update

May 22, 2023

1 CS 449 Final Project Update

Due: May 23, 2023 at 11:59pm

1.1 1. Names and Net IDs

- Kyle Hwang (ksh6947)
- Michael Lin (qlb968)
- Dylan Wu (dwg0364)

1.2 2. Abstract

Write a new abstract that describes the motivation for your project and proposed methods. This should reflect any changes you've made, but you'll explain those changes later.

Our final project seeks to use `pytorch` to replicate the "Attention is All You Need", which introduced the Transformer as a way to improve upon existing sequence transduction language models. We will attempt to implement the model's architecture and train the model on a subset of the WMT 2014 English-French dataset. We will then perform analyses on the training results, and, where appropriate, compare the model performance against results from the paper.

1.3 3. Big Changes

Describe any major changes that have taken your project in a new direction. What made you decide to make these changes? Were some things easier or harder than you expected?

So far, we don't have any big changes. To focus our implementation to value-added parts of the model, we will use an off-the-shelf tokenizer and BLEU score implementations.

1.4 4a. Describe your dataset(s)

This should be an updated version of your datasets section from your proposal. Write this section as if the reader hasn't read your original proposal. List the datasets you're using, where you found them, and what they contain. Be detailed! For each dataset, what does the data look like? What is the data representation? (e.g., what resolution of images? what length of sequences?) How is the data annotated or labeled? Include citations for the datasets. Include at least one citation of previous work that has used this data, or an explanation of why it hasn't been used before

We will be using the WMT (Workshop on Statistical Machine Translation) 2014 English-German and English-French datasets. These datasets provide examples of phrases in

1

English and its translation in German or French (and vice versa). Specific details about the datasets can be found here: https://aclanthology.org/W14-3302.pdf . In summary, the datasets come from formal sources (European Parliment, United Nations, news sources, etc.). They directly translated text from the source language to the target language (i.e. with no intermediary language), using machine translation, and this translation subsequently was followed by an involved manual evaluation process. The dataset does not involve any tokenization; it is purely the source text and its translation. Each example is relatively short: as stated before, they are either a simple phrase or a sentence. The reason for using this dataset is to replicate the work of Vaswani et al. in their influential paper "Attention is All you Need."

Ondrej Bojar et al. "Findings of the 2014 Workshop on Statistical Machine Translation". In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 12–58. url: http://www.aclweb.org/anthology/W/W14/W14-3302.

Ashish Vaswani et al. "Attention is All you Need". In: Advances in Neural Information Processing Systems. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. url: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

1.5 4b. Load your dataset(s)

Load at least a few examples and visualize them as best you can

```python
import json
```

```python
de, fr = {}, {}
with open("./data/de-en/test.json") as f:
    de["test"] = json.load(f)
with open("./data/de-en/train.json") as f:
    de["train"] = json.load(f)
with open("./data/de-en/validation.json") as f:
    de["validation"] = json.load(f)
with open("./data/fr-en/test.json") as f:
    fr["test"] = json.load(f)
with open("./data/fr-en/train.json") as f:
    fr["train"] = json.load(f)
with open("./data/fr-en/validation.json") as f:
    fr["validation"] = json.load(f)
```

```python
for key in de:
    for d in de[key][:2]:
        print(f"de {key} \t", d)
    print()
for key in fr:
    for d in fr[key][:2]:
        print(f"fr {key} \t", d)
    print()
```

de test   {'translation': {'de': 'Gutach: Noch mehr Sicherheit für
Fußgänger', 'en': 'Gutach: Increased safety for pedestrians'}}
de test   {'translation': {'de': 'Sie stehen keine 100 Meter voneinander
entfernt: Am Dienstag ist in Gutach die neue B 33-Fußgängerampel am
Dorfparkplatz in Betrieb genommen worden - in Sichtweite der älteren
Rathausampel.', 'en': 'They are not even 100 metres apart: On Tuesday, the new B
33 pedestrian lights in Dorfparkplatz in Gutach became operational - within view
of the existing Town Hall traffic lights.'}}

de train   {'translation': {'de': 'Wiederaufnahme der Sitzungsperiode',
'en': 'Resumption of the session'}}
de train   {'translation': {'de': 'Ich erkläre die am Freitag, dem 17.
Dezember unterbrochene Sitzungsperiode des Europäischen Parlaments für
wiederaufgenommen, wünsche Ihnen nochmals alles Gute zum Jahreswechsel und
hoffe, daß Sie schöne Ferien hatten.', 'en': 'I declare resumed the session of
the European Parliament adjourned on Friday 17 December 1999, and I would like
once again to wish you a happy new year in the hope that you enjoyed a pleasant
festive period.'}}

de validation   {'translation': {'de': 'Eine republikanische Strategie, um der
Wiederwahl von Obama entgegenzutreten', 'en': 'A Republican strategy to counter
the re-election of Obama'}}
de validation   {'translation': {'de': 'Die Führungskräfte der Republikaner
rechtfertigen ihre Politik mit der Notwendigkeit, den Wahlbetrug zu bekämpfen.',
'en': 'Republican leaders justified their policy by the need to combat electoral
fraud.'}}

fr test   {'translation': {'fr': 'Spectaculaire saut en "wingsuit" au-
dessus de Bogota', 'en': 'Spectacular Wingsuit Jump Over Bogota'}}
fr test   {'translation': {'fr': "Le sportif Jhonathan Florez a sauté
jeudi d'un hélicoptère au-dessus de Bogota, la capitale colombienne.", 'en':
'Sportsman Jhonathan Florez jumped from a helicopter above Bogota, the capital
of Colombia, on Thursday.'}}

fr train   {'translation': {'fr': 'Reprise de la session', 'en':
'Resumption of the session'}}
fr train   {'translation': {'fr': 'Je déclare reprise la session du
Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et
je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes
vacances.', 'en': 'I declare resumed the session of the European Parliament
adjourned on Friday 17 December 1999, and I would like once again to wish you a
happy new year in the hope that you enjoyed a pleasant festive period.'}}

fr validation   {'translation': {'fr': "Une stratégie républicaine pour contrer
la réélection d'Obama", 'en': 'A Republican strategy to counter the re-election
of Obama'}}
fr validation   {'translation': {'fr': 'Les dirigeants républicains
justifièrent leur politique par la nécessité de lutter contre la fraude

électorale.', 'en': 'Republican leaders justified their policy by the need to combat electoral fraud.'}}

1.6 4c. Small dataset

Describe the small dataset you constructed as requested in the project proposal. How many examples did you use? How large was the dataset in total?

```
[ ]: for key in de:
         print(f'len de["{key}"]: \t', len(de[key]))
     for key in fr:
         print(f'len fr["{key}"]: \t', len(fr[key]))
```

```
len de["test"]:          3000
len de["train"]:         20000
len de["validation"]:    3000
len fr["test"]:          3000
len fr["train"]:         20000
len fr["validation"]:    3000
```

According to https://www.tensorflow.org/datasets/catalog/wmt14_translate, the de-en dataset was 1.36GB and the fr-en dataset was 14.65GB. In the original de-en dataset, there were 4,508,785 examples in the training set and in the original fr-en dataset, 40,836,876 examples. From the original de-en and fr-en WMT'14 dataset, we took a sample of 3000 examples for the test split and 3000 for the validation split.

In the original proposal, we were planning on using the streamed version of the dataset, but now, we are going to be using a locally downloaded version of the dataset. A locally downloaded version of the dataset would be faster and more reliable than a streamed version, so that is why we switched to this. As each individual example case is small, there was no need to perform any major preprocessing of the textual data beforehand; we simply had to limit the number of examples we put in each.

1.7 5. Updated Methods

This should be an updated version of your methods section from your proposal. Write this section as if the reader hasn't read your original proposal. For example:

This is a self-supervised learning task, and we will (as much as we can) implement the Transformer architecture in the paper from scratch using `pytorch`. We will then train the model using KL divergence loss on the WMT 2014 English-French dataset to predict English text from French text. This is the loss function used in the paper, which makes sense because at each token position, there is a predicted probability of the token to place there and an actual token that should be placed there. Given KL divergence loss measures the dissimilarity between the prediction distribution and the true labels, this is a good choice of loss function. We will evaluate the model's performance using BLEU score, which is used by the paper's authors.

1.8 6. Proposed Deliverables

Copy the deliverables from your proposal and write a sentence saying whether you completed this goal, made progress on this goal, or abandoned this goal. Unless you completed the goal, give an explanation of how it went.

1.8.1 6.1 Previous Essential Goals

- Have a complete neural network that runs without error. Additionally, to make sure we understand the impact of tuning hyperparameters, we will analyze the effects of varying hyperparameters on the trained model's BLEU score (keeping the number of epochs fixed).

  – We are almost done with the complete transformer, starting with training and hyperparameter tuning soon

- Generally, as we train the model for longer, the BLEU score should be increasing (we will include a chart to show this).

  – Still waiting to complete the model implementation.

1.8.2 6.2 Previous Desired Goals

- Achieve results of translation that is not simply mappings between the two language vocabularies but rather encompasses the context and attention mappings of the whole sentence. This could be achieved by inputting examples where gendered noun and adjective would be correctly translated (i.e., the Transformer can differentiate between genders).

  – "I am a tall man" vs. "I am a tall woman"
  – Still waiting to complete the model implementation.

- Have consistency across languages when training under the same model in terms of BLEU score. Testing with another language dataset (most likely English to French) and have a similar BLEU score performance under the same training settings and time.

  – Still waiting to complete the model implementation.

1.8.3 6.3 Previous Stretch Goals

- Based on the findings from the paper "Reformer: The Efficient Transformer", we would try to implement and quantify the impact of the suggested changes from that paper compared to the Transformer.

  – Still waiting to complete the model implementation.

- Try pretraining the model on English-to-French, then fine tune the model to translate English-to-Spanish (the motivation is that since both French and Spanish are Romance languages, the pretrained model could have already learned important parts of the mapping from English to any Romance language).

  – Still waiting to complete the model implementation.

1.9   7. Results So Far

Show the results of your methods trained on your data. You should at least have empirical results for the small dataset described in 4c. Unless you are doing generation or another task where simple models (e.g., linear regression or logistic regression) aren't applicable, we strongly encourage you to compare your initial results against a simple baseline. If you have additional results, you're welcome to include those as well.

Currently still implementing the model and do not have results yet.

1.10   8. New Deliverables

For any proposed deliverables that you did NOT complete or abandon, copy them into these lists below. Then, if you have abandoned or completed your Desired or Stretch goals, add new goals until you have at least two goals per category.

1.10.1   8.1 New Essential Goals

- Generally, as we train the model for longer, the BLEU score should be increasing (we will include a chart to show this).

1.10.2   8.2 New Desired Goals

- Achieve results of translation that is not simply mappings between the two language vocabularies but rather encompasses the context and attention mappings of the whole sentence. This could be achieved by inputting examples where gendered noun and adjective would be correctly translated (i.e., the Transformer can differentiate between genders).

  - "I am a tall man" vs. "I am a tall woman"

- Have consistency across languages when training under the same model in terms of BLEU score. Testing with another language dataset (most likely English to French) and have a similar BLEU score performance under the same training settings and time.

1.10.3   8.3 New Stretch Goals

- Based on the findings from the paper "Reformer: The Efficient Transformer", we would try to implement and quantify the impact of the suggested changes from that paper compared to the Transformer.

- Try pretraining the model on English-to-French, then fine tune the model to translate English-to-Spanish (the motivation is that since both French and Spanish are Romance languages, the pretrained model could have already learned important parts of the mapping from English to any Romance language).

1.11   9. Hopes and Concerns

What's gone well so far? What's been the hardest thing?

The creation of the model has been going quite well, most of the modules are coming together and working together nicely.

We are most excited in being able to recreate the groundbreaking paper for the transformer models and emulate the responses for the training on translations. It would be very exciting to see the model actually translating new input sentences into a new language and interpreting how accurate it is.

We are slightly nervous about the results in BLEU score that our transformer model is going to output since we are using a significantly smaller dataset, less training time as well as weaker computing power than that of the paper. We are also slighlty concerned that our initial architechture would not run with our personal computers due to processing power limitations.

1.12   10. References

Cite the papers or sources that you used to discover your datasets and/or models, if you didn't include the citation above. For example:

Vaswani, Ashish, et al. "Attention Is All You Need." ArXiv (2017): /abs/1706.03762.

Kitaev, Nikita, et al. "Reformer: The Efficient Transformer." ArXiv (2020): /abs/2001.04451.