Tìm model bé hơn phù hợp với các model decoder

1. **MobileNetV3**

Link:[1905.02244 (arxiv.org)](https://arxiv.org/pdf/1905.02244)

* Là một kiến trúc CNN nên MobileNetV3 chủ yếu tập trung vào xử lý và trích xuất các đặc trưng từ hình ảnh.

A diagram of a block diagram

Description automatically generated

A diagram of a block diagram

Description automatically generated

* Xử lý ảnh trong model:
* Ảnh đầu vào sẽ thường được điều chỉnh kích thước và chuẩn hóa trước khi đưa vào model.
* MobileNetV3 sử dụng các convolutional layers để trích xuất các đặc trưng từ ảnh đầu vào
* Sau khi ảnh đi qua các lớp tích chập ( convolutional layers) và các lớp kết nối (fully connected layers), model sẽ tạo ra một vector đặc trưng đại diện cho các thông tin quan trọng trong ảnh như hình dạng, kết cấu, màu sắc, đối tượng.
* Tạo văn bản: Sau khi có vector đặc trưng thì có thể kết hợp với một số model decoder như LSTM hoặc Transformer để tạo ra từng từ một cho đến khi hoàn thành câu mô tả ảnh

A diagram of a diagram

Description automatically generated

1. DistilBERT

* Train on : 8 16GB V100 trong 90h
* Link : [1910.01108 (arxiv.org)](https://arxiv.org/pdf/1910.01108)
* DistilBERT là một phiên bản nhỏ gọn của mô hình BERT, được thiết kế nhằm giữ lại hiệu suất của BERT nhưng giảm thiểu kích thước và chi phí tính toán.

1. TinyBERT

* Train on: 1 K80
* Link: [1909.10351 (arxiv.org)](https://arxiv.org/pdf/1909.10351)

1. EfficientNet-Lite

* Link: [tpu/models/official/efficientnet/lite/README.md at master · tensorflow/tpu (github.com)](https://github.com/tensorflow/tpu/blob/master/models/official/efficientnet/lite/README.md)
* Train on: T4