nferX: a case study on data-centric NLP in biomedicine

David Chang, Vineet Mathew, Lorenzo Kogler Anele, Roger Jin, Krishna Rao, Bharathwaj Raghunathan, Wui Ip, Zainab Doctor, Colin Pawlowski, Ajit Rajasekharan,
nference, Inc.
{dchang, bharath}@nference.net

Abstract

The growing prevalence of AI in industry and the dominance of a handful of model classes have contributed to a community-wide shift toward more data-centric AI. As an AI-driven biotech company unparalleled in the range and quantity of its biomedical data, nference has built the nferX platform housing many of our ML tools to facilitate both data-driven scientific discovery and healthcare AI product development. In this case study, we provide an overview of the nferX platform and its application in biomedical NLP, with an emphasis on methods for increasing labeling efficiency in our model development pipeline.

1 Introduction

Developing the appropriate tools, processes, and infrastructure for successful AI product development has become a major challenge among companies trying to leverage recent advances in AI for their business aims. The problems and solutions encountered along that journey fall under the umbrella of machine learning operations (MLops), a term that emphasizes the holistic nature of a functional ML product pipeline. Data-centric AI is a movement that has been gaining momentum, especially with the advent of key classes of models (e.g. pretrained language models and CNNs) and libraries (e.g. Huggingface transformers) that take care of most modeling needs, where we shift the focus from model architecture design to actual data-first exploration and development. This is especially important in industry settings in which serious data science work focused on complex, real-world data can yield significant value and insight.

2 Biomedical NLP on the nferX Platform

Given the variety of publically available literature and molecular sources synthesized by the nferX platform, a broad range of biomedically related model-training tasks can leverage the nferX platform to identify token collections of interest and retrieve relevant text fragments. For many biomedical or clinical tasks, it is necessary to fetch sentences containing entities of a particular type (e.g. identifying sentences where a drug interacts with a protein or fetching patients who have been diagnosed with a specific disease). To that end, the **Nucleus Knowledge Graph** can be used to specify token collections of interest, be it drugs, cell lines, or medical devices. After identifying relevant token collections, sentences containing the relevant tokens can be retrieved using the **Get Sentences** API, a service that splits the entire text corpus into fragments and then returns the relevant sentences along with meta-information about the source document (e.g. title, publication date, etc.)

Once the sentences with the relevant entity types have been collected, they are used to create datasets that are then uploaded to **AI Studio**. AI Studio is an nferX application that facilitates model development at scale. Through an intuitive UI, various types of datasets (e.g. text and image) can be tagged by assigned taggers. Data taggers are typically medical students or residents and must pass

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

a threshold of accuracy on a golden standard certification set in order to contribute to a project. In addition, AI Studio's UI also allows users to easily configure, train, and deploy various in-house deep learning models on the uploaded datasets.

The nferX platform provides the infrastructure to not only carry out scientific research [1, 2] and ML projects at scale but also to do so in an accessible way that allows non-ML practitioners to easily integrate ML components into their workflow. One of the main issues, though, is labeling efficiency, and methods to increase labeling efficiency can directly cut down costs and improve the overall pipeline.

3 Clustering-based Sampling for Improved Label Efficiency

To better utilize the tagging efforts that are made in the AI Studio suite, we must curate the data that is passed down to taggers for manual review. By selecting diverse, relevant text fragments, our AI models are able to train more efficiently and generalize to most edge cases in the universe of data. This curation is a key aspect in a data-centric pipeline, as it allows us to allocate resources such that efforts are not wasted in tagging and reviewing very similar sentences.

Clustering is a natural way to coordinate this data filtering and selection process. By grouping together similar data points, we are able to gather information about similarities and patterns in the data, and consequently make better decisions about what is relevant enough to be tagged. The underlying rationale behind this process is simple: we should only pick a few points that belong to the same cluster.

To achieve this, our clustering approach relies on looking at the distribution of the pairwise similarities between data embeddings. For a given point, we use the mean and standard deviation of its similarity distribution with other points to obtain a cutoff value, such that any other point with a larger cutoff belongs to this cluster. Visually, this is equivalent to covering the data space with spheres whose radii depend on how similar a point is to others. This process ensures that outliers are also considered, since we allow for singleton clusters. We tune the radius using a hyperparameter, which allows for some flexibility if one desires a more granular or coarse set (see Appendix).

Using this approach, we are then able to make informed decisions about how to select relevant sentences for tagging. Using cluster-based sampling ensures that all sentences that are passed down for review are relevant and diverse and cover a large proportion of the embedding space. From a dataset of 400,000 samples, we are able to tag 1,000 selected points such that their cluster coverage is 90% of the dataset, as shown in the Appendix. Similar approaches have been used in settings such as active learning [3] to further boost model performance.

4 Discussion and Future Work

The components of our nferX platform and the methods described above address several challenges in MLops and data-centric AI: the synthesis of a wide variety of data sources, data generation and labeling, data quality, model training and iterative refinement, and labeling efficiency. Furthermore, building this platform has not only helped improve model performance, data quality, and efficiency, but also empowered data scientists to maximize their contributions. The next steps in our ML journey will involve exploring domain-specific data augmentation approaches, implementing an artifact and metadata management system to more effectively scale up projects, and developing more sophisticated preprocessing methods to improve the quality of the constructed datasets.

References

[1] Jiho Park, Agustin Lopez Marquez, Arjun Puranik, Ajit Rajasekharan, Murali Aravamudan, and Enrique Garcia-Rivera. Recapitulation and retrospective prediction of biomedical associations using temporally-enabled word embeddings. *bioRxiv*, 2019.

[2] Tyler Wagner, Samir Awasthi, Gayle Wittenberg, AJ Venkatakrishnan, Dan Tarjan, Anuli Anyanwu-Ofili, Andrew Badley, John Halamka, Christopher Flores, Najat Khan, Rakesh Barve, and Venky Soundararajan. Real-time biomedical knowledge synthesis of the exponentially growing world wide web using unsupervised neural networks. *bioRxiv*, 2020.

[3] Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. Cold-start active learning through self-supervised language modeling, 2020.

A Appendix

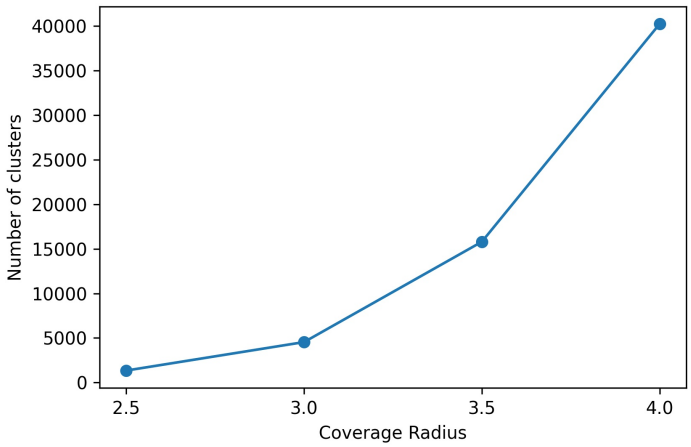

Figure 1: Clustered dataset granularity with different parameter settings.

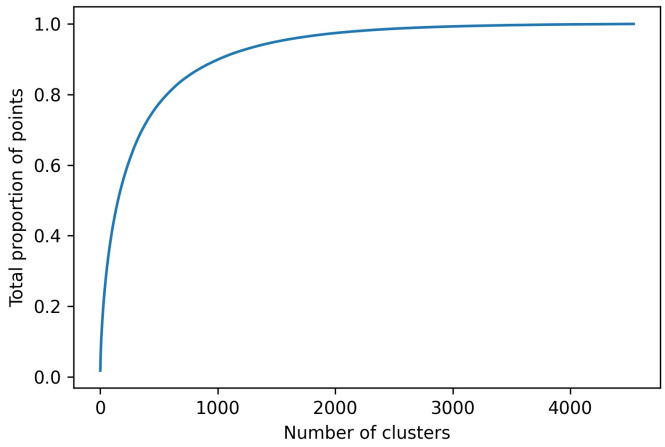

Figure 2: Cumulative dataset coverage by clusters.