CSCI 662: Assignment#5

Due: November 9, 2017 11:59 PM PST

In this assignment, you will design and run Neural Machine Translation (NMT) experiments using Tensorflow.

Warm-up

1. Install Tensorflow 1.3: `https://www.tensorflow.org/install/`
2. Read the first two sections (Introduction and Basic) of the neural machine translation (seq2seq) tutorial on Github (`https://github.com/tensorflow/nmt/tree/tf-1.2`). These explain the basics of building sequence-to-sequence models using Tensorflow.
3. Download the file `hw5-data.tar.gz` from Piazza. This file contains the data that you will be using to train your models.

Installing the Tutorial

Once TensorFlow is installed, you can download the source code of the seq2seq tutorial by running:

```
git clone https://github.com/tensorflow/nmt.git --branch tf-1.2
--single-branch
```

1 Data (10 points)

You will be using a subset of the IWSLT15 Vietnamese-English dataset for your experiments. Data files can be found under (`hw5-data/`). The data is pre-processed and tokenized for you. Remember that we need a parallel corpus to train an MT system. Notice the language suffix in file names (`.vi` for Vietnamese and `.en` for English). All data files are sentence-aligned (i.e. the first sentence in `train.en` is the English translation of the first sentence in `train.vi`, and so on).

Q1. (3 points) How many word tokens are there in `train.en`?

Q2. (3 points) How many word types are there in `train.en`?

Q3. (2 points) How many sentences are there in `train.en`?

Q4. (2 points) What is the average sentence length (in words) in `train.en`?

2 Building an NMT Model (20 points)

In this part of the assignment, you will train a neural machine translation model to translate from Vietnamese (source) to English (target). i.e. Vietnamese is the input language and English is the output language. Follow the instructions in the tutorial to train your NMT model. Use the given hyper-parameter settings in `hw5-data/hparams.json` and the given vocab, train, dev, and test data files.

Q5. (8 points) Show the English translations of the first 5 dev sentences in this format:

<source Vietnamese>

<human reference English translation>

<your system's English translation>

<blank line>

Q6. (3 points) Look at the log file in your output directory. What is the best dev set perplexity your model could get?

Q7. (3 points) What is the BLEU score on the dev set?

Q8. (3 points) What is the test set perplexity?

Q9. (3 points) What is the BLEU score on the test set?

3 Hyper-parameters (35 points)

In this part of the assignment, you will study the effect of hyper-parameter tuning on NMT performance. Pick a hyper-parameters of your choice for analysis. Try at least 5 different values for this parameter and see how this affects MT performance. You should report both test set perplexity and BLEU.

Q10. (10 points) Create a results table that includes hyper-parameter values and their effect on test set perplexity and BLEU.

Q11. (10 points) Plot hyper-parameter values (x-axis) vs. test set perplexity (y-axis).

Q12. (10 points) Plot hyper-parameter values (x-axis) vs. test set BLEU (y-axis).

Q13. (5 points) Did your results match your expectations? Briefly explain your findings.

4 Learning Curve (35 points)

In this part of the assignment, you will study the effect of training data size on NMT performance. Create 5 smaller training data sets by repeatedly cutting the given training dataset in half. Re-train your NMT system using the new training data sets. Use the originally given dev and test sets for evaluation.

Q14. (10 points) Create a results table that includes different training data sizes and their effect on test set perplexity and BLEU.

Q15. (10 points) Plot training data size (x-axis, log scale) vs. test set perplexity (y-axis).

Q16. (10 points) Plot training data size (x-axis, log scale) vs. test set BLEU (y-axis).

Q17. (5 points) How does training data size affect NMT performance? Briefly explain your results.

Summary of Files Provided

- `train.vi` and `train.en`: Training data, formatted as one sentence per line, with line-by-line correspondence between Vietnamese and English translations.
- `dev.vi` and `dev.en`: Development data, formatted as one sentence per line, with line-by-line correspondence between Vietnamese and English translations.
- `test.vi` and `test.en`: Test data, formatted as one sentence per line, with line-by-line correspondence between Vietnamese and English translations.
- `vocab.vi` and `vocab.en`: Vocabulary files for Vietnamese and English.
- `hparams.json`: Initial hyper-parameter settings for your NMT experiments.

What to Turn in

Make a tar file called `FirstName_LastName_Assignment5.tar`. It should unpack into a directory called `FirstName_LastName_Assignment5`. That directory should include:

- **writeup.pdf**: Your answers to questions. Make sure to put all your answers in a **single pdf file.** Any graphs/images outside that single pdf file will **NOT** be graded.
- **Your MT output files:** Submit your system output translations for all your experiments (for the test set only). No need to submit trained models. Make sure to use meaningful names for your files.