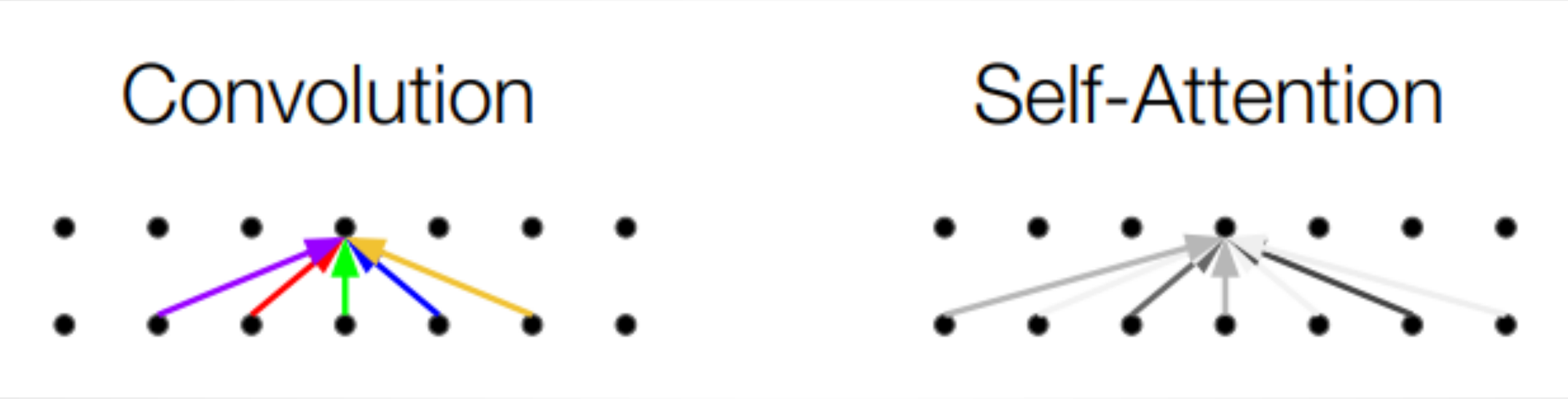
Skipgram

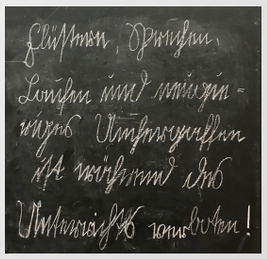
Skipgram	CBOW
(have, a) (a, have) (a, good) (good, a) (good, day) (day, good)	(a, have) (have, a) (good, a) (a, good) (day, good) (good, day)
with window size = 1	



Appendix - Self Attention

Difference between convolution and self attention





Appendix - The Transformer

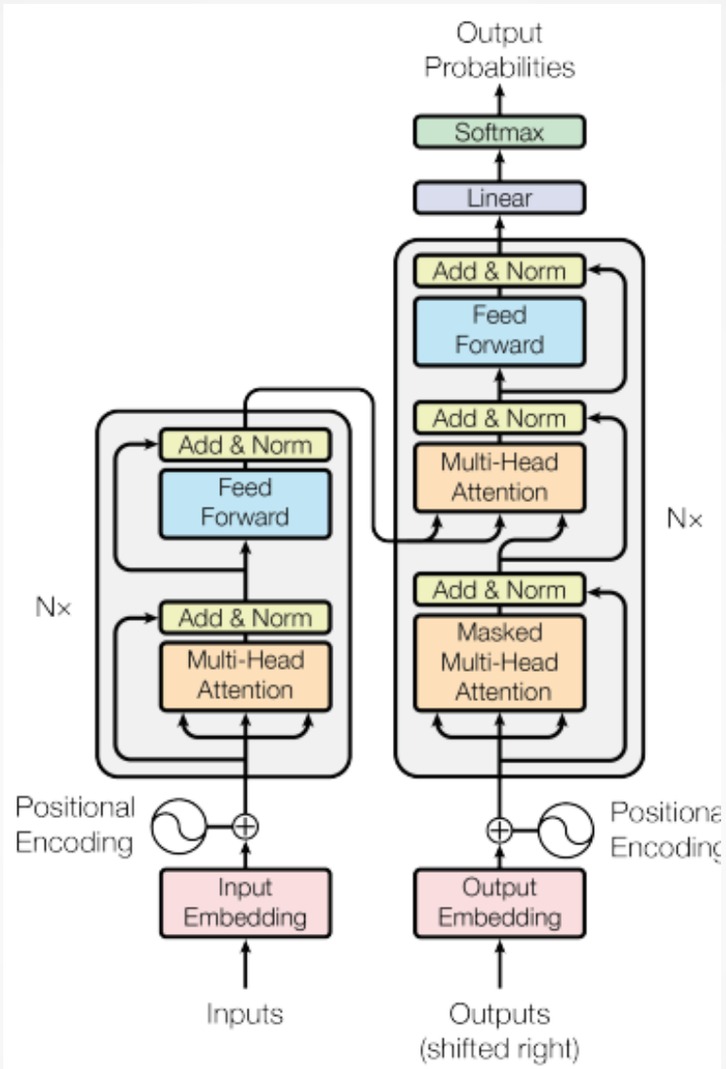
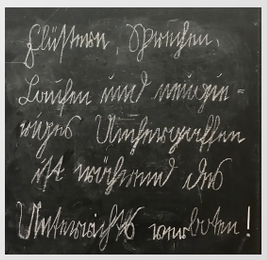


Figure 1: The Transformer - model architecture.

Attention Is All You Need

http://jalammar.github.io/illustrated-transformer/

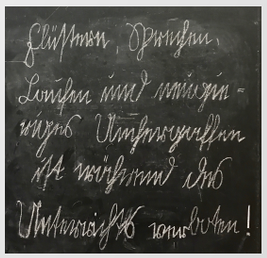


Appendix - Word Embeddings Origin

"The procedure repeatedly adjusts the weights of the connections in the network so as to **minimize a measure of the difference** between the **actual output vector** of the net and the **desired output vector**.

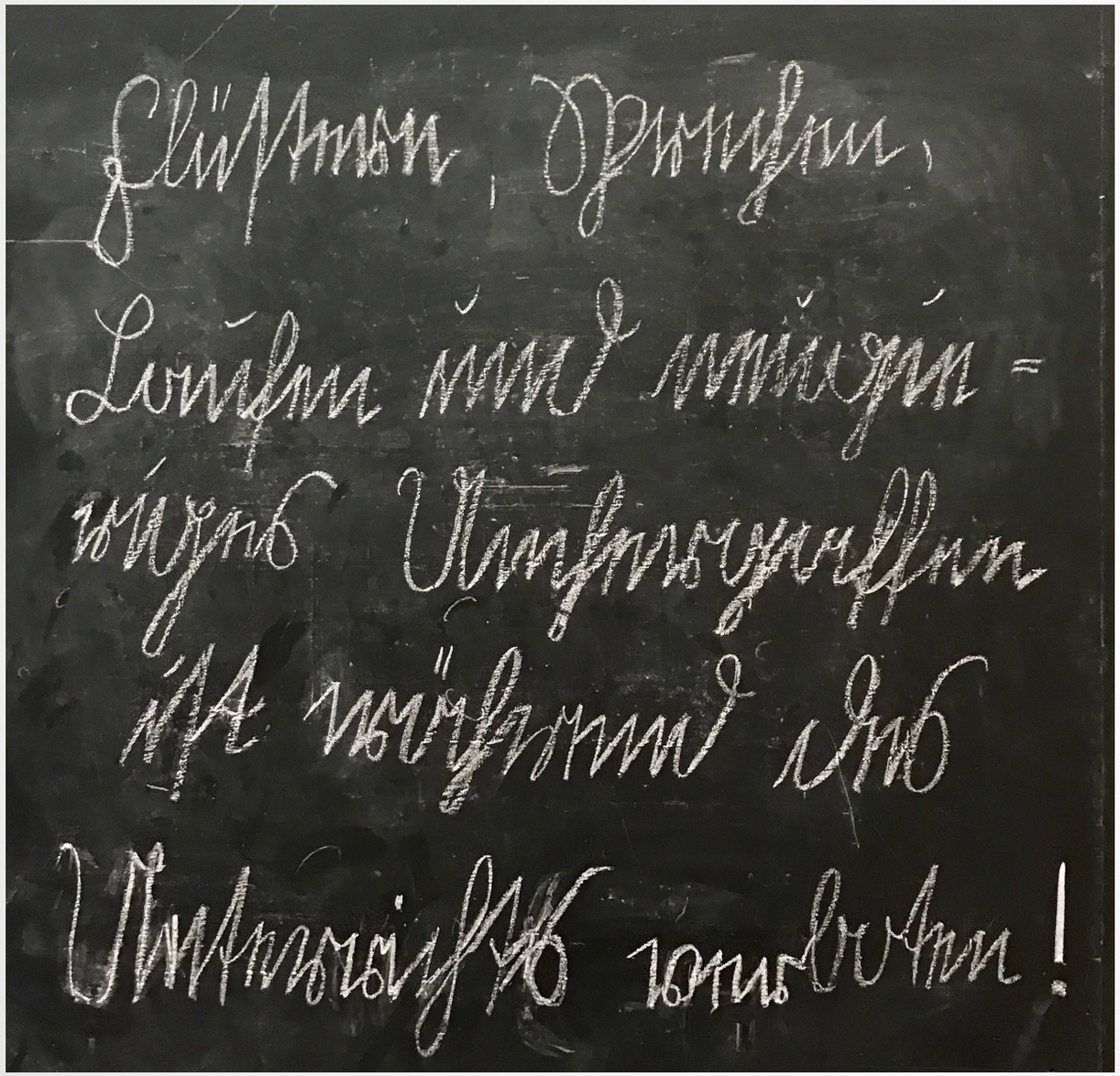
As a result of the weight adjustments, **internal 'hidden' units** which are not part of the input or output come to **represent important features of the task domain**, and the regularities in the task are captured by the interactions of these units"

(David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams, 1986)



Appendix - Motivation

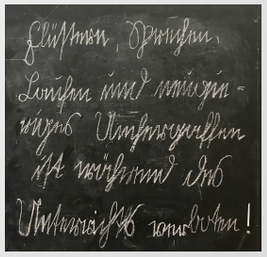
⊡ Machine Translation



Heinrich von Kleist's school (Frankfurt/O)

Flüstern, Sprechen,
Laufen und neugieriges
Umhergaffen ist
während des
Unterrichts verboten !

Whisper, talking,
running and snoopy
gawking around is
forbidden during
class !



Context is Important

⊡ MirrorGAN: Learning Text-to-image Generation by Redescription

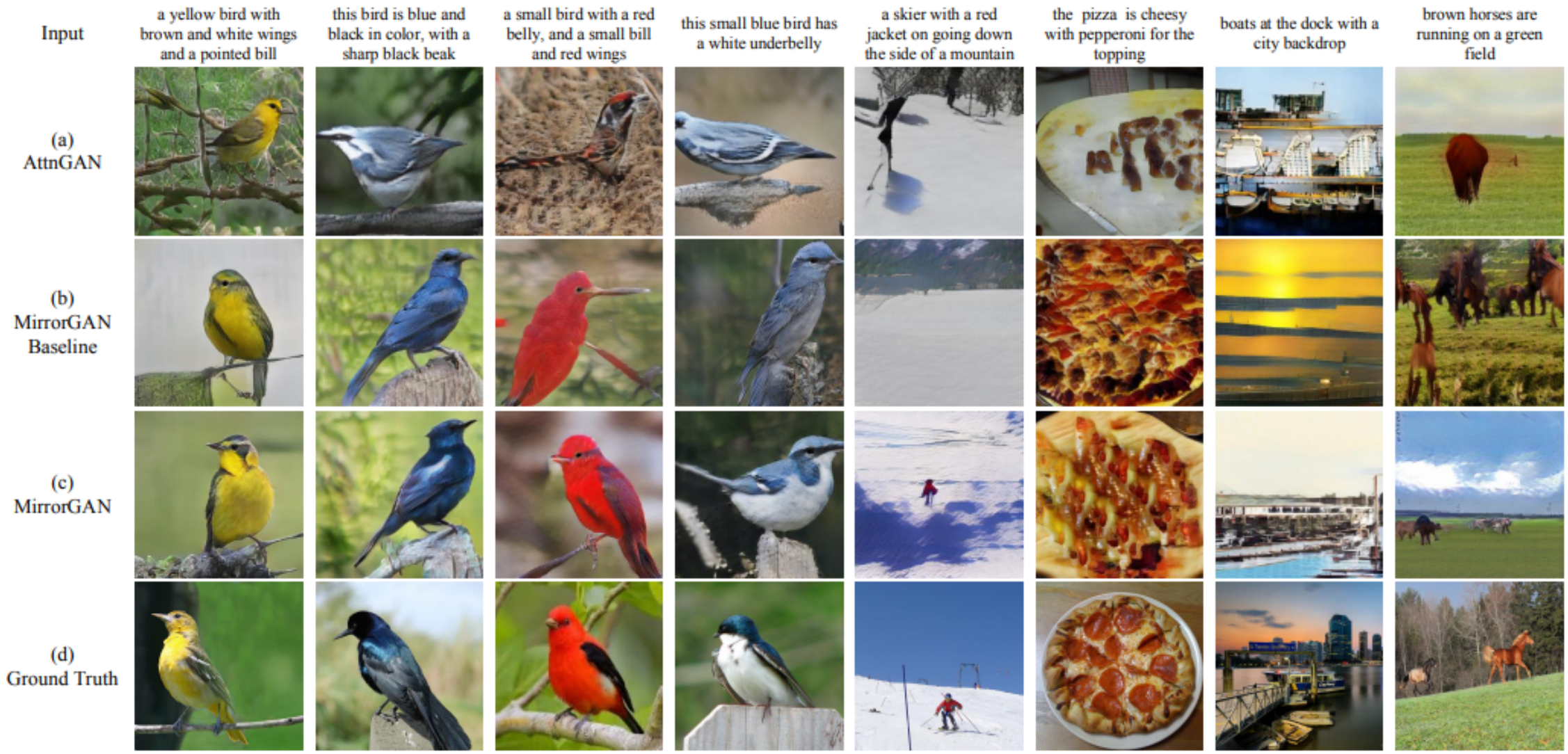


Figure 3: Examples of images generated by (a) AttnGAN [35], (b) MirrorGAN Baseline, and (c) MirrorGAN conditioned on text descriptions from CUB and COCO test sets and (d) the corresponding ground truth.