Massively Multilingual Neural Machine Translation in the Wild - Findings and Challenges

2019 • Multi Domain • Multi Task • Natural Language Processing • Neural Machine Translation • AI • NLP • NMT • Scale

30 Jan 2020

Introduction

- The paper proposes to build a universal neural machine translation system that can translate between any pair of languages.
- As a concrete instance, the paper prototypes a system that handles 103 languages (25 Billion translation pairs).
- Link to the paper

Why universal Machine Translation

- Hypothesis: *The learning signal from one language should benefit the quality of other languages*[1]
- This positive transfer is evident for low resource languages but tends to hurt the performance for high resource languages.
- In practice, adding new languages reduces the effective per-task capacity of the model.

Desiderata for Multilingual Translation Model

- Maximize the number of languages within one model.
- Maximize the positive transfer to low resource languages.

- Minimize the negative interference to high resource languages.
- Perform well ion the realistic, multi-domain settings.

Datasets

- In-house corpus generated by crawling and extracting parallel sentences from the web.
- 102 languages, with 25 billion sentence pairs.
- Compared with the existing datasets, this dataset is much larger, spans more domains, has a good variation in the amount of data available for different language pairs, and is noisier. These factors bring additional challenges to the universal NMT setup.

Baselines

- Dedicated Bilingual models (variants of Transformers).
- Most bilingual experiments used Transformer big and a shared source-target sentence-piece model (SPE).
- For medium and low resource languages, the Transformer Base was also considered.
- Batch size of 1 M tokes per-batch. Increasing the batch size improves model quality and speeds up convergence.

Effect of Transfer and Interference

- The paper compares the following two setups with the baseline:
    - Combine all the datasets and train over them as if it is a single dataset.
    - Combine all the datasets but upsample low resource languages so all that all the languages are equally likely to appear in the combined dataset.

- A target “index” is prepended with every input sentence to indicate which language it should be translated into.

- Shared encoder and decoder are used across all the language pairs.

- The two setups use a batch size of 4M tokens.

Results

- When all the languages are equally sampled, the performance on the low resource languages increases, at the cost of performance on high resource languages.

- Training over all the data at once reverse this trend.

Countering Interference

- Temperature based sampling strategy is used to control the ratio of samples from different language pairs.

- A balanced sampling strategy improves the performance for the high resource languages (though not as good as the multilingual baselines) while retaining the high transfer performance on the low resource languages.

- Another reason behind the lagging performance (as compared to bilingual baselines) is the capacity of the multilingual models.

- Some open problems to consider:

    - Task Scheduling - How to decide the order in which different language pairs should be trained.

    - Optimization for multitask learning - How to design optimizer, loss functions, etc. that can exploit task similarity.

    - Understanding Transfer:

        - For the low resource languages, translating multiple languages to English leads to improved performance than translating English to multiple languages.

- This can be explained as follows: In the first case (many-to-one), the setup is that of a multi-domain model (each source language is a domain). In the second case (one-to-many), the setup is that of multitasking.

- NMT models seem to be more amenable to transfer across multiple domains than transfer across tasks (since the decoder distribution does not change much).

- In terms of zero-shot performance, the performance for most language pairs increases as the number of languages change from 10 to 102.

Effect of preprocessing and vocabulary

- Sentence Piece Model (SPM) is used.
- Temperature sampling is used to sample vocabulary from different languages.
- Using smaller vocabulary (and hence smaller sub-word tokens) perform better for low resource languages, probably due to improved generalization.
- Low and medium resource languages tend to perform better with higher temperatures.

Effect of Capacity

- Using deeper models improves performance (as compared to the wider models with the same number of parameters) on most language pairs.

Related Posts

Hints for Computer System Design 07 Jan 2022

Synthesized Policies for Transfer and Adaptation across Tasks and Environments 29 Mar 2021

Deep Neural Networks for YouTube Recommendations 22 Mar 2021