Interactive Exploration for 60 Years of AI Research

Hendrik Strobelt
MIT-IBM Watson AI Lab
IBM Research
Cambridge, MA 02142
hendrik@strobelt.com

Benjamin Hoover
IBM Research and Georgia Tech
Atlanta, GA 30308
benjamin.hoover@ibm.com

1 Introduction

Research in artificial intelligence has been around for over six decades, and interest in the field is still rapidly growing. A diversification of interests has birthed many sub-fields within AI, making it harder for novices and senior researchers alike to orient themselves and their work within the historical context of ML research. We created an interactive demo to investigate an opinionated selection of papers from the last 60 years. The demo not only reflects on the past, but it also allows users to position abstracts of their own novel ideas into the research landscape carved by the last 60 years of AI publications.

2 Methods

The demo system is based on the S2ORC(Lo et al. [2020]) publication dataset that has been pruned to contain a selection of ML/AI research papers. The selection of papers is then visualized and interactive methods are added to help explore the corpus.

Data Wrangling. The S2ORC data is filtered to contain only paper entries that fulfill two criteria: (1) the publication is categorized as "Computer Science" or "Mathematics" by S2ORC **and** (2) the publication contains at least one match with the following regular expression in the `title` or `abstract` field: `"machine learning|artificial intelligence|neural network|(machine|computer) vision|perceptron|network architecture| RNN | CNN | LSTM | BLEU | MNIST | CIFAR |reinforcement learning|gradient descent| Imagenet "`. This opinionated filtering leads to a subset of ∼300,000 publications. To filter them more we utilize the intra-corpus citation count to include only the top 1 percent cited papers per year. Selecting papers based on yearly statistics is a shallow way to normalize for fluctuating overall citation counts. The resulting `60years` dataset contains 3,309 publications. In addition, each document $d_i = (authors, abstract, title, v)_i$ contains a vector embedding v_i which is derived by applying the `SPECTER` model (Cohan et al. [2020]) on text from the title and the abstract.

Visual Encoding. Each paper is represented as a circle. Its position is determined by a UMAP projection (McInnes et al. [2018]) of v_i, where overlaps are removed by successive applications IPSep-CoLa (Dwyer et al. [2006]). The size of a circle encodes the count of intra-corpus citations of the publication. Its color indicates the quinquennium in which it was published. See Figure 1.

Interaction. The demo provides a rich set of interactions for exploration and discovery. Hovering over a dot reveals details about each paper. Papers can be filtered by quinquennium or by substrings within titles or authors. The user can define a rectangle selection by dragging their mouse, which reveals summary information about the selected subset; e.g., most common title tokens, most common title bigrams or a list of the most cited papers. In addition, users can enter their own abstract which is then encoded using the same sentence embedding and projected as a pin into the landscape of publications (see Figure 1).

Preprint. Under review.

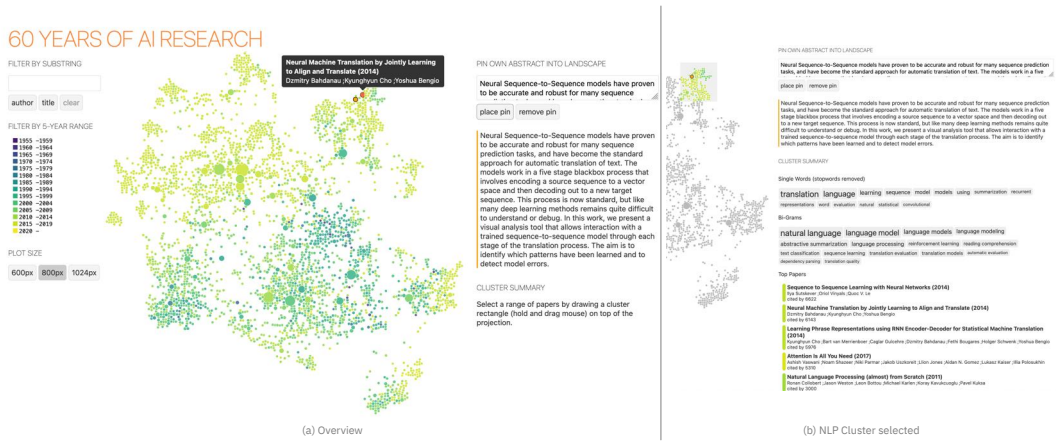

Figure 1: Visualization of the 60years corpus of 3,309 publications in AI/ML. (a) a user defined abstract is positioned on the map (orange) and a neighboring publication is highlighted (red, tooltip). (b) The area around the pin is selected and summarized by most common title tokens/bigrams and most cited articles.

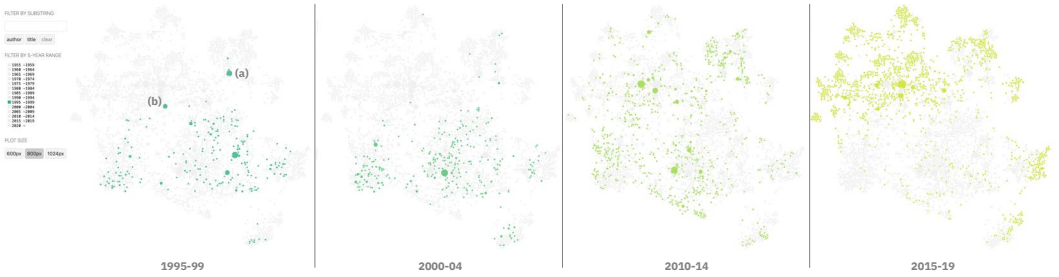

Figure 2: Succession of publications over 5-year periods. Details in the use case section.

3 Use cases

The Delayed Rise of Deep Learning. Figure 2 shows the succession of 5-year periods of publications. In the late 90s, publications like Hochreiter and Schmidhuber [1997] (Figure 2a) or Lecun et al. [1998] (Figure 2b) stick out from the core of ML research. In the succeeding years (2000-09), the area around these deep learning milestones remains quiet. In 2010-14, this neighborhood begins to populate with other strong papers. In 2015-19, the image is nearly inverse to 1995-99, with many publications focusing on deep learning methods and avoiding more traditional ML approaches.

Custom Abstract. In Figure 1a, a user posted their own abstract about visualizing seq2seq models into the pin text box and let the system position the pin for it (orange dot). By hovering over the neighbors, the user gets an idea that the pin is correctly surrounded by NLP publications. When creating a rectangle selection around the pin, terms like "translation" and "natural language" pop up as descriptive terms which matches the users intuition about the context for the paper.

4 Requirements & Links

The demo runs in a web browser (`http://60years.vizhub.ai`) and should be intuitive to use by a broad audience of conference attendees. We will also release the source code to easily replicate our procedure or feed differently filtered (opinionated) corpora into the same interactive visualizer. We will showcase additional use cases, but also emphasize that the selection of papers is biased towards our filtering criteria.

References

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*, 2020.

Tim Dwyer, Yehuda Koren, and Kim Marriott. Ipsep-cola: An incremental procedure for separation constraint layout of graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(5): 821–828, 2006. doi: 10.1109/TVCG.2006.156.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL `https://www.aclweb.org/anthology/2020.acl-main.447`.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL `https://doi.org/10.21105/joss.00861`.