Section 1: Building a Liveness Classifier

Le Tien Quyet

April 2025

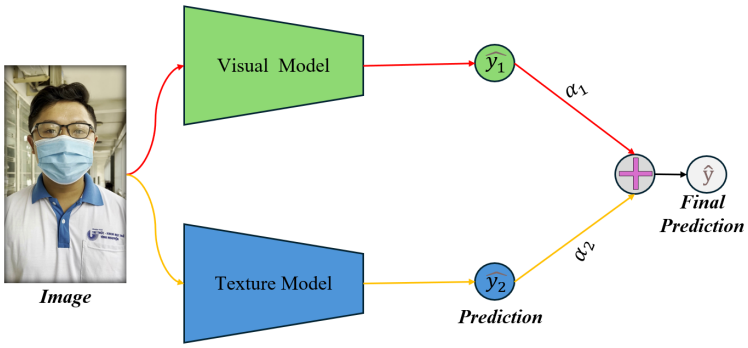
1 Load & Preprocess the Dataset

Trước khi đưa vào mô hình, ảnh đầu vào được xử lý qua nhiều bước chuẩn hóa và tăng cường nhằm tăng độ đa dạng của dữ liệu và đảm bảo phù hợp với yêu cầu của mô hình học sâu, từ đó phát huy tối đa hiệu suất mô hình.
Đầu tiên, toàn bộ ảnh được chuyển đổi về kích thước 224×224. Việc chuẩn hóa kích thước đầu vào này là cần thiết để đảm bảo tính nhất quán, đồng thời tương thích với các mô hình đã được huấn luyện trước trên các tập dữ liệu lớn với kích thước đó.
Tiếp theo, ảnh được tăng cường dữ liệu thông qua biến đổi về độ sáng và độ tương phản ($\pm 15\%$). Điều này giúp mô hình học được các đặc trưng ổn định hơn trước các thay đổi tự nhiên trong nhiều điều kiện môi trường khác nhau.
Cuối cùng, ảnh được chuẩn hóa theo thống kê của tập dữ liệu ImageNet cho từng kênh màu RGB ($\mu = [0.485, 0.456, 0.406], \sigma = [0.229, 0.224, 0.225]$). Việc chuẩn hóa này giúp đưa các giá trị điểm ảnh về cùng một phân phối với dữ liệu mà các mô hình pretrained đã được huấn luyện, từ đó cải thiện tính ổn định của quá trình huấn luyện và đẩy nhanh tốc độ hội tụ.

2 Approach Method

Trong bài toán phân loại ảnh spoof và ảnh normal, yếu tố then chốt nằm ở khả năng phát hiện những khác biệt tinh vi mang tính kết cấu hơn là hình thái tổng thể. Ảnh spoof vốn được tạo ra bằng cách in ảnh, chiếu lại qua màn hình hoặc sử dụng thiết bị nhân tạo nên thường không còn giữ được các đặc trưng kết cấu tự nhiên của da người. Thay vào đó, chúng để lộ ra nhiều dấu hiệu như nhiễu in ấn, mất chi tiết bề mặt, phản chiếu ánh sáng không tự nhiên và xảy ra hiện tượng Moiré (xem Ảnh 1).
Do đó, khả năng nắm bắt và phân tích đặc trưng kết cấu là cực kỳ quan trọng để phát hiện các ảnh spoof tinh vi. Điều này lý giải cho hướng tiếp cận kết hợp giữa một Visual Model để xử lý đặc trưng tổng quát và một Texture Model chuyên biệt, nhằm tăng cường khả năng phát hiện các tín hiệu cục bộ mang tính chất vật lý đặc trưng của ảnh spoof (xem Ảnh 2).

Hình 1: Một số ảnh spoof trong tập dữ liệu huấn luyện



Hình 2: Tổng quan về phương pháp tiếp cận

Đầu ra của mỗi mô hình được biểu diễn tương ứng là:

$$\begin{aligned} \hat{y_1} &= Visual\_Model(Image) \\ \hat{y_2} &= Texture\_Model(Image) \end{aligned} \tag{1}$$

Trong đó: $\hat{y_1}$ và $\hat{y_2}$ lần lượt là xác suất mà mỗi mô hình dự đoán ảnh thuộc lớp spoof.
Để tận dụng ưu điểm của cả hai mô hình, kết quả dự đoán cuối cùng được tính

bằng cách tính tổng có trọng số giữa hai đầu ra trên:

$$\begin{aligned} \hat{y} &= \alpha_1 * \hat{y_1} + \alpha_2 * \hat{y_2} \\ &= \alpha * \hat{y_1} + (1 - \alpha) * \hat{y_2}, \text{ với } 0 \leq \alpha \leq 1 \end{aligned} \tag{2}$$

Việc lựa chọn giá trị α tùy vào mức độ ưu tiên giữa Visual Model và Texture Model. Thông qua đánh giá trên tập dữ liệu kiểm thử để tìm ra giá trị α tối ưu.

2.1 Visual Model

Đầu tiên, sử dụng ResNet-50 làm kiến trúc mạng nền tảng nhờ vào khả năng học đặc trưng mạnh mẽ từ các tầng sâu và thiết kế *residual connections* giúp giảm hiện tượng mất mát thông tin khi mạng trở nên sâu. ResNet-50 đã được chứng minh hiệu quả trong nhiều bài toán thị giác máy tính, đặc biệt trong tác vụ phân loại ảnh. Việc thử nghiệm với ResNet-50 nhằm xây dựng một baseline mạnh để làm cơ sở so sánh với các kiến trúc nhẹ hơn hoặc hiện đại hơn.
Tiếp theo, MobileNetV2 được thử nghiệm nhằm đánh giá tính hiệu quả về mặt tính toán khi triển khai lên các thiết bị thực tế. MobileNetV2 sử dụng kiến trúc *depthwise separable convolution* cho phép giảm đáng kể số lượng tham số và phép tính mà vẫn duy trì độ chính xác ở mức tương đối.
Cuối cùng, Vision Transformer (ViT) với khả năng học đặc trưng và mối liên quan giữa các patches trong ảnh theo cơ chế *self-attention* thay vì phép tích chập truyền thống. Việc sử dụng ViT giúp đánh giá xem kiến trúc phi tích chập có thể tận dụng thông tin kết cấu bề mặt một cách hiệu quả hơn hay không.

2.2 Texture Model

Deep TEN (Deep Texture Encoding Network) tích hợp một Encoding Layer hoạt động bằng cách chuyển đổi mỗi đặc trưng thành một residual vector. Cụ thể, nó tính chênh lệch giữa đặc trưng đó và các vector mã hóa trong từ điển. Sau đó, các residual vector này được trọng số và tổng hợp lại để tạo nên một biểu diễn toàn cục về kết cấu của ảnh.
Nhờ khả năng tập trung vào kết cấu, Deep TEN bổ sung những thông tin mà các mô hình thị giác thông thường có thể bỏ sót, từ đó nâng cao đáng kể hiệu quả bài toán.

3 Detail Setting for Training

Cả Visual Model cũng như Texture Model đều được điều chỉnh lại lớp cuối để phù hợp cho bài toán phân loại nhị phân. Các mô hình được khởi tạo với trọng số từ các mô hình pretrained (ImageNet-1K đối với Visual Model và MINC đối với Texture Model). Trong quá trình huấn luyện, toàn bộ các tham số của mô hình gốc được đóng băng, chỉ có lớp cuối được huấn luyện để thực hiện nhiệm vụ phân loại spoof và normal.
Các mô hình được huấn luyện trong 40 epoch trên 1 GPU P100 (môi trường

của Kaggle), sử dụng batch size = 128 và hàm mất mát Binary Cross-Entrpoy. Dùng thuật toán Adam với learning rate khởi tạo là 0.001 để tối ưu các tham số. Tuy nhiên, trong quá trình chạy thử nghiệm với ResNet-50 thì việc cố định giá trị learning rate không hiệu quả. Do đó, learning rate sẽ được giảm 10% sau mỗi 3 epoch nếu validation loss không cải thiện rõ rệt.

Để đánh giá vai trò đóng góp của từng thành phần, ba giá trị của α được sử dụng: với $\alpha = 1$ (chỉ sử dụng Visual Model), $\alpha = 0$ (chỉ sử dụng Texture Model) và $\alpha = 0.5$ (kết hợp đồng đều cả hai mô hình). Thiết lập này giúp kiểm chứng hiệu quả của từng thành phần riêng lẻ cũng như mức độ cộng hưởng khi tích hợp đặc trưng thị giác và kết cấu bề mặt.