proposal

April 20, 2023

1 CS 449 Final Project Proposal

Due: April 21, 2023 at 11:59pm

1.1 1. Names and Net IDs

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

1.2 2. Abstract

Your abstract should be two or three sentences describing the motivation for your project and your proposed methods.

> 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 compare our results against the performance achieved by the paper.

1.3 3. Introduction

Why is this project interesting to you? Describe the motivation for pursuing this project. Give a specific description of your data and what machine learning task you will focus on.

> We are interested in understanding the implementation details of the Transformer, which underpins many large language models that perform vital tasks such as machine translation and prompt response generation. In addition, we think being able to implement AI research is an important skill to develop.
>
> We will use the same data that the paper used, namely the WMT 2014 English-French dataset. Each example in the dataset consists of a string of text in English and the corresponding string of text in French. The machine learning tasks we will focus on is machine translation.

1.4 4a. Describe your dataset(s)

List the datasets you plan to use, 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 your data, or explain why no one has used your data 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 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)

Demonstrate that you have made at least some progress with getting your dataset ready to use. Load at least a few examples and visualize them as best you can

[ ]:
```
from datasets import load_dataset

'''
    streaming=True is important. It will otherwise download the whole dataset. It will probably take an hour to load both of these in this way.

    Quick Guide:
        IterableDatasetDict : https://huggingface.co/docs/datasets/v2.11.0/en/package_reference/main_classes#datasets.IterableDatasetDict
        IterableDataset     : https://huggingface.co/docs/datasets/v2.11.0/en/package_reference/main_classes#datasets.IterableDataset

        `load_dataset(..., streaming=True)` returns an `IterableDatasetDict`.
        Use 'train', 'test', or 'validation' as keys to access the respective data of type `IterableDataset`.
        On `IterableDataset`,
            use `take(n)` for some n:int > 0 to get `IterableDataset` with the first n examples.
```

```
            use `shuffle()` to shuffle the dataset
'''
dataset_fr = load_dataset("wmt14", "fr-en", streaming=True)
dataset_de = load_dataset("wmt14", "de-en", streaming=True)
```

[ ]:
```
for data in dataset_fr['train'].take(2):
    print(data)
print()
for data in dataset_fr['test'].take(2):
    print(data)
print()
for data in dataset_fr['validation'].take(2):
    print(data)
print()
for data in dataset_de['train'].take(2):
    print(data)
print()
for data in dataset_de['test'].take(2):
    print(data)
print()
for data in dataset_de['validation'].take(2):
    print(data)
```

```
{'translation': {'fr': 'Reprise de la session', 'en': 'Resumption of the
session'}}
{'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.'}}

{'translation': {'fr': 'Spectaculaire saut en "wingsuit" au-dessus de Bogota',
'en': 'Spectacular Wingsuit Jump Over Bogota'}}
{'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.'}}

{'translation': {'fr': "Une stratégie républicaine pour contrer la réélection
d'Obama", 'en': 'A Republican strategy to counter the re-election of Obama'}}
{'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.'}}

{'translation': {'de': 'Wiederaufnahme der Sitzungsperiode', 'en': 'Resumption
of the session'}}
{'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.'}}

{'translation': {'de': 'Gutach: Noch mehr Sicherheit für Fußgänger', 'en':
'Gutach: Increased safety for pedestrians'}}
{'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.'}}

{'translation': {'de': 'Eine republikanische Strategie, um der Wiederwahl von
Obama entgegenzutreten', 'en': 'A Republican strategy to counter the re-election
of Obama'}}
{'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.'}}
```

[ ]:
```
'''
   NOTE: When viewing some examples, it may appear to have some potential␣
 ↪issues as seen below in "voiture\xa0?".
   However, this is simply a matter of how it is printed.
   As seen in the output of this code block, this will work properly when␣
 ↪directly observed.

   But also note in this example that there are going to be some␣
 ↪inconsistencies in the data.
   The \xa0 is present in the French translation but not the English one, and␣
 ↪this results in a space between the word and the question mark.
'''
printing_issue_example = list(dataset_fr['test'].skip(3).take(1))
print(printing_issue_example)
print(printing_issue_example[0]['translation']['fr'])
```

```
[{'translation': {'fr': 'Une boîte noire dans votre voiture\xa0?', 'en': 'A
black box in your car?'}}]
Une boîte noire dans votre voiture ?
```

1.6 4c. Small dataset

Many deep learning datasets are very large, which is helpful for training powerful models but makes debugging difficult. For your update, you will need to construct a small version of your dataset that contains 200-1000 examples and is less than 10MB. If you are working with images, video, or audio, you may need to downsample your data. If you are working with text, you may need to

truncate or otherwise preprocess your data.

Give a specific plan for how you will create a small version of one dataset you'll use that is less than 10MB in size. Mention the current size of your dataset and how many examples it has and how those numbers inform your plan.

By specifying `streaming=True` when initializing the dataset, it returns a version of the dataset that allows for immediate usage of the dataset without having to download the entire thing. This makes it so that it does not require any significant space as the data is streamed. Thus, making our dataset less than 10MB is trivial. See the below code block to see how much the streamed dataset takes up. Constructing a smaller version of our dataset is also trivial: simply specify the number of examples desired in the `take` function. This streamed dataset does not allow to see how many examples there are, but given the capabilities of this `IterableDataset`, this does not seem like this will cause an issue.

[ ]:
```
from sys import getsizeof # returns size in bytes

print("size of dataset_fr\t\t\t", "type: IterableDatasetDict\t",␣
 ↪getsizeof(dataset_fr), "Bytes")
print("size of dataset_fr['train']\t\t", "type: IterableDataset:\t\t",␣
 ↪getsizeof(dataset_fr['train']), "Bytes")
print("size of dataset_fr['train'].take(1000)\t", "type: IterableDataset\t\t",␣
 ↪getsizeof(dataset_fr['train'].take(1000)), "Bytes")
```

```
size of dataset_fr                      type: IterableDatasetDict       208
Bytes
size of dataset_fr['train']             type: IterableDataset:          56
Bytes
size of dataset_fr['train'].take(1000)  type: IterableDataset           56
Bytes
```

1.7 5. Methods

Describe what methods you plan to use. This is a deep learning class, so you should use deep learning methods. Cite at least one or two relevant papers. What model architectures or pretrained models will you use? What loss function(s) will you use and why? How will you evaluate or visualize your model's performance?

This is a 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 Cross-Entropy 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 Cross-Entropy 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. Deliverables

Include at least six goals that you would like to focus on over the course of the quarter. These should be nontrivial, but you should have at least one and hopefully both of your "Essential" goals done by the project update, due in mid-May. Your "Stretch" goals should be ambitious enough such that completing one is doable, but completing both this quarter is unlikely.

1.8.1 6.1 Essential Goals

- Have a complete neural network that runs without error
- After training with the smaller dataset for English translation, have a baseline BLEU score of at least 10

1.8.2 6.2 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 indicitive through a BLEU score of 20 (Transformers reached 28.4 but that is with a bigger training set and much more processing power).
- 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.

1.8.3 6.3 Stretch Goals

- Replicate the resulting data from the original paper up to an error margin of 10% since we are using a smaller dataset.
- Achieving simillar results qualitatively when comparing to the outputs of state of the art translations such as Google translate.

1.9 7. Hopes and Concerns

What are you most excited about with this project? What parts, if any, are you nervous about? For example:

> 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.

1.10 8. References

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

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