Exercise 1: Image Classification

Thu thập dữ liệu để huấn luyện mô hình

Với yêu cầu bài toán là huấn luyện mô hình phân loại ảnh chó và mèo. Em tiến hành thu thập dữ liệu từ Kaggle với tổng cộng 10000 ảnh. Mỗi lớp gồm 5000 ảnh. Nguồn dữ liệu em public ở https://www.kaggle.com/datasets/anhtuandaotran/dog-cat-dataset/data

Xử lí dữ liệu

Với bộ dữ liệu trên, hai lớp chó và mèo cân bằng về số lượng mẫu và số lượng 5000 ảnh là đủ lớn. Do đó, em không sử dụng tăng cường dữ liệu mà tiến hành huấn luyện mô hình luôn. Em chia tập dữ liệu thành 2 tập train và validation với tỉ lệ 8:2, đảm bảo tỉ lệ đúng với cả 2 lớp.

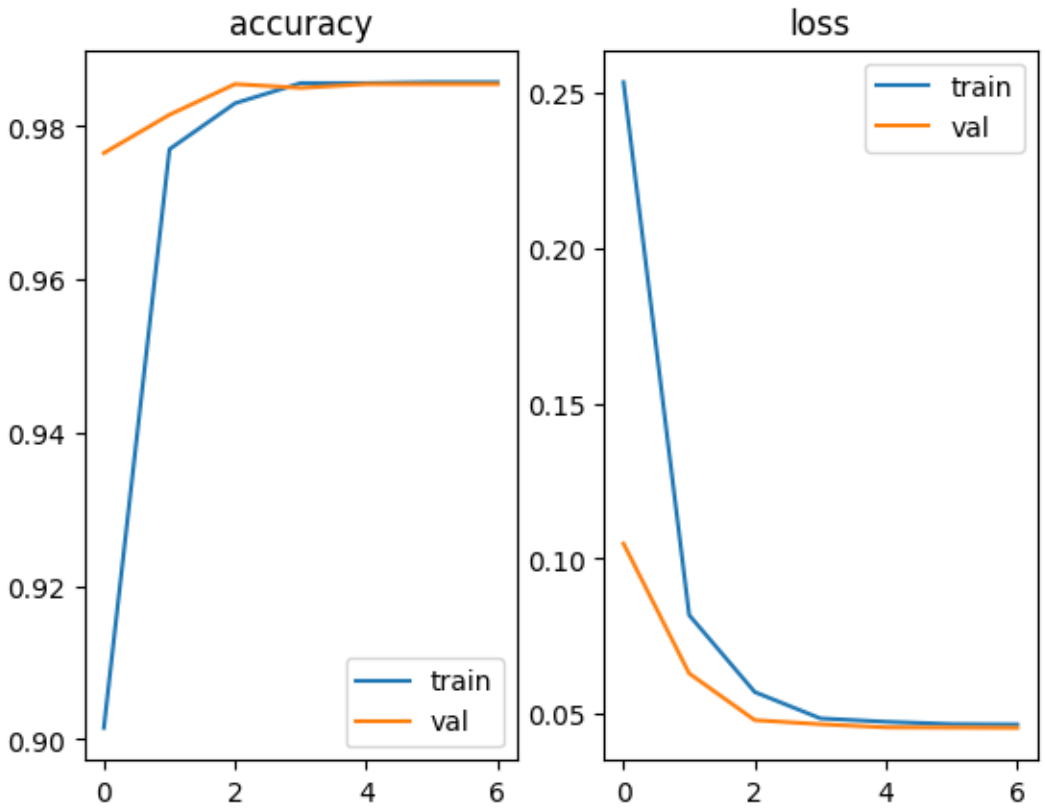
Quá trình huấn luyện mô hình:

Với bộ dữ liệu và yêu cầu đề bài, em sử dụng kĩ thuật Transfer Learning để có được mô hình phân loại ảnh có độ chính xác cao cũng như tính tổng quát. Backbone em sử dụng là ResNet50 được pretrain trên tập ImageNet1K. Ngoài ra, em còn sử dụng các kĩ thuật giảm learning rate trong quá trình huấn luyện, dừng sớm (early stopping) để ngăn chặn mô hình overfit trên tập training.

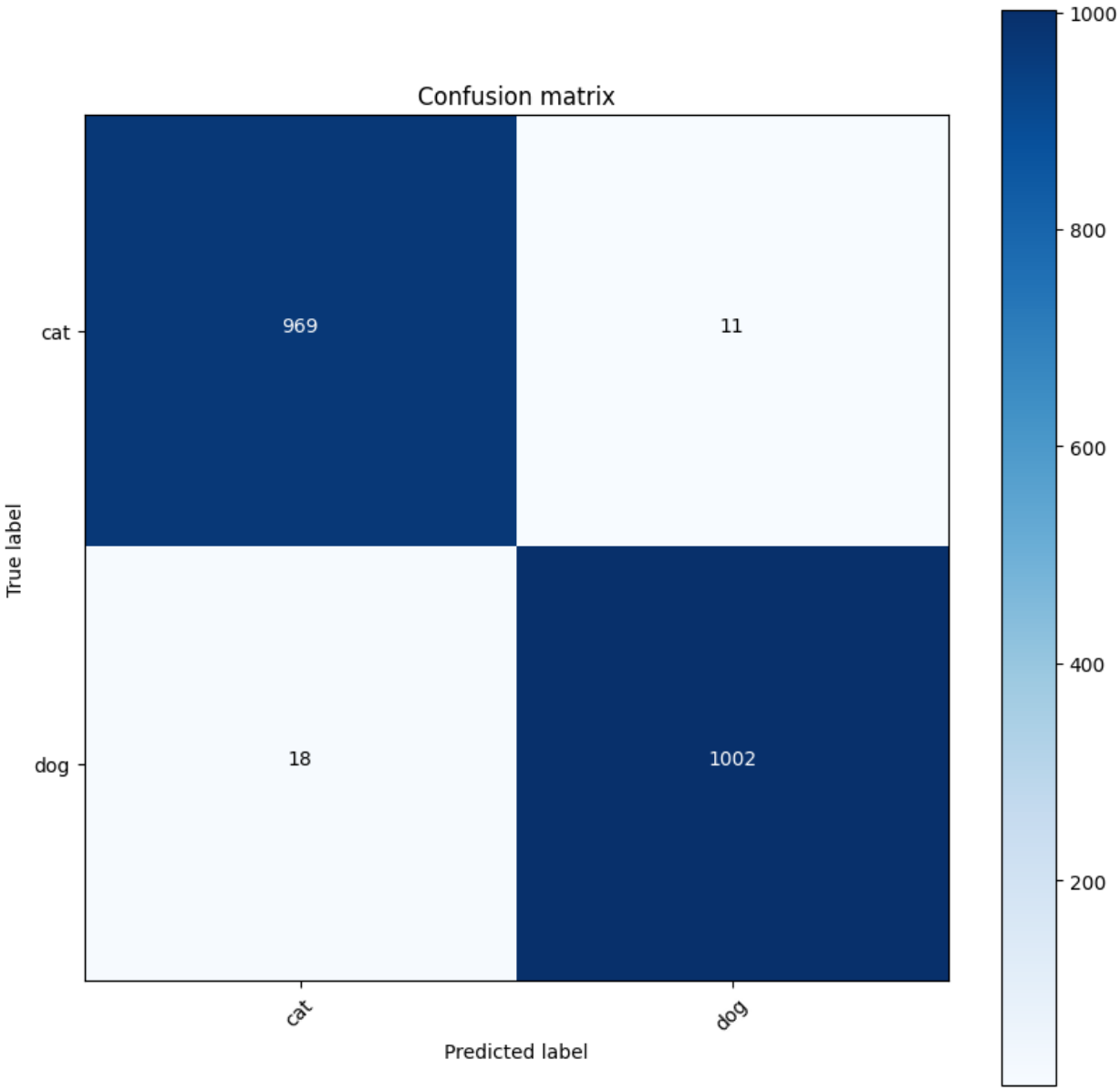
Chọn thang đo đánh giá

Vì bộ dữ liệu mang tính cân bằng, do đó em sử dụng độ đo Accuracy để đánh giá độ chính xác phân loại của mô hình. Ngoài ra, em còn sử dụng thêm Confusion Matrix để đánh giá được chi tiết Precision, Recall và F1 Score cho từng lớp đối tượng.

Kết quả thu được



Hình 1: Kết quả accuracy và loss trong quá trình huấn luyện mô hình



Hình 2: Confusion matrix trên tập validation

	precision	recall	f1-score	support
cat	0.98	0.99	0.99	980
dog	0.99	0.98	0.99	1020
accuracy			0.99	2000
macro avg	0.99	0.99	0.99	2000
weighted avg	0.99	0.99	0.99	2000

Hình 3: Kết quả Precision, Recall và F1 cho từng lớp trên tập huấn luyện

Nhận xét

Mô hình đạt độ chính xác cao (accuracy là 98.58% trên tập train). Ngoài ra, không bị overfit trên tập train (accuracy trên tập validation đạt 98.55%). Điều này chứng tỏ kĩ thuật Transfer Learning và các kĩ thuật em sử dụng có hiệu quả tốt.

Exercise 2: Text to speech

Chuẩn bị dữ liệu:

Thu thập các bộ dữ liệu giọng nói tiếng Việt, bộ dữ liệu nên có tính đa dạng về số lượng người nói; loại giọng nói (accent); nội dung nói nhiều về đề tài, nhiều lĩnh vực trong thực tế. Một số tập dữ liệu tham khảo là: VIVOS, VLSP, VietTTS

Xử lí dữ liệu:

Dữ liệu âm thanh ở dạng audio cần phải xử lí biến đổi thành dạng spectrogram để các mô hình có thể học và đưa ra dự đoán.

Dữ liệu văn bản cần đưa về dạng các token hoặc tách ra thành âm vị.

Xây dựng mô hình chuyển từ văn bản đầu vào ra spectrogram:

Xử dụng các mô hình SOTA hiện nay như Tacontron2, FastSpeech.

Hậu xử lí dữ liệu từ spectrogram sang dạng audio: Vocoder

Sử dụng các Vocoder như Wavenet, MelGAN, WaveFlow.

Đánh giá hiệu suất và fine-tune mô hình

Để có được một mô hình tốt cần phải có chọn được các phương pháp đánh giá chính xác, trực quan. Fine-tune lại mô hình để có kết quả tốt hơn trong các trường hợp, nhất là khi áp dụng vào các domain cụ thể.

Deploy mô hình

Mô hình sau khi đạt được kết quả tốt (về mặt thời gian và độ chính xác) có thể được deploy để sử dụng trong thực tế. Trong quá trình sử dụng, cần phải thường xuyên bảo trì, nâng cấp mô hình.

Nhận xét

Các mô hình Text To Speech trên đều nhận đầu vào là dữ liệu văn bản ở dạng chữ. Nếu muốn chuyển dữ liệu từ file PDF, website, sách,… thì cần phải sử dụng thêm các mô hình OCR để chuyển các định dạng trên thành dạng text, sau đó mới dùng các mô hình text to speech đề xuất.