Hybrid Classical Quantum Embeddings for NLP and AI using Hamiltonians

Damir Cavar, James Bryan Graves, Shane Sparks, Koushik Reddy Parukola

Indiana University at Bloomington - NLP-Lab

Embeddings in AI/NLP

Distributional Semantics and Vector Models:

- Word and text meaning encoded in dense vectors:
  - fastText (Bojanowski et al., 2017; Joulin et al., 2018)
  - GloVe (Pennington et al., 2014)
  - Numberbatch (Speer et al., 2017)
  - BERT (Devlin et al., 2019)
  - OpenAI GPT4 or Claude 4 LLMs and GenAI (byte-pair encoding)
- Generating word embeddings and language models:
  - Costly and time-consuming computation
  - Large training and evaluation data sets
  - Many pre-computed models are freely available
- Use-cases, for example:
  - Generic neural or probabilistic NLP methods for text classification, machine translation, ...
  - Lexical- or text-similarity computation in semantic search

Quantum AI and NLP Questions:

- Can classical embeddings and language models be used in QC for Q-NLP/AI/ML?
- How reliable are encoding approaches for mapping classical to quantum embeddings?
- Is there information loss or deterioration of quality in different mapping approaches?

Our Goals

- Identify reliable mapping approaches from classical to quantum embeddings.
- Compare the similarity metric in classical with quantum similarity scores.

- **Mapping Algorithms:** Amplitude Encoding: using real-number and complex-number mapping
- **Storing Quantum States:**
  - Amplitudes: each real-vector dimension mapped to one amplitude
  - Amplitudes: two real-vector dimensions mapped to one complex-number amplitude (1st dimension as real part, second dimension as imaginary part)
  - Hamiltonian matrix stored on classical computer
  - Reduced Hamiltonians using Principal Component Analysis (PCA)
- **Similarity Measures:** SWAP test (for hardware benchmarking)
- **Code-base and dataset for benchmarking** available as part of the:
  - Natural Language Qu Kit (NLQK, `https://nlqk.ai/`)

Data

- 751 nouns from the SimLex-999 dataset Hill et al. (2015)
- 666 noun-pairs
- embeddings for all nouns were retrieved from OpenAI and VoyageAI
- Hamiltonians: 11 and 12 qubits = $2^{12} \times 2^{12} = 4096 \times 4096$)
- Dimensionality reduction using Principal Component Analysis (PCA), reducing to 256×256

Quantum Word and Text Similarities

- Classical embeddings $\rightarrow$ Quantum embeddings
  - **OpenAI GPT Embeddings**, large 3072-dim. and short 1536-dimensional word vectors
  - **Claude 4 (VoyageAI)** embeddings, 1024 and 2048 dimensions
- Amplitude Encoding
- **SWAP Test** (Buhrman et al., 2001)
  - Two circuits S and T with the same number of qubits
  - Measures the difference between S and T
  - Physical SWAP test of two embedded words or texts as hardware **Benchmarking**, measuring the correlation coefficient to classical and simulated similarity scores

Results

- Plain quantum encoding (with padding, normalization): ca. 0.9 correlation coefficient
- Complex quantum encoding (with padding, normalization): ca. 0.9 correlation coefficient
- Full Hamiltonians: > 0.9 correlation coefficient
- Reduced Hamiltonian: not usable at all

Issues

- Computational complexity of the conversion from classical to quantum embeddings
  - Costly computation of similarity scores: padding, normalization, conversion
- Hamiltonians data size

Conclusion

- Result: classical vector similarity using Cosine Similarity and quantum embedding similarity using Quantum similarities
  - Correlation Coefficient in general approx. 0.90 on average for the pre-computed vector models using the `qasm_simulator`
- There is minimal information loss in the encoding process.
- Classical word embedding models can be used in Q-NLP/ML/AI tasks.
- Compression from classical to quantum significant: 3,000 x 32bit to 12 qubits.
- Mapping of classical real-number vectors to complex-number quantum significant: saving 1 bit, BUT reducing the circuit size by 50%.
  - Circuit depth is linearly dependent on the input-vector dimensionality

Availability

- **Data and Code available:** GitHub repo `https://nlqk.ai/`
- **PyPi Python module:** `pip install nlqk`

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Harry Buhrman, Richard Cleve, John Watrous, and Ronald De Wolf. Quantum Fingerprinting. *Physical Review Letters*, 87(16):167902, September 2001. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.87.167902.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein et al., editor, *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)*, pages 4171–4186. Association for Computational Linguistics, June 2019.

Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. pages 4444–4451, 2017. URL `http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972`.

Natural Language Processing Lab

The NLP-Lab (`https://nlp-lab.org/quantumnlp/`):