**MÔ TẢ KẾT QUẢ**

**Trần Đức Hải – KHDL1 – CNCQKHDL1**

**1. Tổng quan**

Mục tiêu: Xây dựng một mô hình học máy có khả năng học các mẫu và cấu trúc từ một đoạn văn bản ngắn cho trước. Sau đó, sử dụng mô hình đã học để tự động sinh ra một chuỗi văn bản mới, bắt chước phong cách của văn bản gốc.

Phương pháp: Sử dụng Mạng Nơ-ron Hồi quy (SimpleRNN) ở cấp độ ký tự. Mô hình sẽ học cách dự đoán ký tự tiếp theo trong một chuỗi dựa trên 3 ký tự đứng ngay trước nó.

**2. Dữ liệu đầu vào (Input)**

Nguồn dữ liệu: Một chuỗi văn bản duy nhất:

*“This is GeeksforGeeks a software training institute”*

Tiền xử lý:

* Bộ từ vựng: Dữ liệu được phân tích để tạo ra một bộ từ vựng gồm 22 ký tự duy nhất (bao gồm cả khoảng trắng).
* Tạo mẫu huấn luyện: Dữ liệu được quét bằng một cửa sổ trượt có kích thước là 3 ký tự. 3 ký tự trong cửa sổ được dùng làm đầu vào, và ký tự ngay sau đó được dùng làm nhãn.
* Mã hóa: Cả đầu vào và nhãn đều được chuyển đổi thành dạng số và sau đó được mã hóa bằng phương pháp One-Hot Encoding để phù hợp với định dạng đầu vào của mô hình.

**3. Mô hình và Quá trình Huấn luyện**

Kiến trúc mô hình:

* Một lớp SimpleRNN với 50 nơ-ron và hàm kích hoạt ReLU
* Một lớp Dense làm lớp đầu ra, có 22 nơ-ron và hàm kích hoạt Softmax để tính toán xác suất cho ký tự tiếp theo

Quá trình huấn luyện:

* Optimizer: Adam
* Hàm mất mát (Loss Function): categorical\_crossentropy
* Số Epochs: 100

Kết quả huấn luyện: Sau 100 epochs, mô hình đạt độ chính xác khoảng 95.14% trên tập dữ liệu huấn luyện, cho thấy nó đã học rất tốt các mẫu trong câu văn bản ngắn được cung cấp.

**4. Kết quả đầu ra**

Đầu vào cho quá trình sinh văn bản: Chuỗi ký tự chuẩn bị trước (seed text) là "This is G".

Văn bản được sinh ra: Mô hình đã tạo ra 50 ký tự tiếp theo, cho ra kết quả cuối cùng là:

*“This is Geeks a software training instituteare training ins”*

Đánh giá kết quả:

* Thành công: Mô hình đã tái tạo thành công gần như toàn bộ câu văn bản gốc. Điều này chứng tỏ RNN đã học và "ghi nhớ" được các chuỗi ký tự từ dữ liệu đầu vào cực kỳ hiệu quả.
* Hạn chế: Ở phần cuối (...instituteare training ins), văn bản bắt đầu trở nên lặp lại và không có ý nghĩa. Nguyên nhân là do dữ liệu huấn luyện quá nhỏ, khiến mô hình bị overfitting (học thuộc lòng). Khi đã hết các mẫu trong dữ liệu gốc, nó bắt đầu lặp lại các mẫu đã học một cách máy móc.