

# Báo cáo VietAI Final Project

Vinh-Hung Tran, Xuan-Vy Nguyen

September 2019

## 1 Tổng quan

Bài viết này nhằm mục tiêu báo cáo những thành tựu đã đạt được trong các thí nghiệm sử dụng tập dữ liệu CheXpert Dataset của Stanford. Bằng việc đầu vào là các loại ảnh về lồng ngực của từng bệnh nhân và kết quả mong đợi là chuẩn đoán chính xác về loại bệnh đang mắc phải. Mô hình sử dụng kiến trúc DenseNet với kết quả đạt được là trên baseline.

## 2 Giới thiệu

Viêm phổi là một trong những căn bệnh gây tử vong ở mọi nhóm tuổi với số ca lên đến 4 triệu người, chiếm 7% dân số thế giới mỗi năm. Hiện nay, để đào tạo ra những bác sĩ chuyên ngành nhằm chuẩn đoán chính xác căn bệnh này đòi hỏi rất nhiều thời gian cho việc đào tạo cũng như xảy ra nhiều rủi ro chuẩn đoán sai với những bác sĩ chưa có nhiều năm kinh nghiệm.

Mục tiêu của nhóm là học tập và thí nghiệm các phương pháp có thể hỗ trợ chuẩn đoán bệnh dành cho các bác sĩ thông qua ảnh Xquang (bao gồm bệnh Viêm phổi). Sử dụng tập dữ liệu CheXpert của Đại học Stanford kết hợp với DenseNet 121, kết quả đạt được vượt ngưỡng cho phép. Hi vọng những cải tiến trong tương lai sẽ đưa nghiên cứu này đạt được những kết quả cao hơn.

## 3 Các thí nghiệm liên quan

### 3.1 Thử nghiệm với các nhãn uncertain và null

Trong tập dữ liệu được cung cấp có rất nhiều nhãn được đánh uncertain và null. Để xử lý vấn đề này có nhiều cách phải thử. Nhóm lần lượt thử nghiệm: đánh tất cả uncertain và null là negative; loại bỏ null khỏi tập dữ liệu;... Mô hình sử dụng việc huấn luyện này không mang lại kết quả tốt.

### 3.2 Chia dataset

Tập dữ liệu gồm hơn 10 labels, việc chia tập train và validation ngẫu nhiên sẽ dẫn đến tình trạng các labels phân bố không hợp lý. Do đó nhóm sử dụng cách chia dataset ra làm 5 phần với tỉ lệ của một label tại từng phần là bằng nhau. Cách chia này đã cài đặt thành công.

### 3.3 Thêm các thuộc tính vào việc huấn luyện

Ngoài các labels chứa thông tin về các loại bệnh mà bệnh nhân mắc phải, tập dữ liệu còn cung cấp thêm thông tin về mã số bệnh nhân, tuổi, giới tính,... Các thông tin này rất quan trọng vì nó thể hiện mối liên hệ giữa những bức hình theo thời gian. Và các nhãn uncertain hoặc null của bệnh nhân theo timeline có thể cho ra kết quả dự đoán tốt hơn bằng những liên hệ này. Mô hình này chưa được cài đặt thành công.

### 3.4 Sử dụng Mixup

Mixup được nhóm nghiên cứu của Facebook đề xuất và tỏ ra cực kì hiệu quả trong những bài toán có sự phân bố rời rạc giữa các nhãn. Để mô hình có khả năng dự đoán tốt các trường hợp không được cho trong tập train, nhóm đã thêm thuật toán này vào mô hình Retinal Disease Detection ở bài giữa kì và đạt được thứ hạng cao. Trong bài toán lần này, nhóm cũng đề xuất sử dụng. Mô hình này chưa cài đặt nhưng không đem lại hiệu quả tốt.

### 3.5 Huấn luyện riêng Lateral Images

Tập dữ liệu có cung cấp Lateral Images, khiến cho việc huấn luyện chung với Frontal Images sẽ làm giảm khả năng dự đoán chính xác của mô hình. Do đó nhóm đã đề xuất và cài đặt việc huấn luyện riêng cho loại ảnh này. Mô hình này cài đặt chưa thành công.

## 4 Phương pháp sử dụng

### 4.1 Model sử dụng

Nhóm sử dụng mô hình DenseNet 121. Đây là mô hình rất phù hợp cho bài toán classification. Hơn nữa mô hình này còn giải quyết tốt vấn đề giữa accuracy và thời gian training.

### 4.2 AUROC-score

AUROC cho thấy khả năng dự đoán đúng của phương pháp về mặt tổng thể hơn là xét tại một ngưỡng threshold, điều này giúp AUROC thể hiện tính tổng quan khi so sánh với F1 score. Hơn nữa, AUROC cho thấy mối tương quan rõ ràng giữa True Positive Rate và False Positive Rate.

### 4.3 EarlyStopping

Vì thời gian giới hạn trên Kaggle là 9 tiếng, nhóm đã sử dụng thêm kỹ thuật này. Mặc dù nó gây ra hạn chế là không tìm được kết quả tốt nhất nhưng rất phù hợp với khả năng cung cấp phần cứng để huấn luyện dữ liệu. Cuối cùng nhóm vẫn đạt được kết quả trên baseline.

## 5 Kết quả đạt được

Sử dụng phương pháp như trên, nhóm đã đạt được độ chính xác **0.74921**. Dù kết quả chưa cao nhưng đã vượt ngưỡng baseline. Với những thí nghiệm đã thực hiện nhưng chưa thành công,

hi vọng trong tương lai sẽ áp dụng cách thích hợp và đem lại kết quả cao hơn nữa.

## **6 Kết luận**

Với những thí nghiệm và kết quả đã đạt được, nhóm vẫn còn có thể phát triển thêm những thí nghiệm chưa hoàn thành nhằm phục vụ việc nghiên cứu và có thể đạt kết quả cao hơn.