COS 484: Natural Language Processing

Final projects

Fall 2019

Announcements

- Assignment 4 due: Nov 18
- Problem 3: Neural MT (Nov 12 lecture)

Logistics

- Initial proposal due: Nov 11
- Proposal feedback meetings: Nov 12 - Nov 15
- Milestone meetings: first/second week of Dec
- Final project presentations: around Jan 9 - 13
- Final report due: January 14 (Dean's date)
- Teams of 2 - 3 members

Decide the goal

- First, clearly define a **specific** goal/hypothesis of the project (more details the better!)
  - Good: “Adding a convolutional layer to the Transformer architecture will better capture positional information and improve machine translation”
  - Good: “Reimplement XYZ paper and verify the results on ABC tasks”
  - Bad: “Use BERT for analyzing financial news” (too vague/generic)
  - Bad: “Improve question answering for SQuaD” (how?)
- Important to pick an **achievable** goal (we can help!)

Considerations while picking a project

1. **Availability of data**
    - DO NOT TRY TO COLLECT YOUR OWN DATA
        - Significant time (and money) investment (at your own risk!)
2. ML framework - pick your favorite one
    - PyTorch, Tensorflow, Keras, etc.
3. Statistical model or neural network architecture
    - What are the inputs, outputs, functions, parameters?
4. ML Objective (Supervised? Unsupervised?)
5. **Availability of compute**

Finding inspiration

1. Do a thorough literature search - almost certain that someone has attempted something similar before you
    - Google scholar, ACL anthology
    - Don't stop with just one search. Try variants - e.g. "neural style transfer" vs "adapting models for different writing styles"
2. Search "awesome {NLP, RL, computer vision} papers github"
    - Example: https://github.com/mhagiwara/100-nlp-papers
    - Play around with code existing on github and see how readable/usable it is if you want to build off
3. Project page with potential ideas: nlp.cs.princeton.edu/cos484/projects.html

Other sources

- https://paperswithcode.com/sota

Browse > Natural Language Processing

Natural Language Processing

398 leaderboards • 217 tasks • 106 datasets • 3103 papers with code

Machine Translation

Task	Leaderboards	Papers with code
Machine Translation	43 leaderboards	503 papers with code
Transliteration		15 papers with code
Unsupervised Machine Translation	9 leaderboards	9 papers with code
Multimodal Machine Translation		7 papers with code
Low-Resource Neural Machine Translation		6 papers with code

See all 6 tasks

Question Answering

Task	Leaderboards	Papers with code
Question Answering	46 leaderboards	404 papers with code
Open-Domain Question Answering	3 leaderboards	21 papers with code
Answer Selection	2 leaderboards	16 papers with code
Community Question Answering		12 papers with code
Knowledge Base Question Answering	1 leaderboard	7 papers with code

Brainstorming

- Have each team member flesh out 10-20 quick ideas before first meeting
- Filter out list by performing Google searches
  - Data availability
  - Has the same idea been done before (with possibly existing github code)?
  - How long and how much do the models need to be trained?
- Are there little tweaks or other experiments that haven't been done yet in existing work?
- Can you extend idea in one paper with an idea from another?
- Which idea allows for more experimentation/interesting conclusions?

Types of projects

1. Experiment with improving an architecture on a predefined task
2. Case study: Apply an architecture to a dataset in the real world (that has not been done before)
3. Compete in a predefined competition (SemEval 2020, Kaggle, etc.)
4. Stress test or comparison study of known models/architecture (e.g. when are RNNs better than Transformers for task XYZ?)
5. Design a novel NN layer, objective function, optimizer, etc.
6. Multi-domain NLP (RL + NLP, CV + NLP, …)
7. Visualization/Interpretability study of deep learning models
8. …

Reading papers

- Don’t read start to finish in order
- Tables, figures, captions provide a lot of useful information at first glance
- **First pass: Abstract, Introduction, Experiments, Results**
- Plenty of blogs, github repos, etc. that summarize several papers at once in a nice manner

Compute

- Some projects may require more CPU/GPU resources
- Tiger clusters: https://researchcomputing.princeton.edu/systems-and-services/available-systems/tiger
- CS Ionic clusters: https://csguide.cs.princeton.edu/resources/clusters
- Google cloud / Amazon AWS credits / Google Colab (1 free GPU)
- **Request/get access to the above ASAP if you plan on using them!**

Tips for successful projects

1. Clearly divide work between team members for optimal progress
2. **Start early and work on it every 1-2 days rather than rush at the end**
3. Set up work flow ASAP - download data, verify data, set up base code
4. Have running code and fully trained baseline model by milestone
5. Have a clear, well-defined hypothesis to be tested (++ novel/creative hypothesis)
6. Conclusions and Results should teach the reader something
7. Meaningful tables, plots to display the key results

- ++ nice visualizations or interactive demos
- ++ novel/impressive engineering feat
- ++ good results

Come to office hours and talk to us!

Scenarios to avoid

- Data not available or hard to get access to, which stalls progress
- All experiments run with prepackaged source - no extra code written for model/data processing
- Team starts late - only draft of code up by milestone
- Just ran model once or twice on the data and reported results (not much hyperparameter search done)
- A few standard graphs: loss curves, accuracy, without any analysis
- Results/Conclusion don’t say much besides that it didn’t work
  - Even if results are negative, analyze them

Milestone goals

1. Have code up and running
2. Source of data explained correctly, along with true train/test/val split
3. What Github repo, or other code you're basing off of
4. Ran baseline model and have results
5. Brief discussion of initial, preliminary results
6. Reasonable literature review (2+ related papers)
7. 1-2 page progress report (not very formal)

Bonus points available for good milestone reports!