Optimizing Data Usage in Neural Sequence-To-Sequence Learning

Tsz Kin Lam

1 Module Type

- Hauptseminar (8 credit points)

2 Prerequisite for Participation

Good knowledge of statistical machine learning (e.g., by successful completion of courses "Statistical Methods for Computational Linguistics" and/or "Neural Networks: Architectures and Applications for NLP") and experience in experimental work (e.g., software project or seminar implementation project) and basic knowledge of Sequence-To-Sequence Learning.

- Experience in deep learning libraries/frameworks such as PyTorch, OpenNMT, fairseq, and joeynmt
- (To be explicit) basic knowledge in Machine Translation e.g. beam search, byte-paired encoding, Transformer architecture, Masked Language Modelling and evaluation metrics such as BLEU

3 Assessment and Grading

- Regular and active participation: reading research papers and asking questions in class (10% -)
- Other participants should prepare two questions and send them to me and the presenter about 3 days before the presentation
- Oral presentation of (a) selected paper(s) + Comments on the paper (40%)
- Implementation project (50%)

4 Content

Deep learning is the de facto standard for many classification tasks, e.g., natural language processing or image recognition. However, it is also notorious of being

data hungry. This data hungry nature, together with the costly annotation process, has stimulated a lot of research on creating synthetic data, aka data augmentation. Another popular method is to create additional data by crawling data on web, aka data crawling. Both approaches allow substantial increases of training data at little cost. However, synthetic or crawled data can be noisy, e.g., due to misalignments between source and target sentences, or due to a domain mismatch between new data and original training data. This casts doubt on the benefits of such additional data to the final model performance, and is also the place where data selection comes into play.

The focus of this seminar is on optimizing data usage of neural sequence-to-sequence learning in text data. Participants will learn about the recent advances of data selection, data augmentation and their connections to multi-domain scenarios. The application focus will be sequence-to-sequence learning, especially machine translation.

5 Topics will include (but not limited to)

- Data selection/filtering
- Data augmentation and adversarial inputs
- Generalization over multiple domains

6 Oral Presentations

- Select from the **Literature** section below. There are two papers to be presented in some selections
- You may have to read more in order to understand the insights (I can also add some background papers)
- About 45 mins, including your comments
- Followed by Q&A section

6.1 Comments

The presenter should also give his/her opinions e.g.

- what are the limitation of the proposed method?
- how transferable is the proposed method?
- what are the missing experiments?
- is the idea novel enough?

6.2 Literature (subject to change)

Please click the title to access the paper

6.2.1 Data selection/filtering

- 1. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora
  2. Learning a Multi-Domain Curriculum for Neural Machine Translation
- 1. Optimizing Data Usage via Differentiable Rewards
  2. Meta Back-Translation
- Distributionally Robust Multilingual Machine Translation

6.2.2 Data augmentation and Adversarial inputs

- Generating Sentences by Editing Prototypes
- Good-Enough Compositional Data Augmentation
- Learning to Recombine and Resample Data For Compositional Generalization
- Doubly-Trained Adversarial Data Augmentation for Neural Machine Translation
- Counterfactual Data Augmentation for Neural Machine Translation
- An Investigation of the (In)effectiveness of Counterfactually Augmented Data
- Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach

6.2.3 Generalization over multiple domains

- 1. Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation
  2. Improving the Quality Trade-Off for Neural Machine Translation Multi-Domain Adaptation
- Multi-Domain Neural Machine Translation with Word-Level Adaptive Layer-wise Domain Mixing

7 Implementation Project

Deadline: 31st May 2022

- You have your own ideas e.g. investigating the comments you raised, or
- I will impose some questions for investigations
- Code + summary of 4 to 8 pages

In either case, we should finalize the project topic before 15th March 2022
Some libraries for references:

- joeyNMT (https://github.com/joeynmt/joeynmt)
- fairseq (https://github.com/pytorch/fairseq)

8 TODO

- Write me an e-mail by 11th Nov 2021 and send 3 preferred choices from the list if you would like to participate
- Next week: introduction and expectation for each paper

9 Important(?) aspects of the selected papers

Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora

- Cross-entropy difference filtering
- Dual conditional cross-entropy filtering
- Background about data filtering tasks in WMT

Learning a Multi-Domain Curriculum for Neural Machine Translation

- Multi-domain NMT e.g. domain tags
- Static data selection vs Dynamic data selection
- Data noise(?)
- Bayesian Optimization
- Instance-level features

Optimizing Data Usage via Differentiable Rewards; Meta Back-Translation

- Model-agnostic meta-learning
- Similar topics and applications but in ML conferences

- Multilingual NMT
- Diversity and quality of the synthetic data

Distributionally Robust Multilingual Machine Translation

- Comparison of different DROs algorithms esp. drawbacks of Empirical Risk Minimization (ERM)
- Primal-dual methods i.e. min-max problem
- Similarities to policy-gradient methods(?)
- Multilingual NMT
- ML-like paper in a NLP conference

Generating Sentences by Editing Prototypes

- Neural editor and sentence VAE
- Latent variables and ELBO
- Useful to understand "the Learning to Recombine" paper
- about language model

Good-Enough Compositional Data Augmentation

- Compositional Generalization
- Linguistically motivated data augmentation
- With application on semantic parsing and language modeling

Learning to Recombine and Resample Data for Compositional Generalization:

- Compositional Generalization
- Controlled generalization(?)
- R&R: n-protoype models, recombination networks
- The importance of re-sampling and neighbourhoods
- Its analysis

Doubly-Trained Adversarial Data Augmentation for Neural Machine Translation

- White-box attack
- Random deletion and nearest neighbour search as perturbations
- Minimum Risk Training

- How to measure model robustness
- Another evaluation metric for MT: COMET
- Connection between attack and data augmentation

Counterfactual Data Augmentation for Neural Machine Translation

- Structural causal model
- Masked Language Modeling and Translation Language Model
- Alignment

An Investigation of the (In)effectiveness of Counterfactually Augmented Data

- Some concepts in causal learning/inference e.g. spurious features
- Analysis via linear Gaussian model
- Application in entailment task
- On Arxiv, not accepted yet

Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach

- Integration of multi-task learning and heuristic-based data augmentation techniques
- Hallucinations in NMT
- Low-resource NMT
- Relatively simple

Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation

- Catastrophic forgetting
- Elastic Weight Consolidation
- Natural gradient and Fisher information matrix

Multi-Domain Neural Machine Translation with Word-Level Adaptive Layer-wise Domain Mixing

- Modification of the Transformer architecture for multi-domain setting i.e. domain tags in training set
- What to do if there is no domain tag in inference?
- Domain proportion: multi-tag?
- Domain information in token level
- Neither data augmentation nor data selection

10 Schedule

- 11 Nov: Pause
- 18 Nov: Overcoming Catatrophic Forgetting During ... + Improving the Quality Trade-Off for NMT Multi-Domain Adapatation
- 25 Nov: Dual Conditioanl XENT + Learning a Multi-Domain
- 2 Dec: Optimizing Data Usage via ... + Meta Back-Translation
- 9 Dec: Distributionally Robust Multilingual NMT
- 16 Dec: Generating Sentences by Editing Prototypes
- 13 Jan: Good-Enough Compositional Data Augmentation
- 20 Jan: Learning to Recombine and Resample ...
- 27 Jan: Doubly-Trained DA for NMT
- 3 Feb: Counterfactual DA for NMT
- 10 Feb: An Investigation of the (In)Effectiveness of ... (or Shifted upward)
- 17 Feb: Rethinking DA for Low-Resource NMT ... or Multi-DA NMT with Word-Level Adaptive Layer-wise Domain Mixing

11 TODO:

- Write me an e-mail by 11th Nov 2021 and send 3 preferred choices from the list if you would like to participate
- Reading group if no one presenter on that week
    1. Highlight important paragraphs on .pdf for discussions