Proceedings of Machine Learning Research 176:278–282, 2022 NeurIPS 2021 Competition and Demonstration Track

Interactive Corpora Visualization for 60 Years of AI Research

Hendrik Strobelt HENDRIK@STROBELT.COM
IBM Research; MIT-IBM Watson AI Lab, Cambridge, MA 02142

Benjamin Hoover BENJAMIN.HOOVER@IBM.COM
IBM Research; Georgia Institute of Technology, Atlanta, GA 30308

Editors: Douwe Kiela, Marco Ciccone, Barbara Caputo

Abstract

Research in artificial intelligence (AI) has been around for over six decades. In that time, it has experienced rich growth, with on and off interest, as researchers tackle this problem from different angles using inspiration from various fields. However, it is difficult to see an overview of the journey that research in AI has taken in its lifespan. We created a visualization we call "60 Years of AI" that explores a (biased) selection of the most influential publications that have shaped the field. Our visualization shows similar works clustered together throughout time and allows users to input abstracts of new ideas to see where their ideas position in the landscape of the ever-growing field of AI.

1. Introduction

Research in Artificial Intelligence (AI) has progressed a long way in its 60 year history and is perhaps experiencing a peak of interest today. However, modern AI research looks noticeably different from the earliest efforts in the field, prompting the question: how can we visualize the journey AI research has taken? We introduce "60 Years of AI Research" which employs modern NLP technologies to place a small sampling of the most influential papers in this field onto a semantically meaningful space where similar efforts and concepts cluster together.

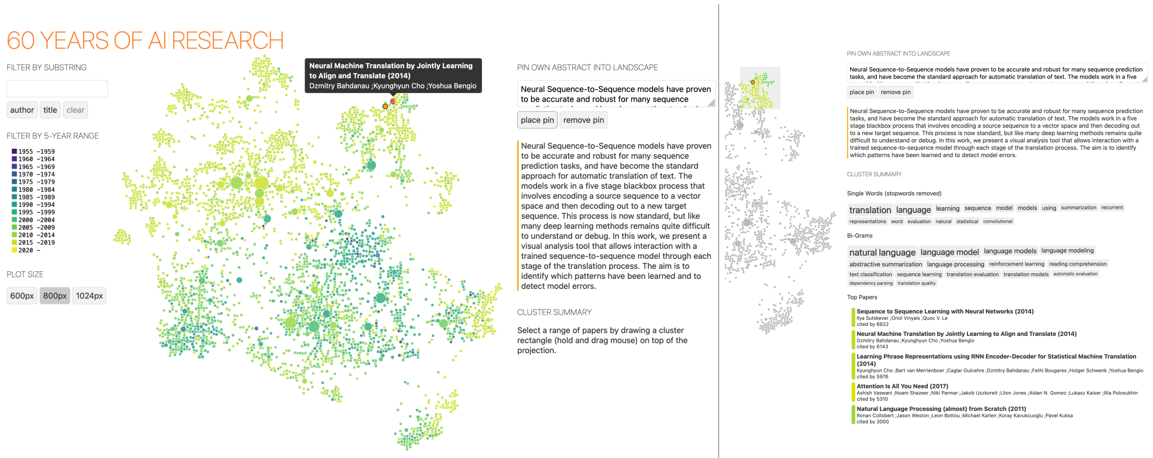
2. Methods

The demo system builds on a pruned subset of the S2ORC (Lo et al., 2020) publication dataset that contains a small selection of ML/AI research papers. This subset is then visualized and paired with interactive methods to help explore the corpus.

2.1. Data Wrangling

The S2ORC data is filtered to contain only papers that fulfill two criteria: (1) publications categorized as "Computer Science" or "Mathematics" by S2ORC **and** (2) publications containing at least one match with the following regular expression in the `title` or `abstract`:

```
machine␣learning␣|␣artificial␣intelligence␣|␣neural␣network
|␣(␣machine␣|␣computer)␣vision␣|␣perceptron␣|␣network␣architecture
|␣RNN␣|␣CNN␣|␣LSTM␣|␣BLEU␣|␣MNIST␣|␣CIFAR␣|␣reinforcement␣learning
|␣gradient␣descent␣|␣Imagenet
```



Neural Machine Translation by Jointly Learning to Align and Translate (2014)
Dzmitry Bahdanau ;Kyunghyun Cho ;Yoshua Bengio

Neural Sequence-to-Sequence models have proven to be accurate and robust for many sequence

Neural Sequence-to-Sequence models have proven to be accurate and robust for many sequence prediction tasks, and have become the standard approach for automatic translation of text. The models work in a five stage blackbox process that involves encoding a source sequence to a vector space and then decoding out to a new target sequence. This process is now standard, but like many deep learning methods remains quite difficult to understand or debug. In this work, we present a visual analysis tool that allows interaction with a trained sequence-to-sequence model through each stage of the translation process. The aim is to identify which patterns have been learned and to detect model errors.

Select a range of papers by drawing a cluster rectangle (hold and drag mouse) on top of the projection.

Neural Sequence-to-Sequence models have proven to be accurate and robust for many sequence prediction tasks, and have become the standard approach for automatic translation of text. The models work in a five

Sequence to Sequence Learning with Neural Networks (2014)
Ilya Sutskever ;Oriol Vinyals ;Quoc V. Le
cited by 6622

Neural Machine Translation by Jointly Learning to Align and Translate (2014)
Dzmitry Bahdanau ;Kyunghyun Cho ;Yoshua Bengio
cited by 6143

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (2014)
Kyunghyun Cho ;Bart van Merrienboer ;Caglar Gulcehre ;Dzmitry Bahdanau ;Fethi Bougares ;Holger Schwenk ;Yoshua Bengio
cited by 5976

Attention Is All You Need (2017)
Ashish Vaswani ;Noam Shazeer ;Niki Parmar ;Jakob Uszkoreit ;Llion Jones ;Aidan N. Gomez ;Lukasz Kaiser ;Illia Polosukhin
cited by 5310

Natural Language Processing (almost) from Scratch (2011)
Ronan Collobert ;Jason Weston ;Leon Bottou ;Michael Karlen ;Koray Kavukcuoglu ;Pavel Kuksa
cited by 3000

(a) Overview

(b) NLP Cluster selected

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.

This opinionated filtering leads to a subset of ∼300,000 publications. We additionally utilize the intra-corpus citation count to include only the top 1% cited papers per year, though selecting papers based on yearly statistics is a shallow way to normalize for fluctuating overall citation counts. The resulting `60-years` dataset contains 3,309 publications. In addition, each document $d_i = (authors, abstract, title)_i$ is associated with a vector embedding v_i which is derived by applying the `SPECTER` model (Cohan et al., 2020) on the combined text from the title and the abstract.

2.2. 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 application of IPSep-CoLa (Dwyer et al., 2006). The size of a circle encodes the count of intra-corpus citations of each publication. Its color indicates the quinquennium in which it was published. See Figure 1.

2.3. Interaction

The demo provides a rich set of interactions for exploration and discovery:

Users can position (pin) their own abstract. To identify the neighborhood of their own work, a user can provide an abstract and the system runs the same embedding model and projection model to determine a position. See Figure 1 (orange circle).

The user can select a cluster of documents by dragging the mouse to create a rectangle selection. A selected cluster is summarized by the most common title tokens and

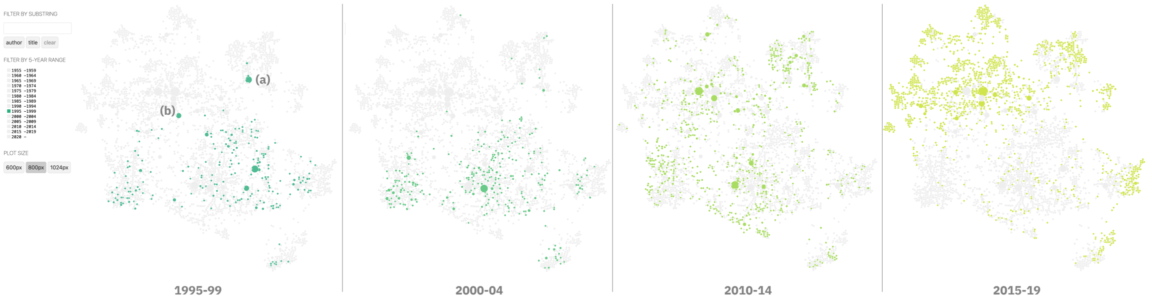


Figure 2: Succession of publications over 5-year periods. Details in the use case section.

title bigrams, a list of most cited papers, and a histogram of the yearly distribution of publications in the cluster.

Papers can be filtered by publication year or by substrings in titles or authors. The yearly filter can be animated to observe progression patterns.

Hovering over a circle reveals details about each paper. If a URL or DOI is available in the dataset, the publication can be opened directly by clicking.

3. Use cases

3.1. The Delayed Rise of Deep Learning

Figure 2 shows the succession of 5-year publications periods. 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 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 that of 1995-99, where many publications focus on deep learning methods and avoid more traditional ML approaches.

3.2. Custom Abstract

In Figure 1a, a user posted a custom abstract about visualizing seq2seq models 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 matching the user’s intuition about the context for the paper.

4. Generalized Use of the Visualization

The described system and interactive methods run in a web browser and can be applied to new collections of academic documents to visualize more than just our opinionated selection of papers from the past 60 years of AI research.

For example, Figure 3 shows how we applied the visualization to the entire anthology of papers published at the NeurIPS conference[1] of size 13,000 documents or the IEEE VIS

[1] https://neuripsav.vizhub.ai/

visualization conference[2]. In these corpora, citation information is absent and all papers are displayed as identically sized circles.

5. Conclusion

While search and filter operations are efficient tools to find particular publications in large corpora, they often fail to provide an extensive overview. For observing a longitudinal trend and the partitioning of AI research in the last 60 years we think that our demo could be a starting point for investigation. The filter criteria we used are driven by practical considerations and provide an equally truthful and false view on the large landscape of papers that contributed to the development of AI research. The methods and system are open-sourced[3] and can be applied to more corpora.

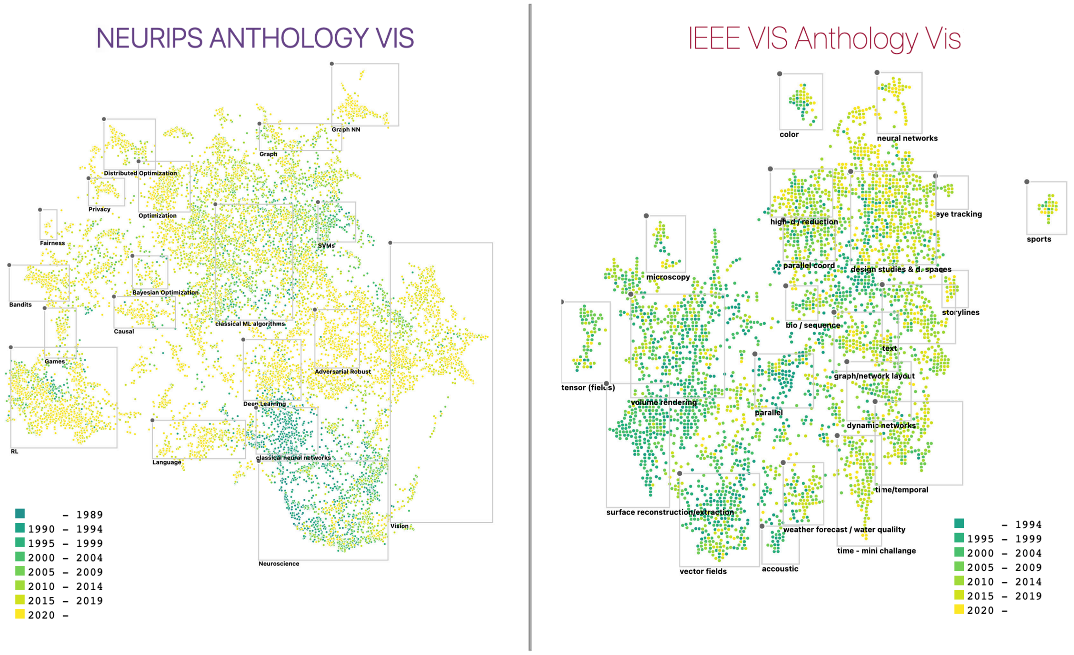


Figure 3: The visualization system applied to the papers of the NeurIPS Conference (left) and the IEEE Vis Conference (right).

Acknowledgments

Special thanks to Konrad Kording, who inspired the idea to spent time on an overview system like our demo. Part of this work was funded by the MIT-IBM Watson AI Lab.

[2] https://visav.vizhub.ai/
[3] https://github.com/HendrikStrobelt/visual-anthology-backend

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.