How quantum theory contributes to NLP?

Benyou Wang

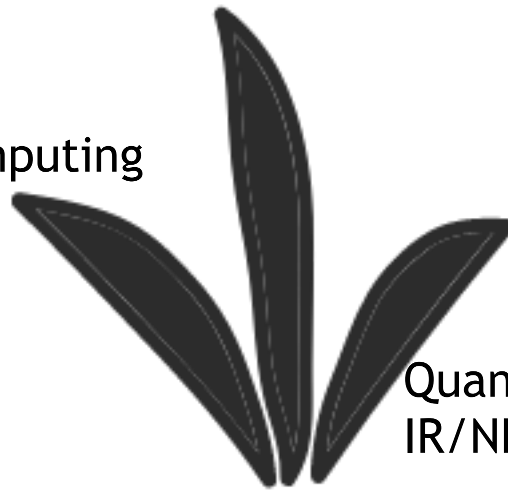
Supervised by Massimo Melucci and Emanuele Di Buccio

University of Padua

QCAI 2020, Virtually, previously Tianjin, Dec 22 2020

Quantum Physics

Quantum Computing



Quantum **inspired/motivated/enhanced**
IR/NLP/ML

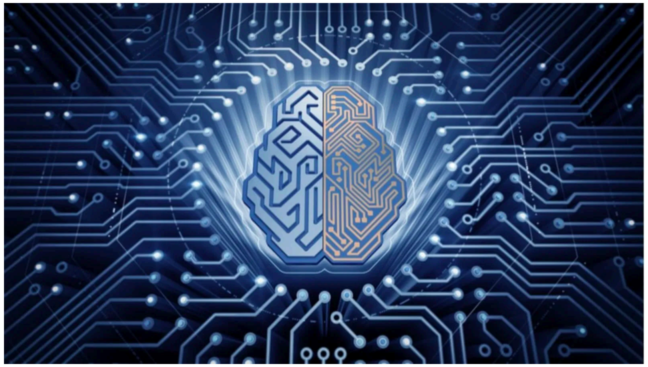
Mathematical Fundamental

Sharing the similar way to **probabilistically** describe the world

Quantum theory **outside** Physics

Using quantum ways to process information

- Quantum computing
  - [Michael A. Nielsen, Isaac L. Chuang. 2011. Quantum Computation and Quantum Information, 10th edition. Cambridge University Press]
  - Arute .et.al. Quantum supremacy using a programmable superconducting processor. Nature. 23 October 2019.
- Social science and cognition science
  - [Jerome R. Busemeyer and Peter D. Bruza. 2013. Quantum Models of Cognition and Decision. Cambridge University Press]
  - [E. Haven and A. Khrennikov. 2013. Quantum Social Science. Cambridge University Press.]
- **Information retrieval**
  - [Van Rijsbergen. 2004. The geometry of information retrieval. Cambridge University Press.]
  - [Massimo Melucci. 2016. Introduction to information retrieval and quantum mechanics. Springer Berlin Heidelberg.]



- *Quantum IR can formulate the different IR models (**logic, vector, probabilistic**, etc.) in a unified framework.*

- Quantum IR does not rely on quantum computing/cognition, but share the same mathematical foundation to **probabilistically** describe the world

- 遇事不决，量子力学?

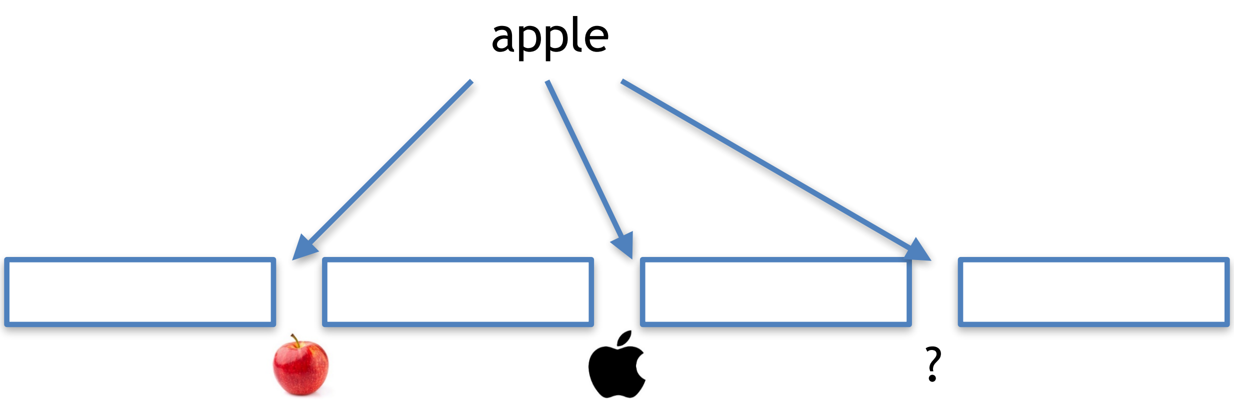
量子波动速读：小学生蒙眼一秒钟阅读10万字?

Quantum Theory and Natural Language

- Motivations
    - Analogy between QT and NLP
    - (Quantum) probability theory in vector space
    - Paradigm with big models and big data
- Applications in NLP
    - Efficiency
    - Effectiveness
    - Interpretability
    - Semantic cognition

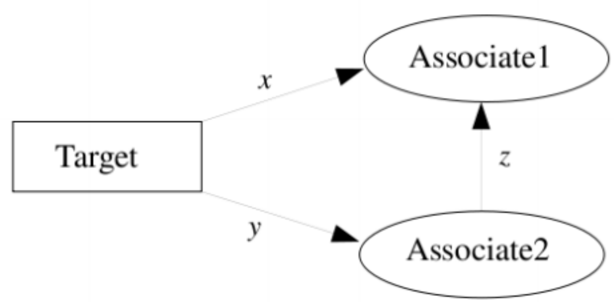
Analogy - superposition

- Ambiguity for words: apple



Analogy - entanglement

- Word association



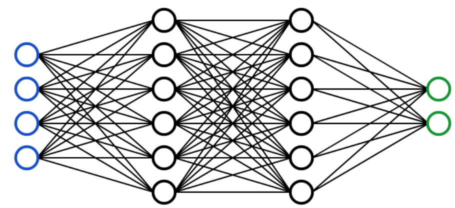
Two associate words are either *both* recalled or *both* not recalled

Quantum Theory and Natural Language

- Motivations
  - Analogy between QT and NLP
  - (Quantum) probability theory in vector space
  - Paradigm with big models and big data
- Applications in NLP
  - In pre-trained language model
  - Complex-valued representations
  - Tensor network language model

From bag-of-words assumption
To word vector based neural networks

Q C A I



discrete

Vector spaces

Set-based Probability Theory

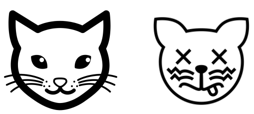
Probability Theory in vector space?

Neural networks usually transform a discrete token index to a vector.
We need a probability theory to describe uncertainty in **vector spaces**.

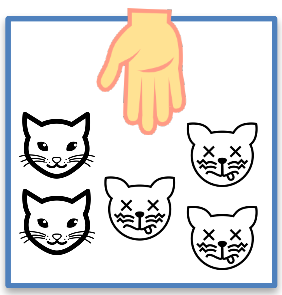
Quantum Probability Theory

a probability theory defining on **vector spaces**

Set-based Probability Theory



alive dead



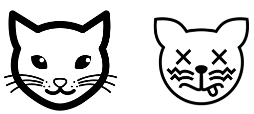
Q: Should the randomly-chosen cat dead or alive ?

A: 0.4 to be alive and 0.6 to be dead

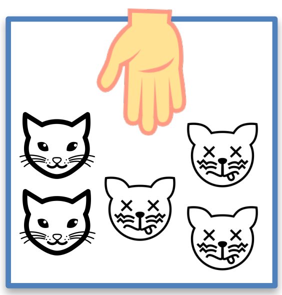
Quantum Probability Theory

a probability theory defining on **vector spaces**

Set-based Probability Theory



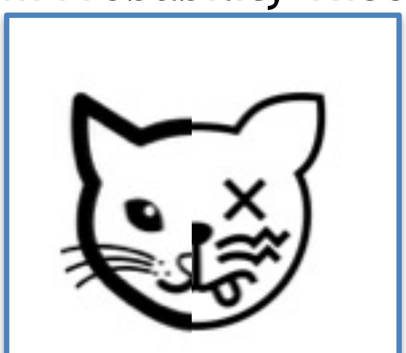
alive dead



Q: Should the randomly-chosen cat dead or alive ?

A: 0.4 to be alive and 0.6 to be dead

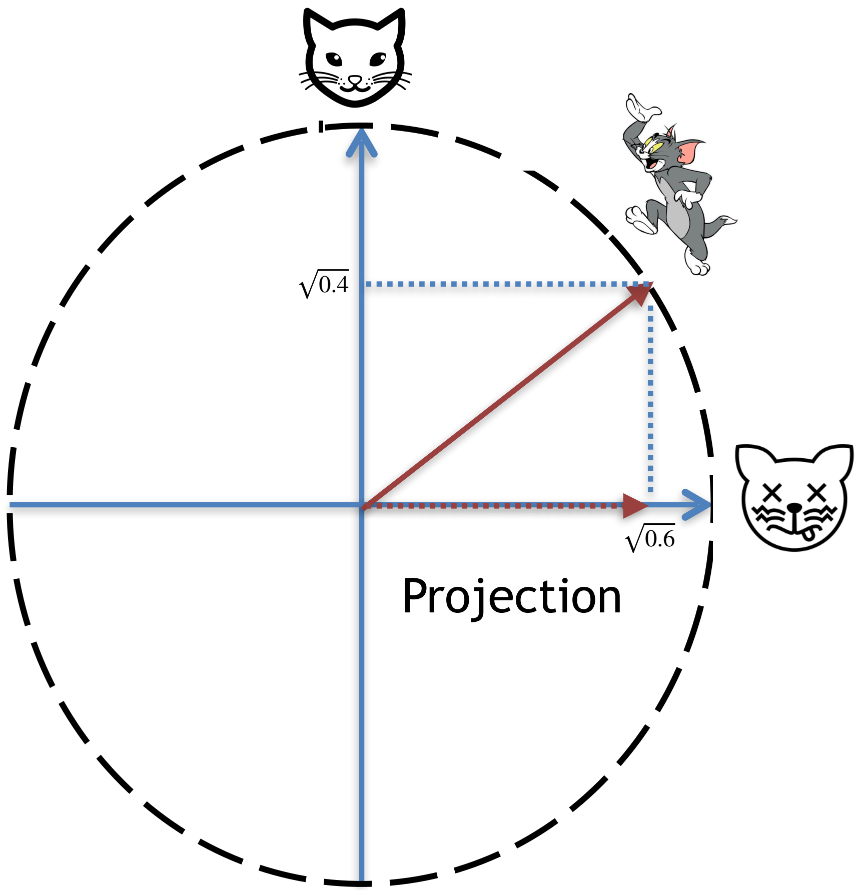
Quantum Probability Theory - **vector-based**



Q: Are these cat dead or alive?

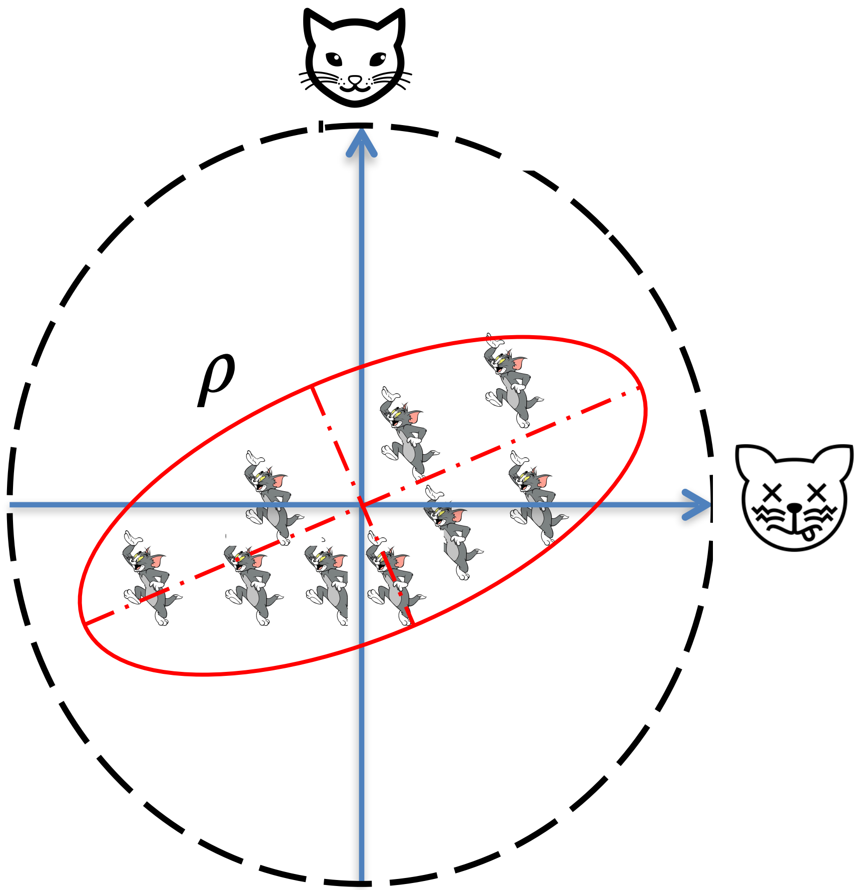
A: 0.501 to be alive and 0.499 to be dead

Probability theory in vector spaces
for single object



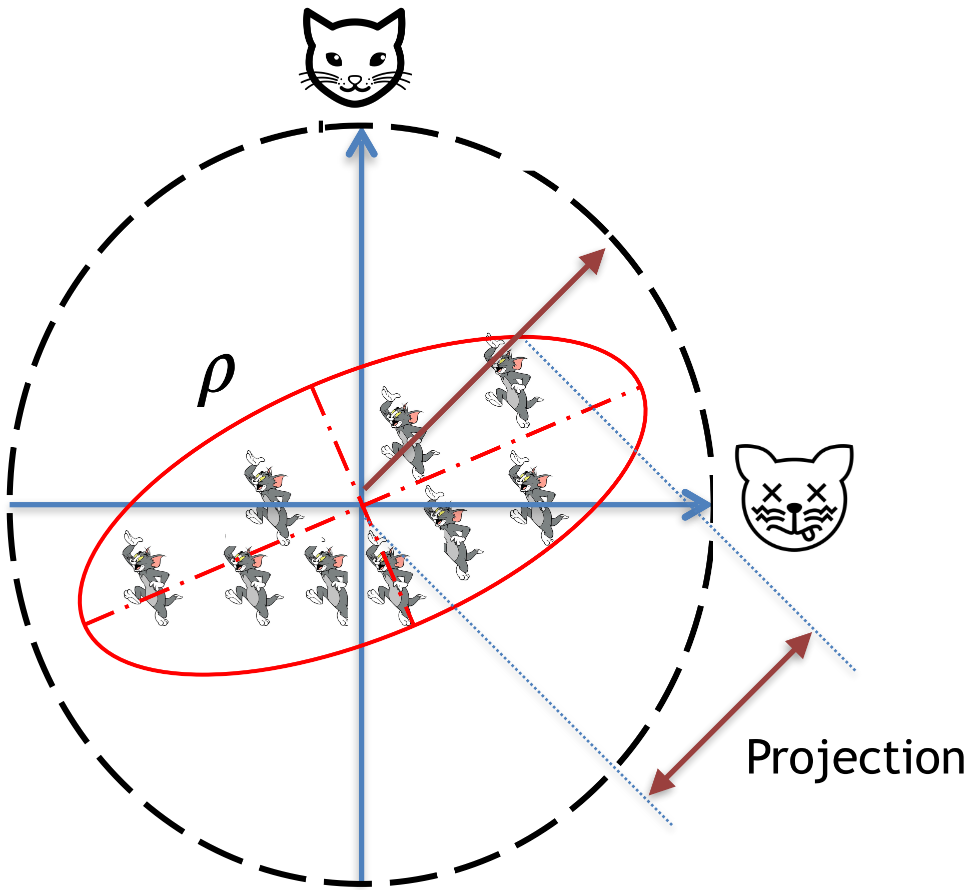
Square of the projection length denotes the probability

Probability theory in vector spaces
for many objects



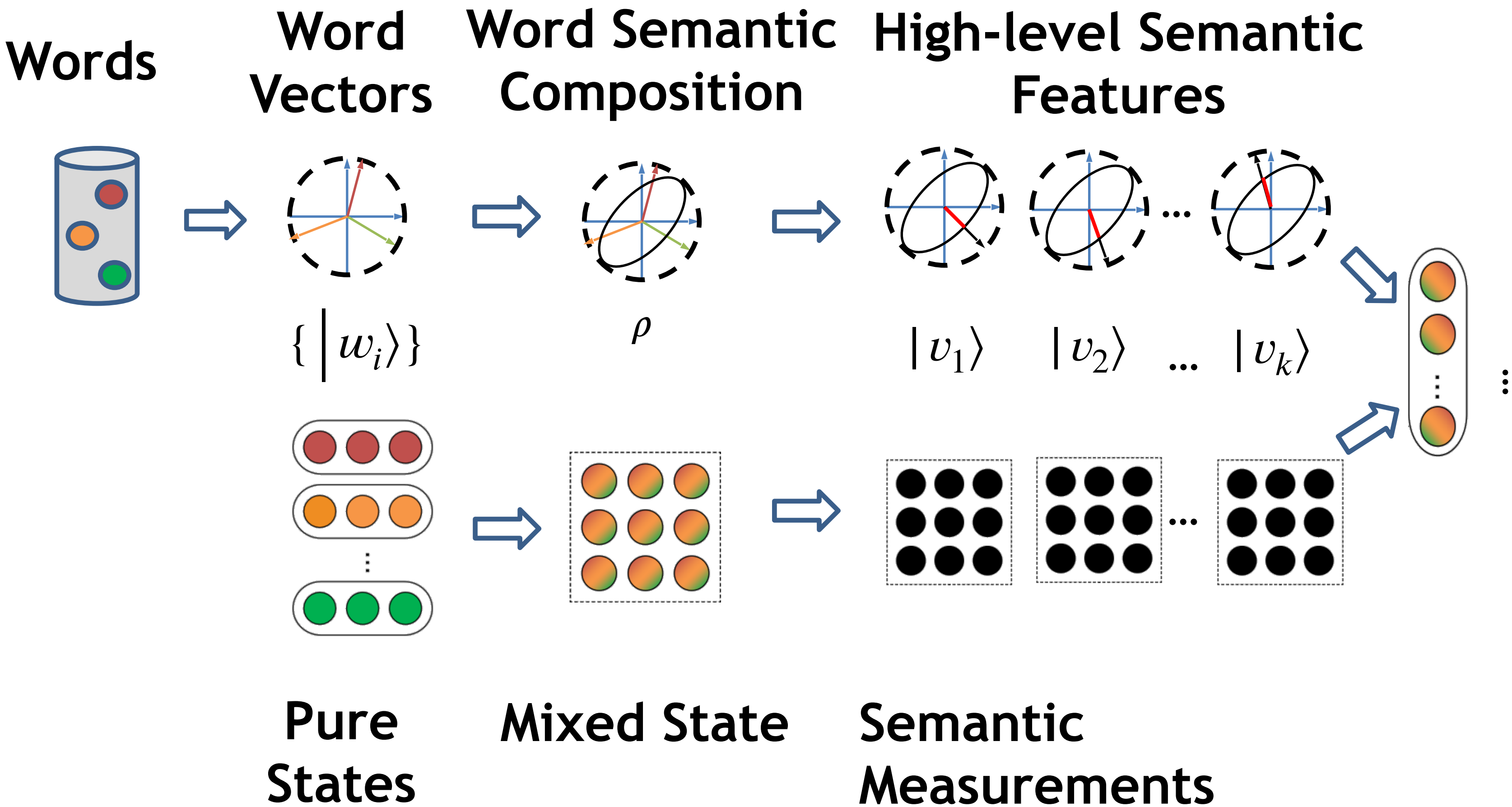
Square of the projection length denotes the probability

Probability theory in vector spaces
for many objects



Square of the projection length denotes the probability

Semantic Hilbert Space



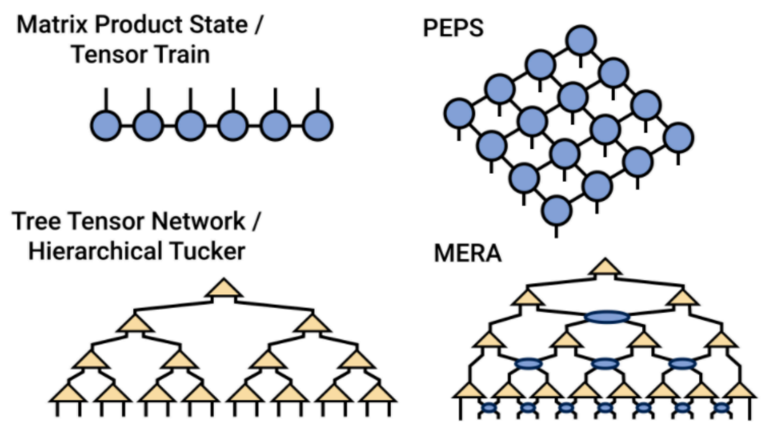
Benyou Wang*, Qiuchi Li*, Massimo Melucci, and Dawei Song. *Semantic Hilbert Space for Text Representation Learning*. In *WWW2019*
Li, Qiuchi*, **Benyou Wang***, and Massimo Melucci. "CNM: An Interpretable Complex-valued Network for Matching." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4139-4148. 2019. NAACL 2019 best explainable paper

Quantum Theory and Natural Language

- Motivations
  - Analogy between QT and NLP
  - (Quantum) probability theory in vector space
  - Paradigm with big models and big data
- Applications in NLP
  - In pre-trained language model
  - Complex-valued representations
  - Tensor network language model

BIG Tensors in Physics

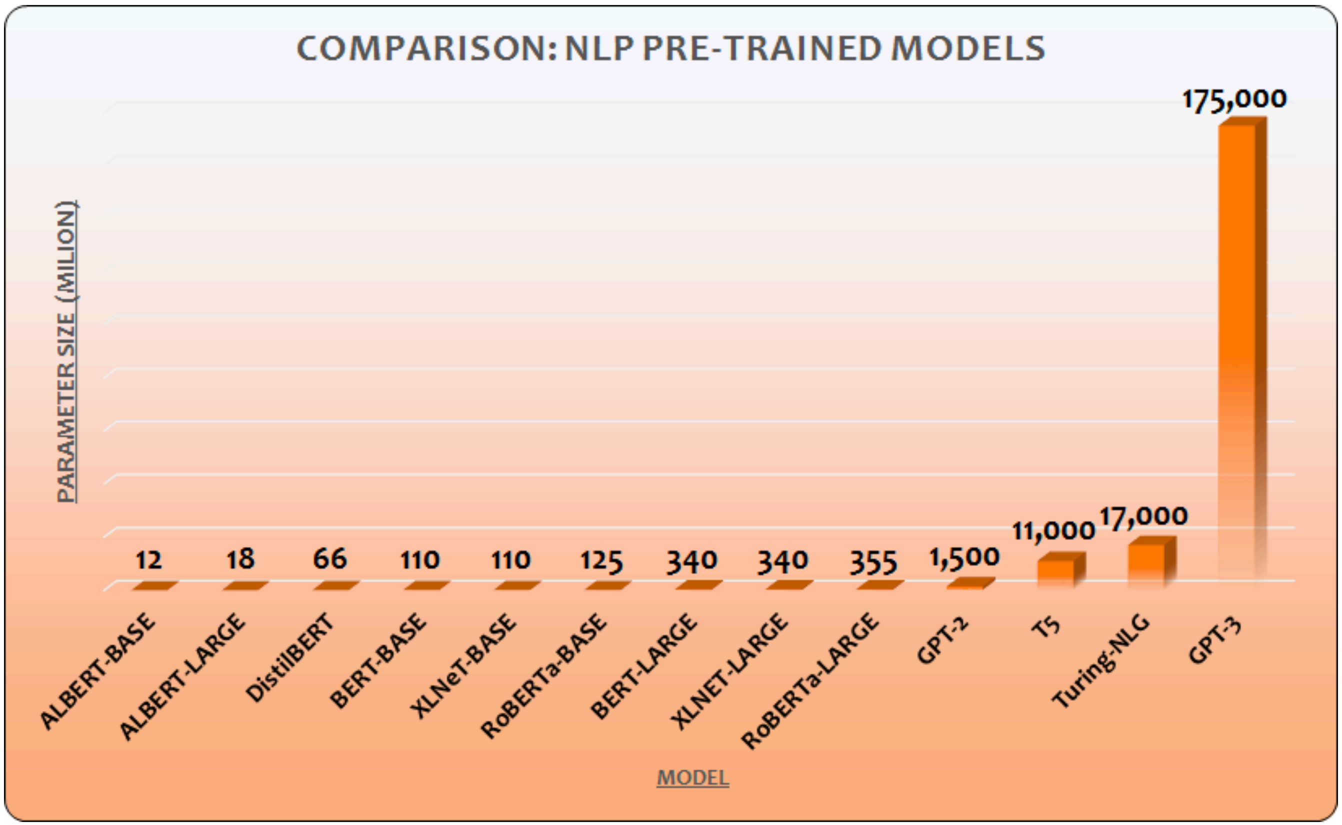
Quantum many body wave functions are represented by usually (exponentially) large tensors



For example, tensor networks are factorizations of very large tensors (quantum many body wave function) into networks of smaller tensors.

BIG Tensors in NLP

Pre-trained language model and massive plain text



https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122

BIG Data

Self-supervised learning with contrastive loss: **without human annotators**

- Baseline: **LM**(GPT,ELMo), **MLM**(BERT), **NSP**(BERT)
- Whole Word Masking (BERT)、SpanBERT (Joshi et al. 2019)
- RTD (**Replaced Token Prediction**): Electra(Clark 2020)
- SOP (**Sentence Order Prediction**): ALBERT(Lan et al. 2020)
- DAE (Denoising Autoencoder (DAE): BART(Mike et al. 2019)
- Multi-task Learning: MT-DNN(Liu et al. 2019)
- **Generator and Discriminator**: Electra(Clark 2020)

Data examples

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Source: From Qun Liu's talk in CCL 2020

https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122

BIG models

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Quantum Theory and Natural Language

- Motivations
  - Analogy between QT and NLP
  - (Quantum) probability theory in vector space
  - Paradigm with big models and big data
- Applications in NLP
  - In pre-trained language model
  - Complex-valued representations
  - Tensor network language model

BIG models with **low rank** bottleneck

low rank bottleneck

- Embedding: low rank
- Architecture bottleneck
- Softmax bottleneck for interference

Softmax bottleneck

Softmax activation for prediction (predict next word)

$$P_\theta(x|c) = \frac{\exp \mathbf{h}_c^\top \mathbf{w}_x}{\sum_{x'} \exp \mathbf{h}_c^\top \mathbf{w}_{x'}}$$

The rank of A with N words is limited to be equal or less than M

$$\mathbf{H}_\theta = \begin{bmatrix} \mathbf{h}_{c_1}^\top \\ \mathbf{h}_{c_2}^\top \\ \cdots \\ \mathbf{h}_{c_N}^\top \end{bmatrix};\ \mathbf{W}_\theta = \begin{bmatrix} \mathbf{w}_{x_1}^\top \\ \mathbf{w}_{x_2}^\top \\ \cdots \\ \mathbf{w}_{x_M}^\top \end{bmatrix};\ \mathbf{A} = \begin{bmatrix} \log P^*(x_1|c_1), & \log P^*(x_2|c_1) & \cdots & \log P^*(x_M|c_1) \\ \log P^*(x_1|c_2), & \log P^*(x_2|c_2) & \cdots & \log P^*(x_M|c_2) \\ \vdots & \vdots & \ddots & \vdots \\ \log P^*(x_1|c_N), & \log P^*(x_2|c_N) & \cdots & \log P^*(x_M|c_N) \end{bmatrix}$$

where $\mathbf{H}_\theta \in \mathbb{R}^{N\times d}$, $\mathbf{W}_\theta \in \mathbb{R}^{M\times d}$, $\mathbf{A} \in \mathbb{R}^{N\times M}$, and the rows of $\mathbf{H}_\theta$, $\mathbf{W}_\theta$, and $\mathbf{A}$ correspond to context vectors, word embeddings, and log probabilities of the true data distribution respectively.

Low rank bottleneck in multihead

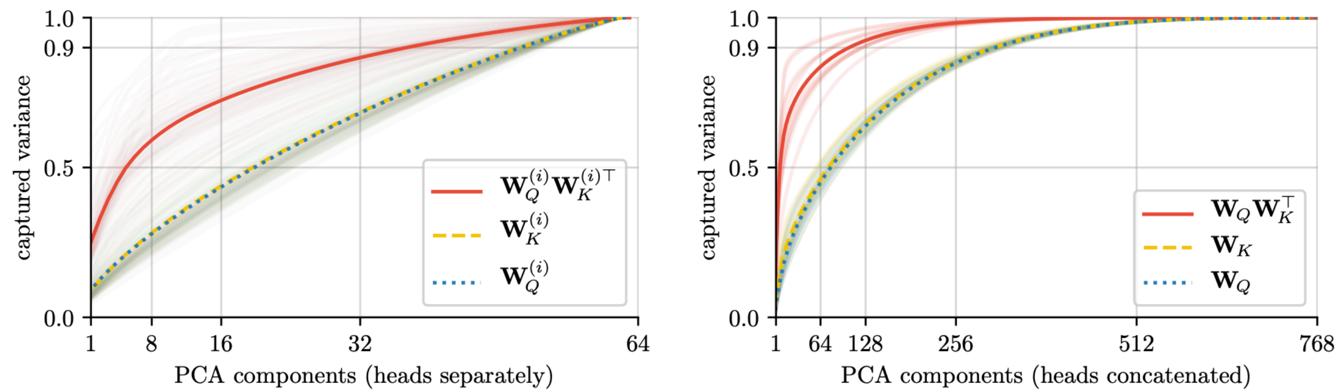


Figure 1: Cumulative captured variance of the key query matrices per head separately (*left*) and per layer with concatenated heads (*right*). Matrices are taken from a pre-trained BERT-base model with $N_h = 12$ heads of dimension $d_k = 64$. Bold lines show the means. Even though, by themselves, heads are not low rank (*left*), the product of their concatenation $\boldsymbol{W}_Q\boldsymbol{W}_K^\top$ is low rank (*right, in red*). Hence, the heads are sharing common projections in their column-space.

How to make full use to BIG models ?

Cordonnier, Jean-Baptiste, Andreas Loukas, and Martin Jaggi. "Multi-Head Attention: Collaborate Instead of Concatenate." *arXiv preprint arXiv:2006.16362* (2020).

Effectiveness and efficiency in pre-trained models

In MHA, main parameters: $W \in \mathbb{R}^{layers \times 4 \times head\_num \times D_{model} \times D_{head}}$

In GPT3: 96; 4; 96; 12288; 128

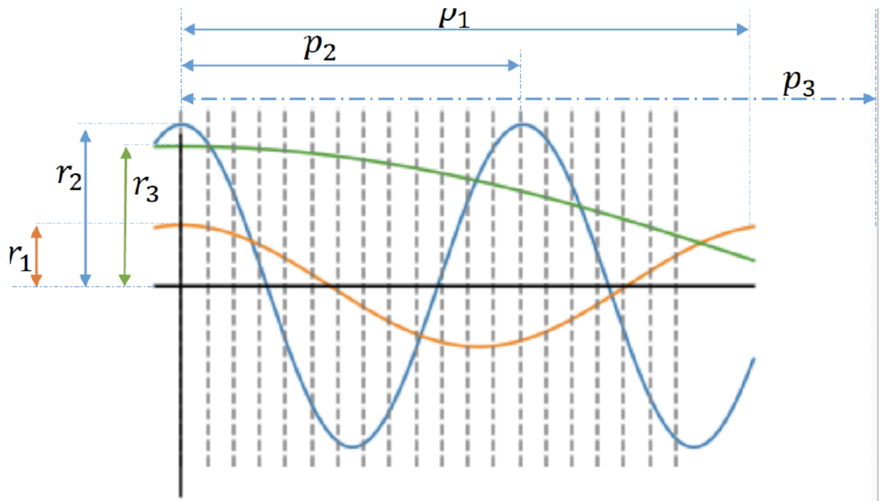
Can we compress it but also keep a relatively-high rank?

Quantum Theory and Natural Language

- Motivations
  - Analogy between QT and NLP
  - (Quantum) probability theory in vector space
  - Paradigm with big models and big data
- Applications in NLP
  - In pre-trained language model
  - Complex-valued representations
  - Tensor network language model

Complex-valued representation in NLP

- Complex word vectors to model words and their positions, this is crucial for bag-of-word network architecture.



- Model a word with individual positions as complex-valued functions (with visualisation of only real part)

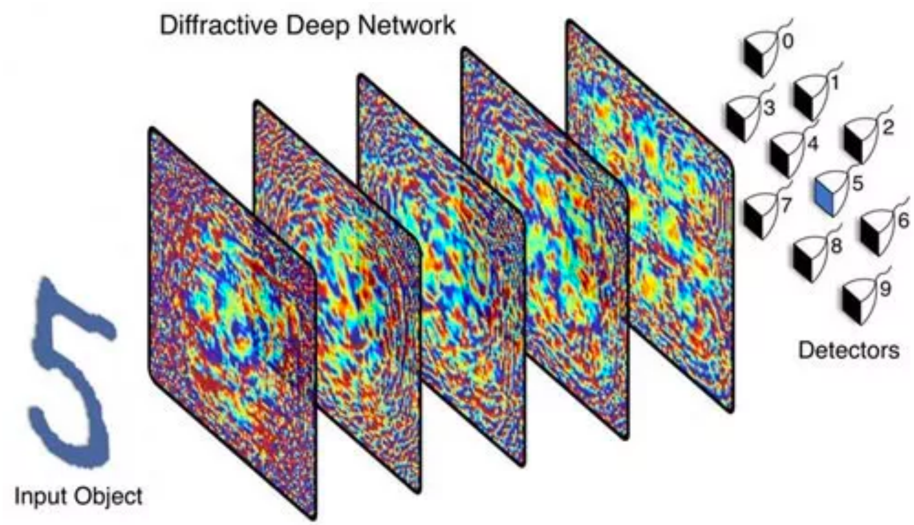
Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., & Simonsen, J. G. (2019, September). Encoding word order in complex embeddings. In *International Conference on Learning Representations*.

asymmetrical relation in KB

- In complex vector space
  - $< x, y > \neq < y, x >$ due to the conjugate transpose
- In knowledge base (KB)
  - Some relations are asymmetrical: *is_the_father_of*
  - Some relations are symmetrical: *is_a_friend_of*

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. International Conference on Machine Learning (ICML).

Physical complex Neural networks



Lin, Xing, et al. "All-optical machine learning using diffractive deep neural networks." *Science* 361.6406 (2018): 1004-1008.

Quantum Theory and Natural Language

- Motivations
  - Analogy between QT and NLP
  - (Quantum) probability theory in vector space
  - Paradigm with big models and big data
- Applications in NLP
  - In pre-trained language model
  - Complex-valued representations
  - Tensor network language model

Tensor network language model

A language model is a mapping

$$f : \underbrace{\mathbb{N}, \mathbb{N}, \cdots, \mathbb{N}}_{\text{n}} \to \mathbb{R}^{+}$$

which aims to give a probability to any N-gram term, resulting a N-order tensor with each dimension of vocabulary size.

See Google N-gram dataset for ground truth.

Tips: Neural Architecture Search (NAS) to select the bond dimensions

This work is in progress

Thanks

Wang@dei.unipd.it

Reference

Benyou Wang, Qiuchi Li, Massimo Melucci, and Dawei Song. *Semantic Hilbert Space for Text Representation Learning.* In *the web conference 2019.*

Li, Qiuchi， Benyou Wang**,** and Massimo Melucci. "CNM: An Interpretable Complex-valued Network for Matching." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4139-4148. 2019.

Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., & Simonsen, J. G. (2019, September). Encoding word order in complex embeddings. In *International Conference on Learning Representations*.