Course Project

COMP7607: Natural Language Processing - University of Hong Kong

Fall 2022

The final project will offer you the chance to engage with the current state of NLP research and to apply your newly acquired skills to an in-depth NLP application. In this project, you will be in a team of 3-5 students and reproduce an ACL/EMNLP/NAACL paper from the past two years. [1]

Teams: Your team will be composed of 3-5 students. Please submit your team's information via Moodle. If you can not have a team of 3 or more students for the project, you can post on Moodle forum to find project partners. If you are still unable to form a team, you can fill in this form. You can start preparing your project proposal as soon as you have assembled your team. (Find your teammates early!)

Topics: You will choose a paper from ACL{2022, 2021}, EMNLP{2021}, and NAACL{2022, 2021} (main and findings). You can browse all papers in proceedings on `https://aclanthology.org/`. You can find sample papers in Section 2.2 (not exhaustive!). Here are some tips for paper choosing:

- You should find the problem tackled in the paper interesting.
- You should be able to access the data you will need to reproduce the paper's experiments.
- In many cases, the authors may have made code available; this may be a blessing or a curse. You should definitely peruse a paper's codebase before deciding on that paper.
- Your project should not focus on new pretraining techniques for language models. Such experiments are too large-scale to feasibly execute even if you have access to significant other compute resources.
- Fine-tuning BERT-Base on a dataset can often be done effectively with more limited resources but will still typically require GPUs. If BERT-Base is still too big for your GPU resources, you can use distillBERT or a similar small pre-trained model.
- The tasks of machine translation and summarization usually rely on training on particularly large datasets. There are good projects you can do in these domains, but you may wish to focus on low-resource settings or more traditional models, as large-scale neural approaches won't be feasible to explore unless you have access to significant GPU resources.

1 Deliverables

1.1 Project proposal (5%)

You will need to submit a 1-2 page final project proposal. The proposal should outline the following:

- Names, school mail, and UID of all team members

[1] Many designs and materials from CSE517@UW, COS484@Princeton, CS388@UTAustin, and CS-4650@Gatech with special thanks!

- The title of the paper you choose
- A brief introduction (5-10 sentences) about the research problem this paper tackled
- One of the type of study you choose to perform:

  Baselines: reproduce the baselines in the paper and run them with different hyperparameters, ablations, etc. You are required to come up with at least one new baseline, which can be a modification of existing ones. You are not required to reproduce all results if there are many experimental results.

  Model ablation/analysis study: reproduce results of various model ablations from the paper. In addition, you will propose at least one new ablation/model variant and empirically test it. You may also propose different analyses on the model (e.g., test on new datasets, error analysis). You are not required to reproduce all results if there are many experimental results.

- Describe some considerations in choosing a paper to reproduce: [1] Check if the original codebase exists(not required). [2] Check that the dataset it uses is publicly available and of a reasonable size. [3] An estimate of the computational requirements for reproducing the paper. You should estimate how many hours per training run and the computational requirements.
- An (initial) timeline of this project and (initial) individual contribution.

1.2 Presentation (10%)

At the end of the semester, we will schedule project presentations for all the projects in the class. Each team will get a chance to present their work and get feedback from their instructors and peers. Your team should prepare a 5-min presentation with up to 5 slides and choose a member to present your work.

1.3 Final report (35%)

Each team will need to submit a final project report. Your report may consist of 4-8 pages of content, up to one page for individual contribution, and unlimited pages of references and appendixes. The final report will include a complete description of work undertaken for the project, including abstract, introduction, related works, methods, experimental details (complete enough for replication), comparison with past work, and thorough analysis. In addition, the final report should include an appendix section to describe the contributions of each member. You should complete your report using the ACL template. We recommend that you collaborate on writing reports using Overleaf and coding using GitHub.

2 Resources

2.1 Computational Resource

You may use the following platforms for GPU resources(Please prepare early!):

- Notebook[2]: Google Colab, Kaggle, Amazon SageMaker Studio Lab
- Internal GPU Resources: HKU GPU Farm
- Google Cloud/Amazon AWS free credits

[2]Refer to this blog for comparison.

2.2 Sample papers

These are examples of some papers from recent *CL conferences (e.g. ACL, NAACL, EMNLP). This is NOT an exhaustive or prescriptive list, they are just meant to give you an idea of how to pick papers. If you pick a paper from this list, please make sure to carefully consider your constraints (timeline, computational resources, etc.) and include them in your project proposal.

- ACL 2021 Prefix-Tuning: Optimizing Continuous Prompts for Generation
- ACL 2021 SimCSE: Simple Contrastive Learning of Sentence Embeddings
- EMNLP 2021 Improving and Simplifying Pattern Exploiting Training
- ACL 2022 BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models
- EMNLP 2021 Generating Datasets with Pretrained Language Models
- EMNLP 2021 ExpBERT: Representation Engineering with Natural Language
- NAACL 2021 On the Inductive Bias of Masked Language Modeling
- ACL 2021 Modeling Fine-Grained Entity Types with Box Embeddings
- NAACL 2021 Low-Complexity Probing via Finding Subnetworks
- ACL 2021 Implicit Representations of Meaning in Neural Language Models
- NAACL 2021 Reading and Acting while Blindfolded: The Need for Semantics in Text Game Agents
- ACL 2021 Self-Attention Networks Can Process Bounded Hierarchical Languages

2.3 Framework, Toolkit, Sample Code

- Framework: PyTorch, Tensorflow, JAX
- Toolkit: Huggingface {Transformers, Datasets}, PyTorch Lightning
- Code Examples: HuggingFace on Colab & Studio Lab, HuggingFace Code Examples

2.4 Sample Reports

These are examples of research proposals and reports. You can refer to them, but please do not repeat the paper in these samples.

- Proposal examples
- Report examples