

Team members: Almas Abdibayev, Sungil Ahn

Paper: Optimal Data-Dependent Hashing for Approximate Near Neighbors

Why we chose this paper:

Almas: I picked this paper since over the summer I have worked on an NLP research project where as a part of the pipeline I needed to compute Jaccard distances efficiently. That is when I came across LSH. I am also briefly familiar with Ilya Razenshteyn's work and so I was interested in going in depth to understand his work.

Sungil: as someone with limited computer science background, I wasn't familiar with a lot of the topics the papers covered. However, I did come across ANNs in a class I took (76) so the topic was familiar to me. Still, my experience with ANNs in that class only amounted to simply *using it*, so I was interested in learning about the method used to efficiently solve ANNs (LSH) and seeing how it could be further improved upon.

What we'd do in the project:

Almas: Aside from the division of labor (dividing chapters) based on our respective skillsets, I am interested in application of these methods to deep learning. More specifically, "The Case for Learned Index Structures" showed that cache optimized hashing can be outperformed by correctly tuned neural networks. The paper we will be reporting on also focuses on data-dependent hashing, so I'd be interested in extending the ideas of this paper with deep learning methods since one the strengths of using neural networks is learning subtle statistical regularities present in the data. This in turn can be used to improve the heuristics used for hashing the sketches and speeding up the approximate nearest neighbors search.

Sungil: Of course, we both have to read and understand the paper; however, reporting and presenting the analysis is – in my opinion – a creative endeavour, one that I am used to. Thus, I would focus on organizing the bits of information we can get off the paper and any other relevant information (such as previous attempts at solving ANNs efficiently) and producing coherent reports/presentations based on it. Additionally, as I am a math major, I could help Almas understand any confusing "math bits".