Анализ статьи "Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis"

Сергей Егоров

1. Окружение для эксперимента

Авторы любезно оставили ссылку на репозиторий статьи. Для воспроизведения эксперимента потребуется следующее окружение:

Библиотеки:

1. Смотрим необходимые модули в 'requirements.txt':

```
accelerate==0.19.0
datasets==2.7.1
evaluate==0.3.0
faiss_gpu==1.7.2
nltk==3.8
numpy==1.23.4
openai==0.27.1
rank_bm25==0.2.2
requests==2.28.1
scikit_learn==1.2.1
sentence_transformers==2.2.2
torch>=1.13.1
tqdm==4.64.1
transformers
```

2. Также нужно поставить 'OpenCL':

```
pip install openicl
```

И 'sacrebleu':

```
git clone --single-branch --branch adding_spm_tokenized_bleu  https://github.com/ngoyal2707/sacrebleu.git
cd sacrebleu
python setup.py install
```

Данные:

Авторы использовали FLORES-101. Датасет содержит 3001 предложение на разные темы из английской Википедии, а также перевод этих предложений на 101 язык. Итого авторы используют 102 языка и 606 пар перевода.

Так же для валидации BLOOMZ авторы собрали свой датасет News2023 из 1000 предложений из новостей на английском, и их перевод на китайский. Этого датасета в репозитории нет;(

Техника:

- Здесь довольно мало конкретики, потому что авторы не приводят точных параметров экспериментов. С другой стороны, моно попробовать посчитать, какие ресурсы им понадобились. Начнем с **ChatGPT** и **GPT-4** - авторы упоминают, что для обучения использовали только первые 100 предожений для каждой пары языков. В среднем, на предложение уходит 20 токенов стоимостью 0.03$ для ввода и для 0.06$ вывода, получаем:

$$\text{Total tokens for 606 pairs} = 100 \times 20 \times 606 = 1,212,000 \text{ tokens}$$

$$\text{Total Cost} = 1,212,000 \times \frac{\$0.03}{1,000} \times 3 = \$109.08 \text{ на 1 итерацию для каждой модели.}$$

- Для остальных моделей (места на детальные рассчеты не хватит, так что только результаты) время на 1 эпоху на 1 GPU NVIDIA A100:

  - **OPT-175B** ≈ 13 часов
    LLaMA2-7B ≈ 1.5 часа для FP16 с FlashAttention2 и Unsloth
    Falcon-7B ≈ 1.5 часа для FP16
    XGLM-7B ≈ 1.5 часа
    BLOOMZ-7.5B ≈ 1.5 часа
    M2M-12B ≈ 1.5 часа
    NLLB-1.3B ≈ 1 час

  Это все в предположении что у моделей близкое количество параметров и скорость обработки токенов примерно одинаковая.

Что если не соблюсти технические требования?

```
datasets, numpy, scikit_learn, torch, tqdm, transformers #ML база, без этого никуда.

nltk, openai, sentence_transformers, rank_bm25 #NLP специфика.

accelerate, evaluate, faiss_gpu, requests #настройка экспериментов. Здесь лучше ничего не забыть,

#иначе можно сильно отклониться от курсса экспериментов в статье
```

Очень демократичный набор модулей плюс нет подробного описания настройки экспериментов кроме генерации экземпляров предложений (Фигура 6 в статье).

2. Воспроизведение результатов

Наиболее вероятно удастся воспроизвести базовые результаты перевода для широко распространенных языков, таких как английский, французский, немецкий, испанский. Это связано с тем, что для этих языков существует много данных и модели обычно лучше обучены на этих парах, что выделяют и сами авторы.

Риски:

- Так как неизвестно, на чем обучались сегодняшние **GPT** модели, то можно пересечься с тестовой выборкой как в примере с **BLOOMZ** (Фигура 4 в статье).
- В репозитории подробно прописан код генерации тестовых экземпляров и подсчета метрик, только вот отсутствие данных настройки обучения на downstream задаче оставляет ресерчера в полной неопределенности сколько и как дообучать модели.
- Если модель была тонко настроена, то повторение процесса тонкой настройки может быть сложным из-за неопределенности гиперпараметров.
- В статье среди LLM **GPT-4** выглядит настоящим монстром, только авторы даже упоминают дороговизну его использования и снова не приводят конкретных экспериментов. Для неолатинских языков (romance) GPT-4 работает вообще лучше чем Google Translate. Может, просто стоит всё вложить в дообучение GPT-4 и получить новую SOTA'у для мультиязычного перевода?)

3. Потенциальные применения результатов

Сервисы и цели:

- Сервисы автоматического перевода, такие как Google Translate, могут использовать результаты для улучшения качества перевода на редкие языки.
- Внедрение в системы мультиязычных виртуальных помощников для улучшения обработки и перевода естественного языка. Авторы отмечают, что модели могут быть более гибкими и менее зависимыми от точных инструкций, чем считалось (Таблица 4 в статье).

Польза:

- Увеличение доступности контента на разных языках, особенно для языков с малым количеством ресурсов.
- Повышение качества перевода в специализированных областях, таких как медицина или право, где важно точное понимание терминов.
- Для индоевропейских языков - более "очеловеченный"перевод за счет сценариев обучения на разных не связанных с прямым переводом шаблонах.

4. Дополнительные рассуждения

Направления для улучшения:

- Исследование методов адаптации модели под конкретные языковые пары для улучшения качества перевода. Авторы деллают акцент на то, что перевод лучше выполняется на английский и для индоевропейских языков. То есть по ощущунию на результаты перевода крайне силльно влияет бэкшраунд обучения модели - на примерес английским, как самым распространенным языком в данных.
- Оптимизация вычислительных ресурсов для снижения затрат на обучение и инференс моделей.