

Báo cáo thực hành Bài Tập 1:

Phát Triển Mô Hình LLAMA2

Nhóm: 21

Thành viên: Nguyễn Huy Hoàng(22022584)

Nguyễn Tông Quân (22022635)

Vũ Trung Hiếu(22022515)

Nguyễn Xuân Trình(22022558)

Mục tiêu:

Bài tập này yêu cầu thực hiện một số thành phần quan trọng của mô hình ngôn ngữ Llama2 để hiểu rõ hơn về kiến trúc của nó. Cụ thể, chúng tôi sẽ thực hiện phân loại câu trên tập dữ liệu SST và CFIMDB bằng mô hình này.

Thực hiện:

Chúng tôi đã thực hiện các thành phần cốt lõi của Llama2 theo hướng dẫn trong file llama.py, classifier.py và optimizer.py. Chúng tôi đã tải trọng số được huấn luyện trước từ stories42M.pt, một mô hình ngôn ngữ 8 lớp với 42 triệu tham số được huấn luyện trên tập dữ liệu TinyStories.

Kết quả:

Mô hình đã được kiểm tra trong 3 trường hợp sau:

1. Hoàn thành văn bản:

Mô hình đã tạo ra các đoạn văn bản hoàn chỉnh với câu đầu vào "I have wanted to see this thriller for a while, and it didn't disappoint. Keanu Reeves, playing the hero John Wick, is". Kết quả cho thấy mô hình có khả năng tạo ra văn bản mạch lạc và đúng ngữ pháp, mặc dù nội dung có thể không hợp lý do được huấn luyện trên tập dữ liệu truyện thiếu nhi. Ví dụ với temperature 0.0 cho ra câu chuyện về John Wick chơi xe đồ chơi, còn temperature 1.0 cho ra câu chuyện John Wick làm một đứa trẻ.

```
load model from stories42M.pt
Temperature is 0.0
I have wanted to see this thriller for a while, and it didn't disappoint. Keanu Reeves, playing the hero John Wick, is this day. He was playing with his toy car, driving it around
the living room. Suddenly, he heard a loud crash. He had broken the car and was very sad.
John was angry and he shouted at his little brother. He was only three years old and he was only three. He was only three years old. He was very ups
-----
Wrote generated sentence to generated-sentence-temp-0.txt.
load model from stories42M.pt
Temperature is 1.0
I have wanted to see this thriller for a while, and it didn't disappoint. Keanu Reeves, playing the hero John Wick, is it!" As John toddled up to the sweet aroma, he held his mum
's hand tight, hoping he would finally reach the top step.
But as he stepped, he felt so heavy, like an apple. He had made a mistake! His mum apologisedDen handlely leaving, so he started to cry.
His m
-----
Wrote generated sentence to generated-sentence-temp-1.txt.
```

2. Phân loại tình cảm zero-shot:

Mô hình được sử dụng để phân loại tình cảm trên tập dữ liệu SST-5 và CFIMDB mà không cần huấn luyện thêm. Kết quả thu được không tốt, tương đương với việc chọn ngẫu nhiên một lớp. Cụ thể, độ chính xác trên tập SST-5 là **0.213** (dev) và **0.224** (test), trên tập CFIMDB là **0.502** (dev) và **0.213**(test).

Zero Shot Prompting Zero-Shot Prompting for SST:

```
python run_llama.py --option prompt --batch_size 30 --train data/sst-train.txt --dev
data/sst-dev.txt --test data/sst-test.txt --label-names data/sst-label-mapping.json --dev_out
sst-dev-prompting-output.txt --test_out sst-test-prompting-output.txt --use_gpu
```

```
load 1101 data from data/sst-dev.txt
load 2210 data from data/sst-test.txt
eval: 100% | 111/111 [00:04<00:00, 25.17it/s]
eval: 100% | 221/221 [00:08<00:00, 26.62it/s]
dev acc :: 0.213
test acc :: 0.224
```

Zero-Shot Prompting for CFIMDB:

```
python run_llama.py --option prompt --batch_size 10 --train data/cfimdb-train.txt --dev
data/cfimdb-dev.txt --test data/cfimdb-test.txt --label-names data/cfimdb-label-
mapping.json --dev_out cfimdb-dev-prompting-output.txt --test_out cfimdb-test-
prompting-output.txt
```

```
load 245 data from data/cfimdb-dev.txt
load 488 data from data/cfimdb-test.txt
eval: 100%|██████████████████████████████████████████████████████████| 25/25 [03:30<00:00, 8.40s/it]
eval: 100%|██████████████████████████████████████████████████████████| 49/49 [06:37<00:00, 8.12s/it]
dev acc :: 0.502
test acc :: 0.213
PS C:\TTNTNN\ASMI-Development-LLM-model-main>
```

3. Tinh chỉnh mô hình:

Chúng tôi đã thực hiện tinh chỉnh mô hình Llama2 bằng cách thêm một đầu phân loại trong classifier.py. Sau khi tinh chỉnh, kết quả phân loại được cải thiện đáng kể. Độ chính xác trên tập SST-5 đạt **0.411**(dev) và **0.408** (test), trên tập CFIMDB đạt **0.820** (dev) và **0.482** (test).

Classification Finetuning for SST

```
load model from finetune-5-2e-05.pt
load 1101 data from data/sst-dev.txt
load 2210 data from data/sst-test.txt
eval: 100% | 14/14 [00:02<00:00, 5.07it/s]
eval: 100% | 28/28 [00:05<00:00, 4.88it/s]
dev acc :: 0.411
test acc :: 0.408
```

```
python run_llama.py --option finetune --epochs 5 --lr 2e-5 --batch_size 10 --train
data/cfimdb-train.txt --dev data/cfimdb-dev.txt --test data/cfimdb-test.txt --label-names
data/cfimdb-label-mapping.json --dev_out cfimdb-dev-finetuning-output.txt --test_out
cfimdb-test-finetuning-output.txt
```

```
load model from finetune-5-2e-05.pt
load 245 data from data/cfimdb-dev.txt
load 488 data from data/cfimdb-test.txt
eval: 100%|██████████████████████████████████████████████████████████████████████████████| 25/25 [01:26<00:00, 3.47s/it]
eval: 100%|██████████████████████████████████████████████████████████████████████████████| 49/49 [02:44<00:00, 3.35s/it]
dev acc :: 0.820
test acc :: 0.482
PS C:\TTNTNN\ASMI-Development-LLM-model-main>
```

Việc tinh chỉnh mô hình cho tác vụ cụ thể (phân loại tình cảm) mang lại kết quả tốt hơn đáng kể so với phương pháp zero-shot. Điều này cho thấy tầm quan trọng của việc tinh chỉnh mô hình cho từng tác vụ cụ thể để đạt được hiệu suất cao. Kết quả trên tập CFIMDB sau khi tinh chỉnh tốt hơn đáng kể so với tập SST-5, có thể do đặc thù của từng tập dữ liệu.

Qua bài tập thực hành, chúng tôi đã hiểu rõ hơn về kiến trúc và cách hoạt động của mô hình ngôn ngữ Llama2. Chúng tôi đã thực hiện thành công các thành phần cốt lõi của mô hình và đánh giá hiệu suất của nó trong các tác vụ khác nhau. Kết quả cho thấy việc tinh chỉnh mô hình là cần thiết để đạt được hiệu suất tốt trong các tác vụ cụ thể.