Final Project Topics

Your final project proposal is due April 18 Before then, you need to find/finalize your group members, pick a topic, and download the dataset you want to work with.

If you already have a group, you should meet with them to discuss project ideas. Then either as a group or individually, you should fill out this survey with some thoughts about what you're interested in.

Your project in total will be worth at least 32% of your grade, broken down as follows:

- 8% for the proposal, due April 18
- 8% for the update, due May 11
- 16% for the report, due during finals

If you want to focus more on the project, you can choose to increase its weight by:

- 8% for a literature review which replaces two of your readings
- 8% for extensions that replace one homework assignment

If you want to replace a homework with project work, you need to submit a "bonus proposal" describing what you hope to do.

Three general kinds of projects

- Write code to reproduce a landmark research paper
  For example, you might write code to implement a Transformer model from scratch as introduced in the paper, “Attention is All You Need.”

- Find a dataset you're interested in, then figure out what model to use on it. For example, suppose you're interested in a dataset of restaurant reviews, and want to predict the numerical rating based on the review text. Maybe you decide you want to compare the performance of an LSTM and a

Transformer model.

- Find a model you're interested in, then apply it to a new dataset.
  For example, maybe you want to understand how a GAN works. You might decide to download a pretrained model and explore how its performance changes when being fine-tuned on a new data source.

Models and concepts you might be interested in

This is by no means an exhaustive list, but we will cover each of these in at least one lecture.

1. Convolutional Neural Networks, e.g.,
2. Recurrent Neural Networks, e.g., LSTMs
3. Transformer Models, e.g., GPT
4. Diffusion Models, e.g., Dall-E
5. Autoencoders
6. Reinforcement Learning

Datasets you might be interested in

This is obviously not an exhaustive list, but hopefully provides some interesting options and inspiration for other resources.

1. Medical and Genetics Applications

- 10x Genomics https://www.10xgenomics.com/resources/datasets
- Links to an external site.
- Links to an external site.
-
- Genotype-Tissue Expression https://www.gtexportal.org/home/datasets
- Links to an external site.
-
- Protein Classification http://scop.mrc-lmb.cam.ac.uk
- Links to an external site.
-
- Medical image analysis
  https://paperswithcode.com/datasets?mod=medical&page=1
- Links to an external site.
- Links to an external site.
-

2. Natural Language

- Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Database_download
- Links to an external site.
-

- Sentiment prediction: http://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences
- Links to an external site.
-
- Reddit: https://huggingface.co/datasets/reddit
- Links to an external site.
-
- Hate speech on Twitter https://data.world/thomasrdavidson/hate-speech-and-offensive-language
- Links to an external site.
-
- Machine Translation: https://huggingface.co/datasets/wmt16
- Links to an external site.
-

3. Image classification

- Many options: https://huggingface.co/datasets?task_categories=task_categories:image-classification&sort=downloads
- Links to an external site.
-
- So many options: https://paperswithcode.com/datasets?task=image-classification
- Links to an external site.
-

4. Finance and economics

- World Bank Open Data: https://data.worldbank.org/
- Links to an external site.
-
- International Monetary Fund Data: https://www.imf.org/en/Data
- Links to an external site.
-
- Stock prices from pandas_datareader
- Links to an external site.
-

5. Other things

- Climate Modeling: https://wiki.climatechange.ai/wiki/Climate_Modeling_and_Analysis#Data
- Links to an external site.
-

- Energy and Emissions: https://wiki.climatechange.ai/wiki/Electricity_Systems#Data
- Links to an external site.
-
- Lie Detection from Videos: http://iab-rubric.org/index.php/bag-of-lies
- Links to an external site.
    - Note: we already have access to this data

If you have something else in mind, there are:

- another 18,789 datasets available at https://huggingface.co/datasets
- Links to an external site.
-
- another 7,624 datasets available at https://paperswithcode.com/datasets
- Links to an external site.
-