**BÁO CÁO BTL NHÓM 1 MÔN NLP**

**Thành viên: Nguyễn Tiến Dũng – 23021498**

**I. Giới Thiệu:**

Mục tiêu bàn tập lớn là xây dựng mô hình NLP dịch giữa hai ngôn ngữ Việt-Lào. Sử dụng mô hình **Tranformer** làm nền tảng (Lý do: Hiện nay các mô hình **Transformer** là lựa chọn tiêu chuẩn cho dịch máy, thay thế hoàn toàn RNN/LSTM trước đây.).

**II.Triển khai thực hiện:**

1. **Cài đặt công nghệ:**

* Python 3.12: Yêu câu để chạy các thư viện cần thiết.
* Pytorch: Framework sây dựng mô hình học sâu.
* Cuda: Tăng tốc quá trình training.

1. **Transformer Encoder-Decoder:**

* Các thuật toán:
* **Encoder:** Nhận đầu vào là câu nguồn (Việt), qua nhiều lớp attention và feed-forward.
* **Decoder:** Sinh câu đích (Lào), từng bước một, sử dụng masked attention để không nhìn trước tương lai.
* **Attention:** Sử dụng Multi-Head Attention với RoPE.
* **FeedForward:** Hai lớp Linear với ReLU ở giữa.
* **LayerNorm & Residual:** Ổn định và tăng tốc độ hội tụ.
* **Scaled Dot-Product Attention:**  là phép tính attention hiệu quả, giúp mô hình học được mối liên hệ giữa các từ trong câu.

1. **DataPipeline và DataLoader**

* SentencePiece Tokenizer:

+ Mục đích: Tạo tokenizer chung cho tiếng Việt và tiếng Lào.

+ Thuật toán: Unigram Language Model hoặc BPE (Byte Pair Encoding) (ở đây mặc định là Unigram).

+ Tạo file vocab và model cho cả hai ngôn ngữ, giúp mô hình học được các subword chung.

* Data Pileline:

+ VLDataset: Encode từng câu bằng SentencePiece, thêm token BOS/EOS, trả về tensor cho DataLoader.

1. **Training:**

* Mixed Precision:

+ Sử dụng torch.cuda.amp để tăng tốc và giảm bộ nhớ GPU.

* Loss:

+ CrossEntropyLoss, bỏ qua token PAD.

* Optimizer:

+ Adam với các tham số chuẩn cho Transformer.

* Scaler: Tối ưu hóa quá trình train.

1. **Dịch:**

* Thuật toán:

+ Dịch từng token một, mỗi lần sinh ra token tiếp theo dựa trên output trước đó (greedy decoding).

+ Kết thúc khi gặp EOS hoặc đạt max\_len.

1. **Đánh giá BLEU**

* Sử dụng thư viện **sacrebleu** để tính điểm BLEU trên tập dev.

**III.Kết quả:**

1. **Kết quả dịch:**

sentence = "Tôi yêu mèo."

Dịch: ຂ້າພະເຈົ້າ ຮັກ.

Translate qua gg dịch: Tôi yêu.

sentence = "Tôi yêu tường đại học công nghệ."

Dịch: ຂ້າພະເຈົ້າ ຮັກ ພິວ ເຕີ ພິວ ເຕີ.

Translate qua gg dịch: Tôi thích máy tính.

1. **Điểm blue:**

* Sử dụng tệp Dev để đánh giá:
* import sacrebleu
* # Load dữ liệu test
* with open("data/Dev/dev2023.vi", encoding="utf-8") as f:
* src\_sentences = [line.strip() for line in f if line.strip()]
* with open("data/Dev/dev2023.lo", encoding="utf-8") as f:
* ref\_sentences = [line.strip() for line in f if line.strip()]
* # Dịch toàn bộ
* hyp\_sentences = [translate\_vi\_to\_lo(sent, model, sp) for sent in src\_sentences]
* # Tính BLEU
* bleu = sacrebleu.corpus\_bleu(hyp\_sentences, [ref\_sentences])
* print(f"BLEU: {bleu.score:.2f}")
* Điểm BLUE: 21.56

1. **Đánh giá:**

* Mô hình dịch khá ổn khi dữ liệu train vẫn còn ít (100k câu).
* Điểm BLUE có thể sẽ cải thiện thêm nếu train lâu hơn (Hiện tại mới chỉ train 10 epoch).
* Có thể sử dụng Beam Search để tăng chất lượng dịch.