Identify Emerging Trends In Machine Learning Using Published Literature

Yunwei Hu, Cheng Zhan, Jielin Xu, Yu Zeng, Junjie Yu, Guo Yu, Licheng Zhang

This work is done by the authors in their spare time, and is not related their employment.

Motivation

As the field of Machine Learning (ML) is growing rapidly, it becomes ever more difficult to digest the incoming information and thereby identify emerging topics that will have a long-term impact, even for ML practitioners. We propose a ML based system that analyzes the published literature to identify trends.

Data

- Use arXiv dataset for published literature
- Open access to scholarly articles, from the vast branches of physics to the many subdisciplines of computer science. Rich corpus of information.
- Fee, open machine-readable arXiv dataset on Kaggle: a repository of 1.7 million articles
  - features include: article titles, authors, categories, abstracts, and published dates.
- Cross link with Semantic Scholar API for citation data.

Preprocessing

- Scoring papers based on citation.
- Calculate the z-core for each paper. $Z = \frac{x - \mu}{\sigma}$
- The z-score has been shown as good indicator of the impact of a paper.

Methodology

The workflow that we have developed automatically uncovers important research trends by periodically analyzing the arXiv publications and other open academic literature search engines.

Traditional Topic Modeling methods rely on counting words and extract similar word patterns to infer topics.

Pre-Trained Models (BERT) are expected to provide more accurate representations of texts than bag-of-words.

Inspired by

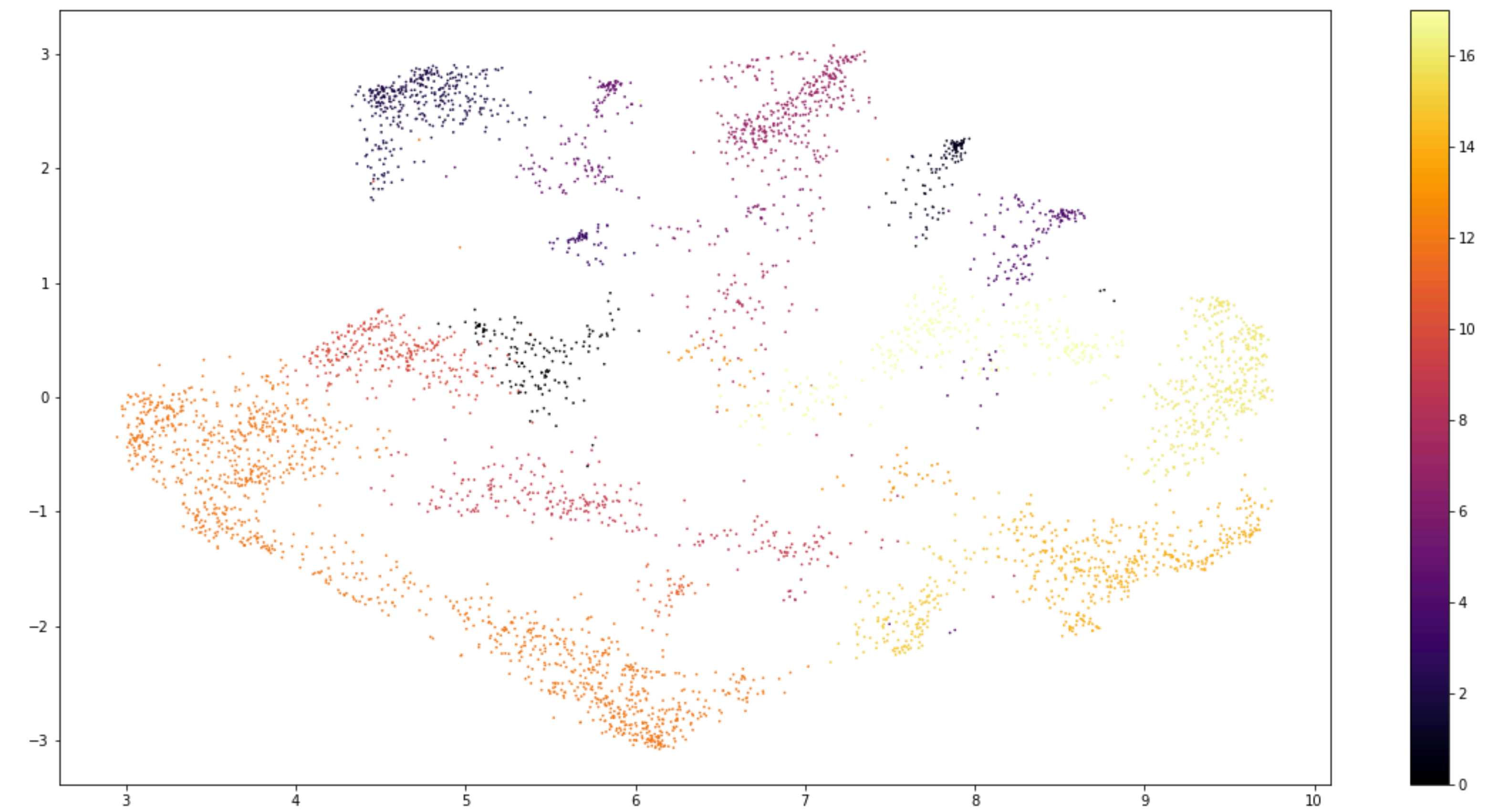
- Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *arXiv preprint arXiv:2008.09470*

- *https://github.com/MaartenGr/BERTopic*

Topic Modeling Steps	Application in Indentifying Machine Learning Trends
Get list of Documents	Get papers from arXiv (cs.LG, stat.ML, and cs.AI) Get citations data from Semantic Scholar Calculate Z-score Filter z-score >0 Use titles and abstract as the doc list
Encoding	Use sentence transformer encoder SciBERT is trained on papers from the full text corpus of semanticscholar.org - Generate a vector of 512 for each paper.
Find Cluster	Dimension Reduction: **UMAP** - Hyperparameter: n-component, etc - Dimensions depends on the clustering tool. (HDBSCAN can do well on up to around 50 or 100 dimensional data, but performance can see significant decreases beyond that.) Clustering: HDBSCAN - Density based
Find Topics for each Cluster	Class based TF-IDF • Get the top words per cluster based on their c-TF-IDF scores. $\text{c-TF-IDF}_i = \frac{t_i}{w_i} \cdot \log \frac{m}{\sum t_j}$
(Optional) Optimize results	Topic Generation - From topic words to themes - Reduce topics (based on distance between clusters.)

Visualization of Topic Clusters

Illustration only. Visualization used 2-d UMAP, while clustering was calculated in higher dimensional space.



18 Topic Clusters Identified

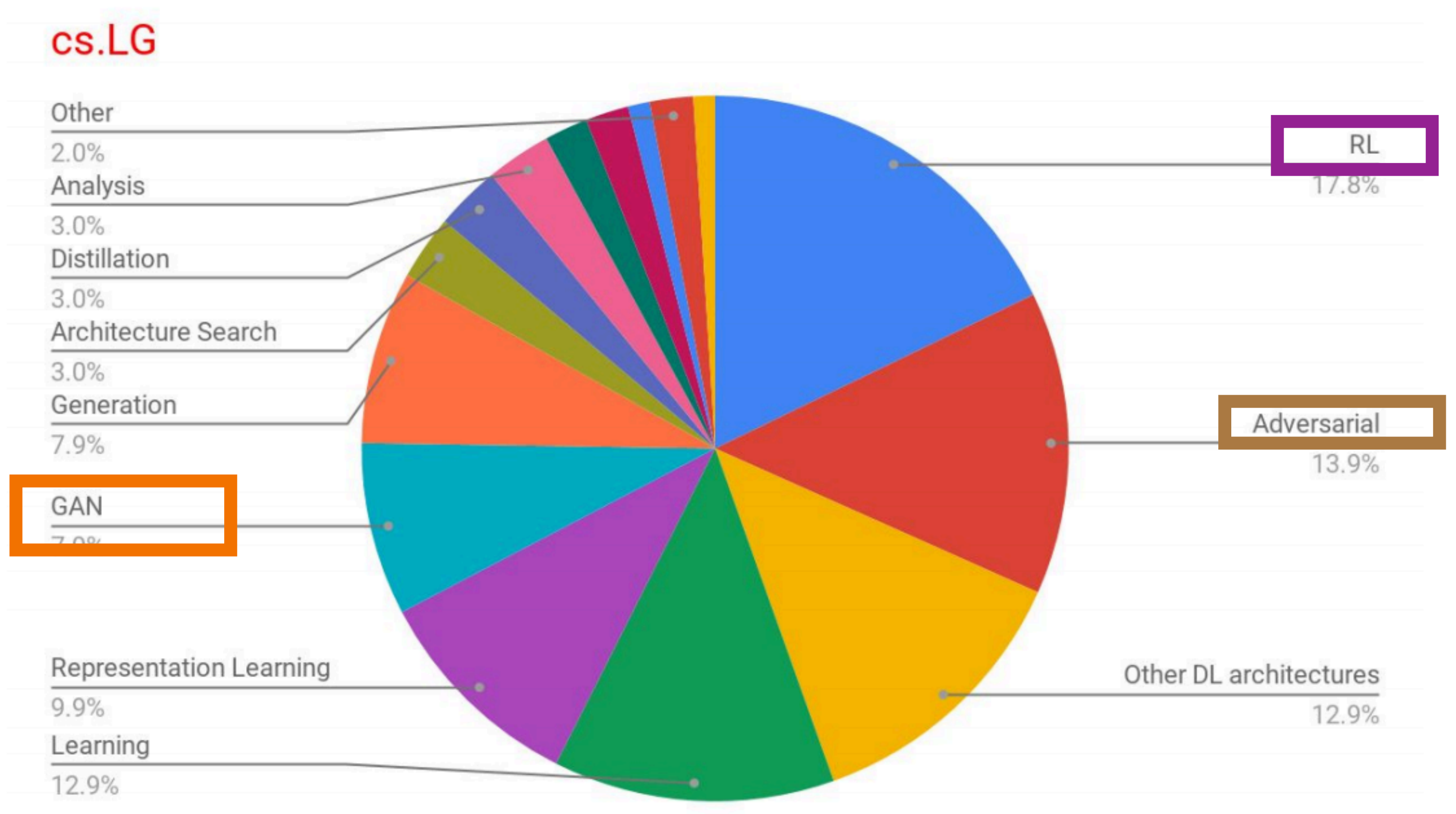
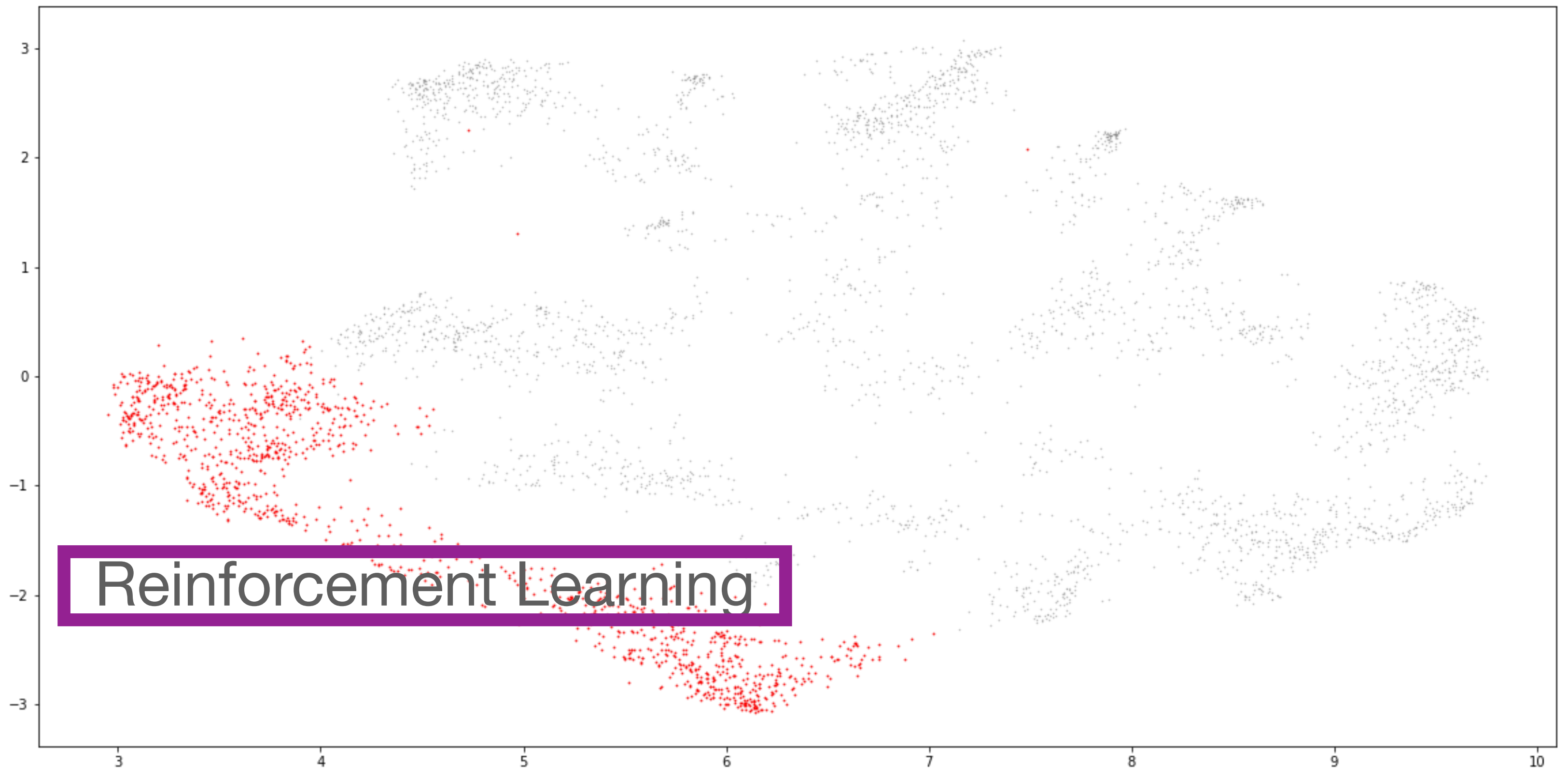
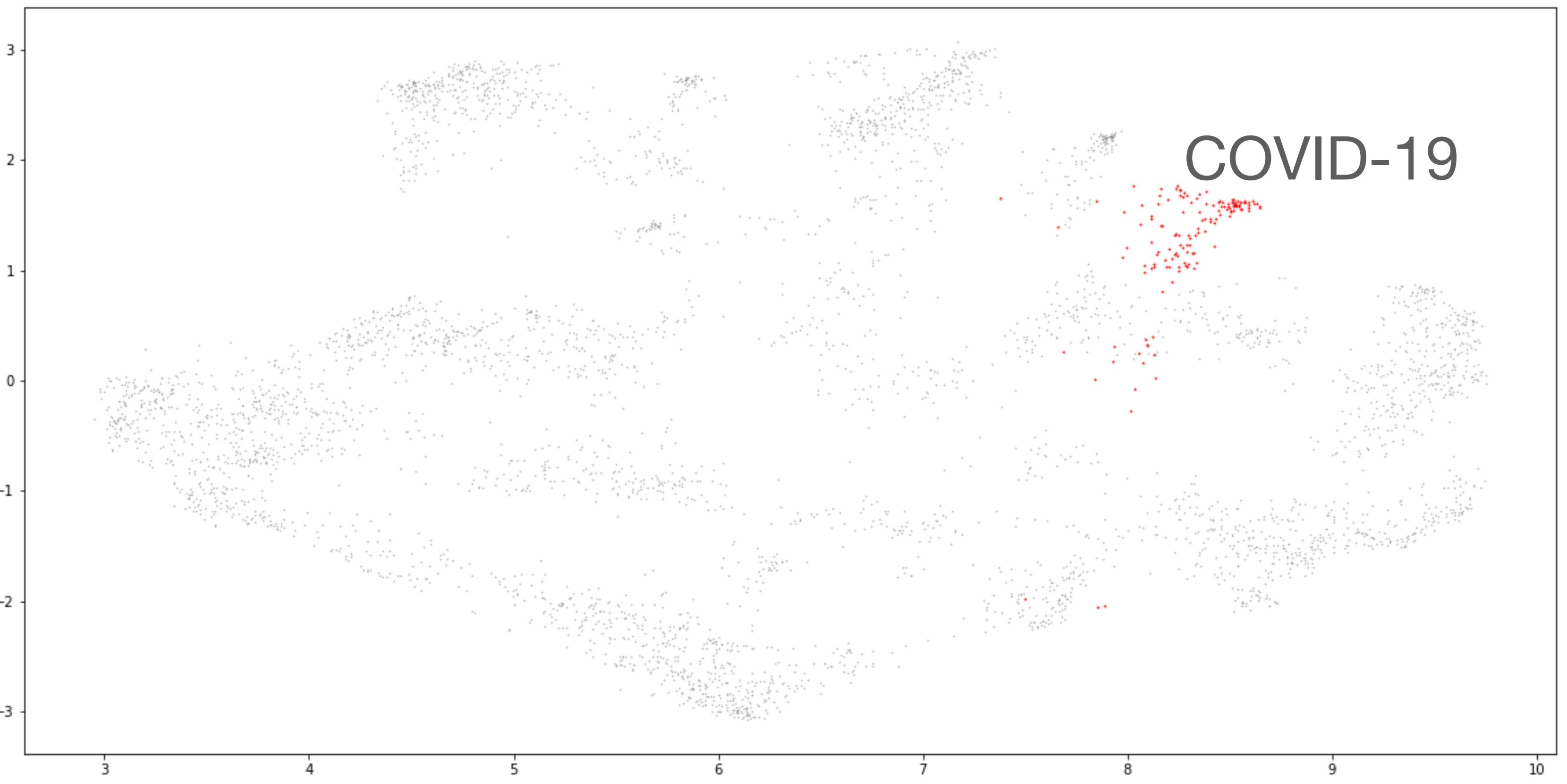


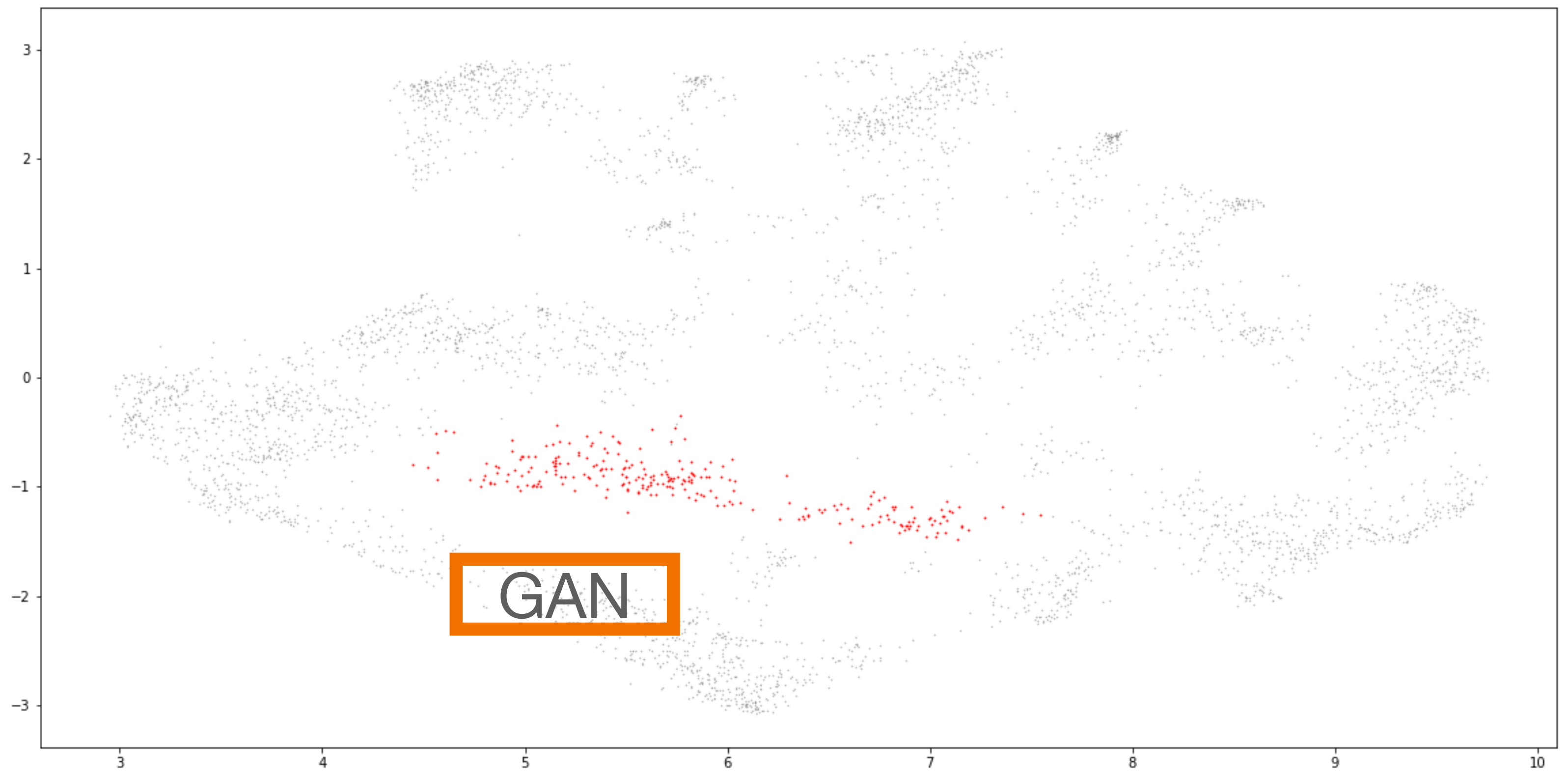
Fig. 2. method distribution in top-100 list in cs.LG.

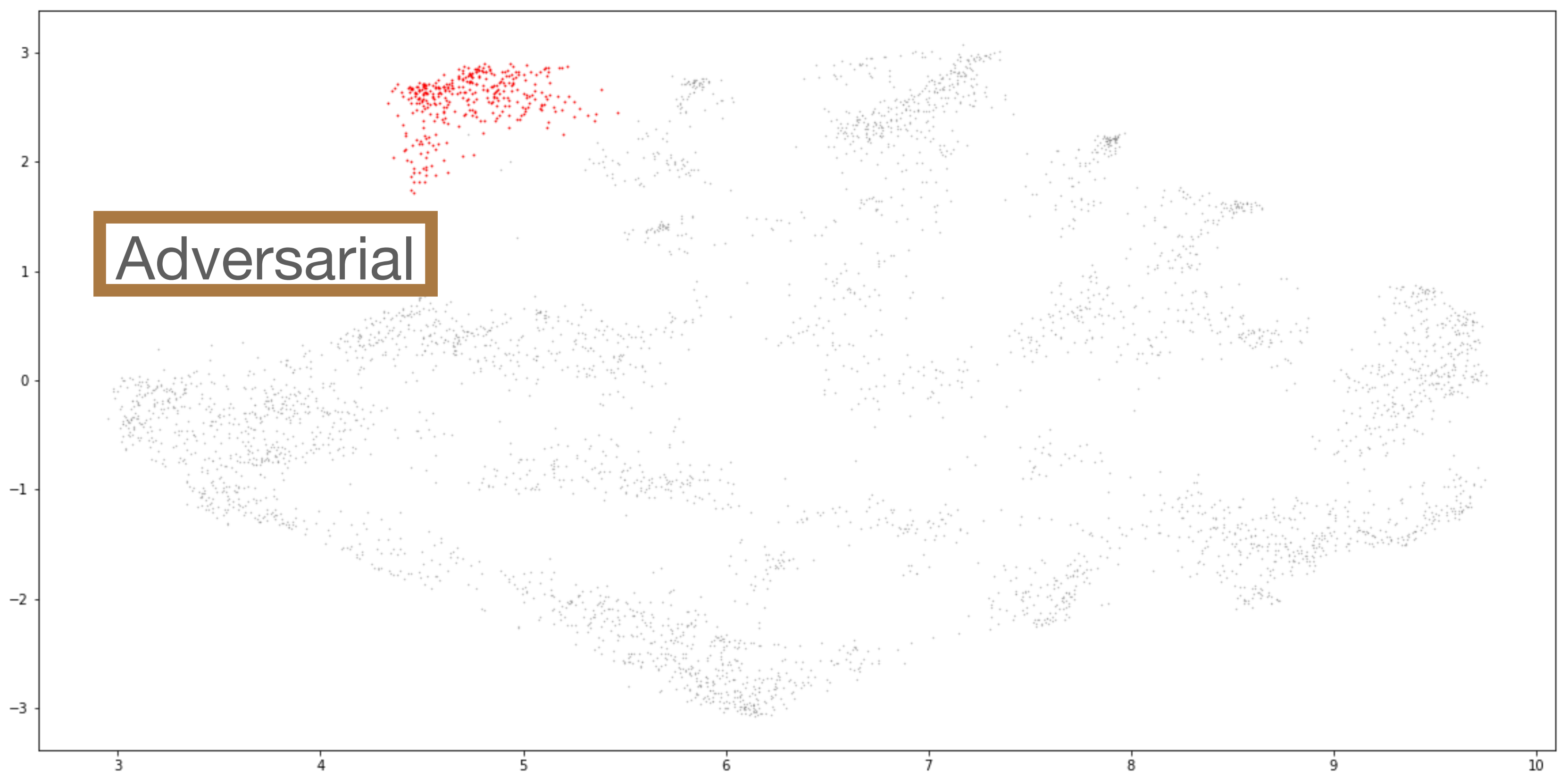


Trends identified by Eger et al. in 2019[1] by manually examine 100 papers.

"Figure 2 shows the method distribution in the top-100 list of cs.LG. The most important methods here are **Reinforcement Learning** (RL) and **Adversarial** techniques. … Generation is likewise prominent, through Generative Adversarial Net- works (**GAN**s)"







1. S. Eger, C. Li, F. Netzer, I. Gurevych, Predicting Research Trends From Arxiv, https://arxiv.org/abs/1903.02831

Work in Progress

This project is still evolving. Please check our GitHub for latest:
https://github.com/huyunwei/TopicModeling
We are making experiments in following area:

Metric

- Quantitative metric to measure the results
- Use the metric to automate model selection and hyper parameter tuning

Data Source

- The arXiv is the most popular open platform, while many influential papers were published in top conferences.
- Title only clustering, or full text clustering

Encoding

- Domain specific pre-trained BERT would be preferred.
- Other Encoding methods
- Knowledge Graph

Dimension Reduction and Clustering

- Other Dimension Reduction methods
- Other clustering methods (Gaussian Mixture)
- Clustering without dimension reduction.
- Supervised learning

References

1. S. Eger, C. Li, F. Netzer, I. Gurevych, Predicting Research Trends From Arxiv, arXiv:1903.0283, https://arxiv.org/abs/1903.02831
2. M. Dong, X. Cao, M. Liang, L. Li, G. Liu, H. Liang, Understand Research Hotspots Surrounding COVID-19 and Other Coronavirus Infections Using Topic Modeling, https://doi.org/10.1101/2020.03.26.20044164
3. Bianchi, S. Terragni, D. Hovy, Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence, arXiv:2004.03974
4. M. E. J. Newman, Prediction Of Highly Cited Papers, Europhys. Lett. 105, 28002 (2014), https://arxiv.org/abs/1310.8220
5. Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *arXiv preprint arXiv:2008.09470*
6. *Maarten Grootendorst. BERTopic,* https://github.com/MaartenGr/BERTopic