Final Project

The University of Texas at Austin
CS395T/EE381V: Spoken Language Technologies
Fall 2022
Instructor: David Harwath

Assignment Date: October 11, 2022

Due Date: December 5, 2022

Overview: The final project will give you an opportunity to independently explore speech technology via a research project.

The final project can full under either of the following 2 categories:

1. Choose a problem in speech processing, propose a modeling solution, implement the model and evaluate it experimentally.Your project does not need to be an original research idea and can build upon existing work, but you are more than welcome to investigate a novel idea as well. Groups up to 3 students permitted. Final project report to be 4-8 pages using the NeurIPS format.
2. Write a survey paper that summarizes the research literature in a specific sub-area of speech and language processing. Groups up to 2 students permitted. Final project report to be 6-10 pages using the NeurIPS format.

The NeurIPS paper template is available on Overleaf at `https://www.overleaf.com/latex/templates/neurips-2020/mnshsmqkjsqz`. In addition to the final write-up, students must turn in a 1/2 page project proposal, a 1/2 page checkpoint report, and present their work to the rest of the class. **The final project is worth a total of 400 points, comprising 40% of the course grade.**

To turn in: You should submit a single .pdf writeup of your final report to Canvas. You do not need to turn in your code or anything else besides your final report.

Collaboration Policy: You may collaborate in groups of no larger than 3 students. **Each group should only submit one shared final report.**

Late policy: Because the final project is due at the end of the term, late days cannot be applied to it.

Logistics and Deliverables

The final project will have four deliverable components:

- Project Proposal: A 1/2 to 1 page proposal that details the members of your project team and outlines what you will do for your project. This is due on Canvas on October 18. **(worth 25 points)**
- Progress Report: A 1/2 to 1 page update that describes the progress you have made so far, and outlines your plan for completion of the project. This is due on Canvas on November 8. **(worth 25 points)**
- Final Report: A 4-8 page report (for a coding/implementation project) that introduces the problem you worked on, references related/background work, describes the technical details of your approach, presents experimental results, and draws conclusions from those results. This is due on Canvas on December 5. **(worth 300 points)**
- Final Presentation: During the regular class time on November 29 and December 1, project groups will present their work to the class using a Powerpoint slide deck (or similar). These presentations should be short - aim for approximately 7 minutes. **(worth 50 points)**

Grading Rubric

The proposal and progress reports should be easy points; as long as you turn documents that clearly articulate your plan and progress (respectively) you should get full credit for these.

The final presentation will be graded according to clarity of your talk (20 points), coverage of the material in your writeup (20 points), and quality of your slides (10 points)

For an implementation-based project, the final report will be graded according to:

- Definition and explanation of the problem you are working on (25 points)
- Discussion of background and citations to related work (25 points)
- Detailed description of the technical aspects of what you did (models, algorithms, etc.) (100 points)
- Discussion of the experiments you performed, including 1) a description of the dataset(s) you used, 2) motivation for why you chose to perform the specific experiments you did, 3) presentation of the experimental results in charts/tables/graphs, 4) discussion of the experimental results in the text (100 points)
- A concluding discussion that summarizes your results, highlights the key takeaways, and highlights areas for future work. (25 points)
- Overall style, organization, grammatical correctness, etc. of the writeup. (25 points)

For a survey-based project, the final report will be graded according to:

- Definition and explanation of the problem you are surveying (25 points)
- A thorough and well-organized discussion of prior work on your topic (150 points)

- A concluding discussion that summarizes the progress that has been made so far on the problem you are surveying, highlights the key takeaways and/or trends in the research, and highlights areas for future work. (75 points)
- Overall style, organization, grammatical correctness, etc. of the writeup. (25 points)

Project Ideas

I have chosen several "default" projects that you may choose from, but you may also propose your own project idea.

Default project 1: Investigating models of visually-grounded speech

In this project, you will build neural models that attempt to directly map spoken image captions to the images they describe, without performing speech recognition as a first-pass step. You can use the Flickr8k, Places Audio, or SpokenCOCO datasets for this project. All are open-source, but vary in size - Flickr8k is several gigabytes in size, whereas the others are hundreds of gigabytes. You should choose the dataset(s) that the computational resources you have at your disposal allow. For your project, you can use the code linked to below as a starting point, and propose some novel improvement or tweak to the model or learning algorithm in an attempt to boost the retrieval accuracy.

Relevant papers:

1. David Harwath and James Glass, "Deep Multimodal Semantic Embeddings for Speech and Images," Proceedings of ASRU 2015 (`https://www.cs.utexas.edu/~harwath/papers/Harwath_ASRU-15.pdf`)
2. David Harwath, Antonio Torralba, and James Glass, "Unsupervised Learning of Spoken Language with Visual Context," Proceedings of NeurIPS 2016 (`https://www.cs.utexas.edu/~harwath/papers/Harwath_ASRU-15.pdf`)
3. David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass, "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input," International Journal of Computer Vision, 2019 (`https://www.cs.utexas.edu/~harwath/papers/Harwath_IJCV_2019.pdf`)
4. David Harwath, Wei-Ning Hsu, and James Glass, "Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech," Proceedings of ICLR 2020 (`https://www.cs.utexas.edu/~harwath/papers/learning_hierarchical_discrete_linguistic_units_from_visually_grounded_speech.pdf`
5. Gabriel Ilharco, Yuan Zhang, and Jason Baldridge, "Large-scale representation learning from visually grounded untranscribed speech", Proceedings of CoNLL 2019 (`https://arxiv.org/abs/1909.08782`)

Relevant datasets: `https://groups.csail.mit.edu/sls/downloads/placesaudio/`

Relevant code base: `https://github.com/wnhsu/ResDAVEnet-VQ`

Default project 2: Investigating speech recognition models on the Librispeech dataset

In this project, you will benchmark state-of-the-art speech recognition models on the Librispeech dataset, which is an open-source speech recognition dataset based on public-domain audiobooks. To do so, you should use a toolkit such as Kaldi (for HMM-DNN hybrid models) or either Flashlight or ESPNet (for end-to-end neural network models). These toolkits ship with recipes for Librispeech that more or less work out of the box, so after getting a baseline recognizer up and running you should choose an improvement based upon a recently published paper to integrate into the recognizer recipe. You are also free to propose and pursue a novel improvement! Librispeech is a large dataset (1000 hours of speech), but there are several well-defined subsets that are much smaller (10-100 hours). You should work with the largest subset of the data that your computational resources will allow.

Relevant papers and other documentation:

- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," Proceedings of ICASSP 2015 (`https://www.danielpovey.com/files/2015_icassp_librispeech.pdf`)
- Look through the Librispeech leaderboard papers here: `https://github.com/syhw/wer_are_we`

Relevant datasets: You shouldn't have to manually download Librispeech, but for reference, the Librispeech dataset is available at `https://www.openslr.org/12`. If you are using Kaldi, ESPnet, or Flashlight, the recipes already include scripts that download the data and format it.

Relevant code bases:

- Kaldi: `https://kaldi-asr.org/`
- ESPNet: `https://github.com/espnet/espnet`
- Flashlight: `https://github.com/facebookresearch/flashlight/tree/master/flashlight/app/asr`

Proposing your own project

If you would like to define your own final project, you are more than welcome to do so. The two main constraints are that the project must have something to do with speech, audio, or language technology, and be of a size and scope appropriate for the project. This means that the project should not be trivial, but should be doable within a 1-month time frame. Additionally, you will need to make sure that you can procure sufficient computational resources and datasets for your project.

Writing a Survey Paper

If you prefer not to work on an implementation-based final project, you can instead opt to write a research survey paper. For this project, you should choose a particular problem in speech processing, search the web for relevant research papers on that problem, and write a

report that organizes and summarizes all of the work that you found on the problem. This report should be longer in length, between 6 and 10 pages.