



BÁO CÁO KHÓA LUẬN TỐT NGHIỆP








PHÂN LỚP ĐA ĐỐI TƯỢNG DỰA TRÊN MÔ HÌNH HỌC SÂU



Giảng viên hướng dẫn:
TS. Bùi Tiến Lên

Sinh viên thực hiện:
Hồ Đăng Cao
Đỗ Đức Duy

NỘI DUNG

-  Giới Thiệu Bài Toán
-  Phương Pháp Giải Quyết
 -  Công trình liên quan
 -  Cải tiến đề xuất
 -  Thực nghiệm và Đánh giá
-  Kết Luận Chung
-  Hướng Phát Triển



GIỚI THIỆU BÀI TOÁN



**Mô hình phân
lớp đa đối
tượng**

Salmon

Bean

Mushroom

⋮

Rice



Giới Thiệu Tập Dữ Liệu

- Nguồn gốc: từ cuộc thi Food Recognition Benchmark 2022.
- Mô tả: các bức ảnh về bữa ăn hằng ngày được phân loại và gán nhãn.
- Số lượng ảnh:
 - Tập huấn luyện: 54392.
 - Tập đánh giá: 946.
- Số lượng nhãn: 323.



Vấn đề 1: Vấn đề kỹ thuật.



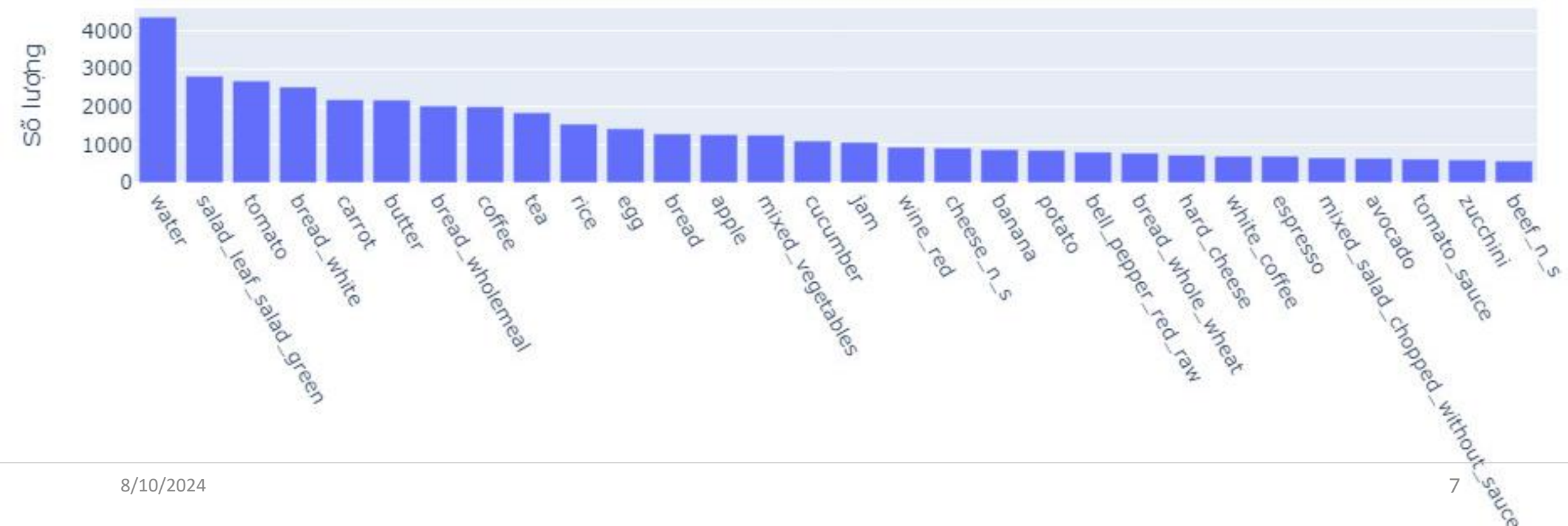


Giới Thiệu Vấn Đề

Vấn đề 2: Số lượng nhãn ở các bức ảnh có sự chênh lệch lớn.

Số lượng nhãn trên hình ảnh	1	2	3	4	5	6	7	8	9	10	11	13
Tỉ lệ % trong tập dữ liệu	62	18.68	10.66	5.01	2.12	0.9	0.38	0.15	0.06	0.03	0.01	0

Vấn đề 3: Số lượng ảnh ở mỗi nhãn có sự chênh lệch lớn.

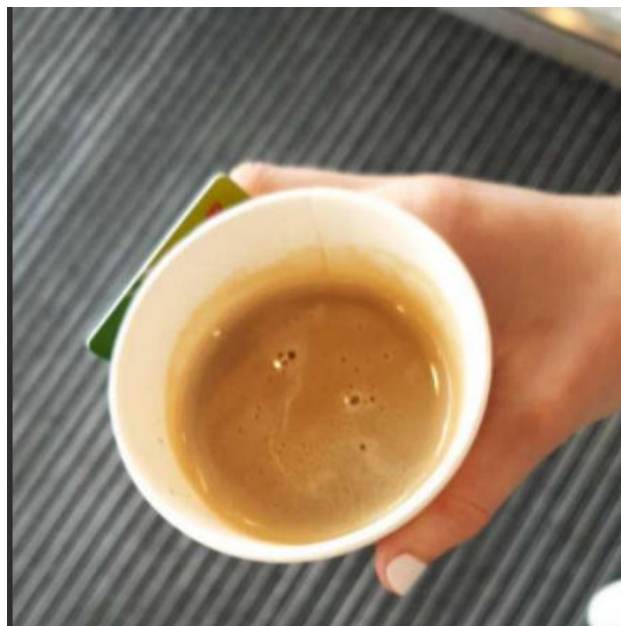




Vấn đề 4: Các **nhãn** có nghĩa thuộc **cùng trường từ vựng**.

coffee – espresso

bread – bread_white



coffee



espresso



PHƯƠNG PHÁP GIẢI QUYẾT

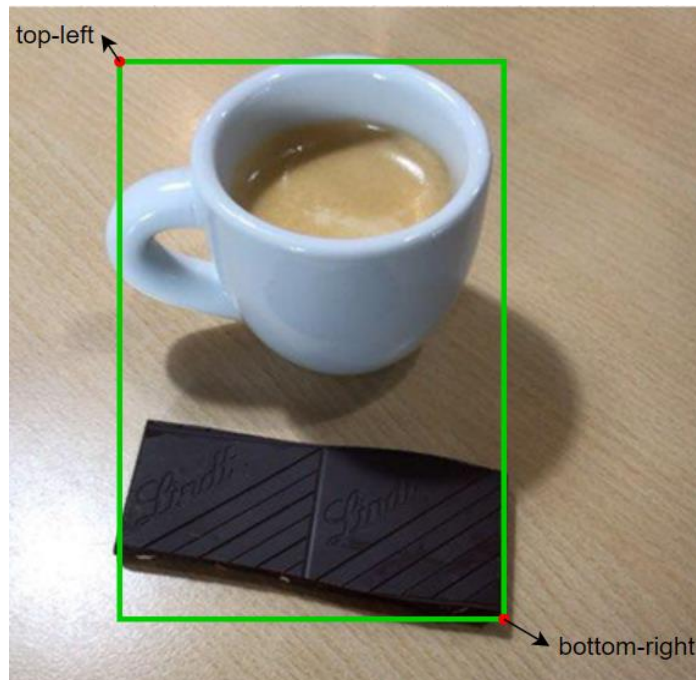
VẤN ĐỀ 1



Cắt ảnh đầu vào



ảnh gốc



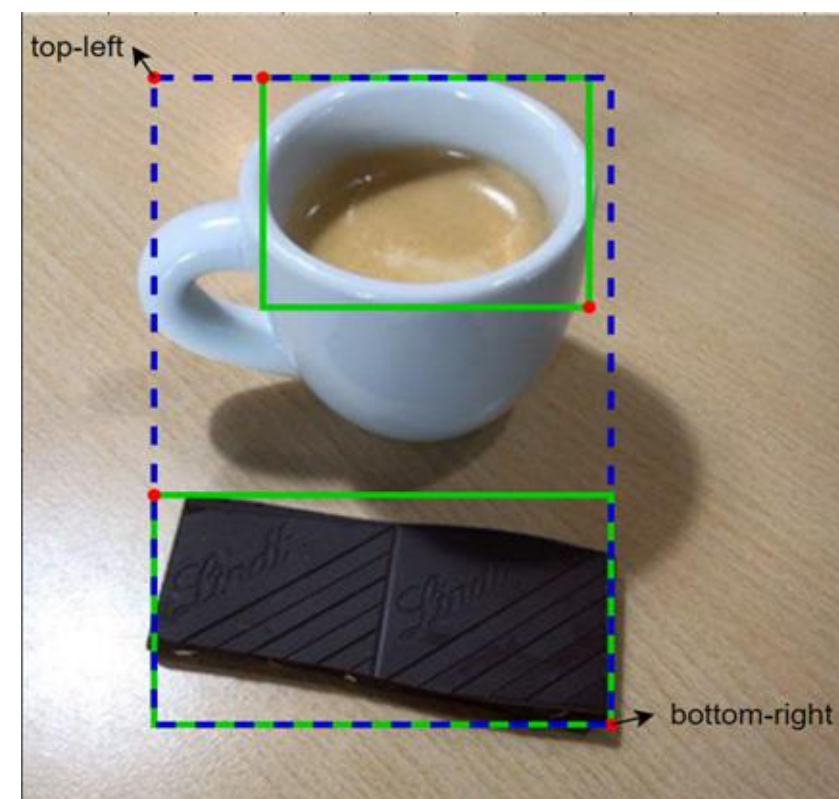
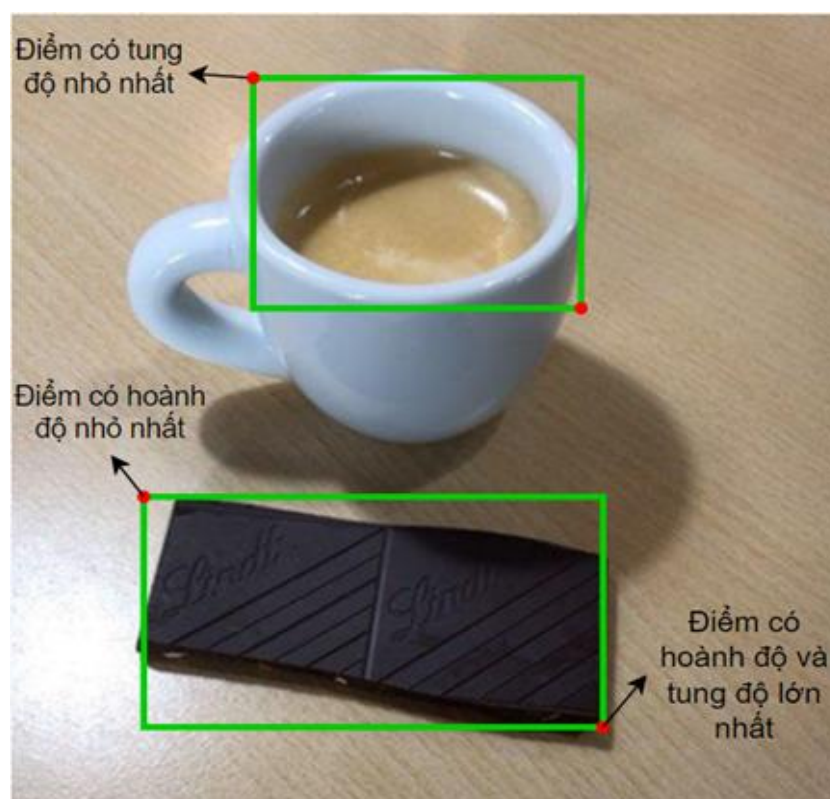
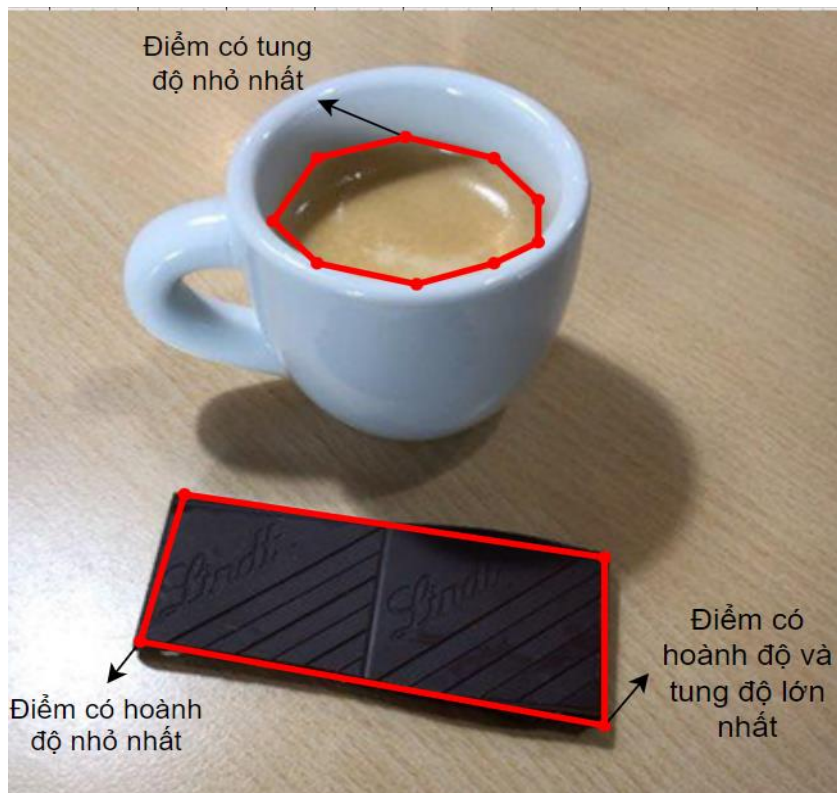
xác định vùng cắt



ảnh sau khi cắt



Xác định vùng cắt



Xác định từ các **phân vùng**.

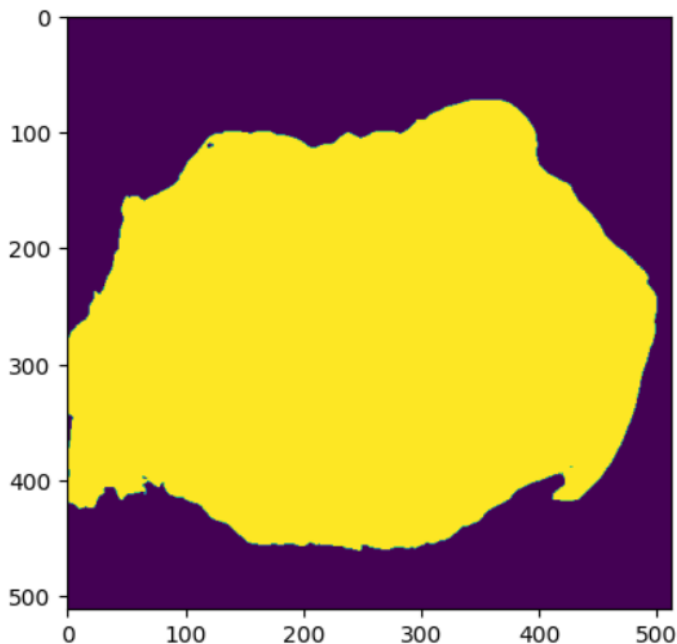
Xác định từ các **khung chứa**.



Dùng mô hình huấn luyện sẵn



ảnh gốc



ảnh sau khi qua mạng U-Net



ảnh sau khi cắt

CÁC VẤN ĐỀ CÒN LẠI

- Đơn nhân dương.
- C-Tran.
- Mô hình kết hợp.



Đơn Nhận Dư



Tập dữ liệu

- \mathbf{z}_n : vector nhãn quan sát của ảnh thứ n .

1	\emptyset	\emptyset	...	\emptyset
---	-------------	-------------	-----	-------------

Chỉ có 1 nhãn dương.

Vector dự đoán \mathbb{f}_n

- $\mathcal{L}_{BCE}^+(\mathbb{f}_n, \mathbf{z}_n)$

Giải pháp:

- Bổ sung nhãn âm.
- Phạt dự đoán nhiều nhãn dương.



Các hàm mất mát

Bổ sung nhãn âm

Giả sử các nhãn không được quan sát là âm: $\mathcal{L}_{AN}(\mathbf{f}_n, \mathbf{z}_n)$ → nhiều nhãn làm giảm độ chính xác.

- Thêm **trọng số**

Hàm mất mát giả sử nhãn âm yếu: $\mathcal{L}_{WAN}(\mathbf{f}_n, \mathbf{z}_n)$

- Kết hợp **làm mịn nhãn** (LS) cho mỗi lớp

Hàm hàm mất mát mới: $\mathcal{L}_{AN-LS}(\mathbf{f}_n, \mathbf{z}_n)$



Các hàm mất mát

Phạt dự đoán nhiều nhãn dương

Điều chuẩn dương kì vọng: $\mathcal{L}_{EPR}(\mathbf{F}_B, \mathbf{Z}_B) = \frac{1}{|B|} \sum_{n \in B} \mathcal{L}_{BCE}^+(\mathbb{f}_n, \mathbf{z}_n) + R_k(\mathbf{F}_B)$

Bộ phận ước lượng nhãn $g(x_n; \phi) = \tilde{y}_n$.

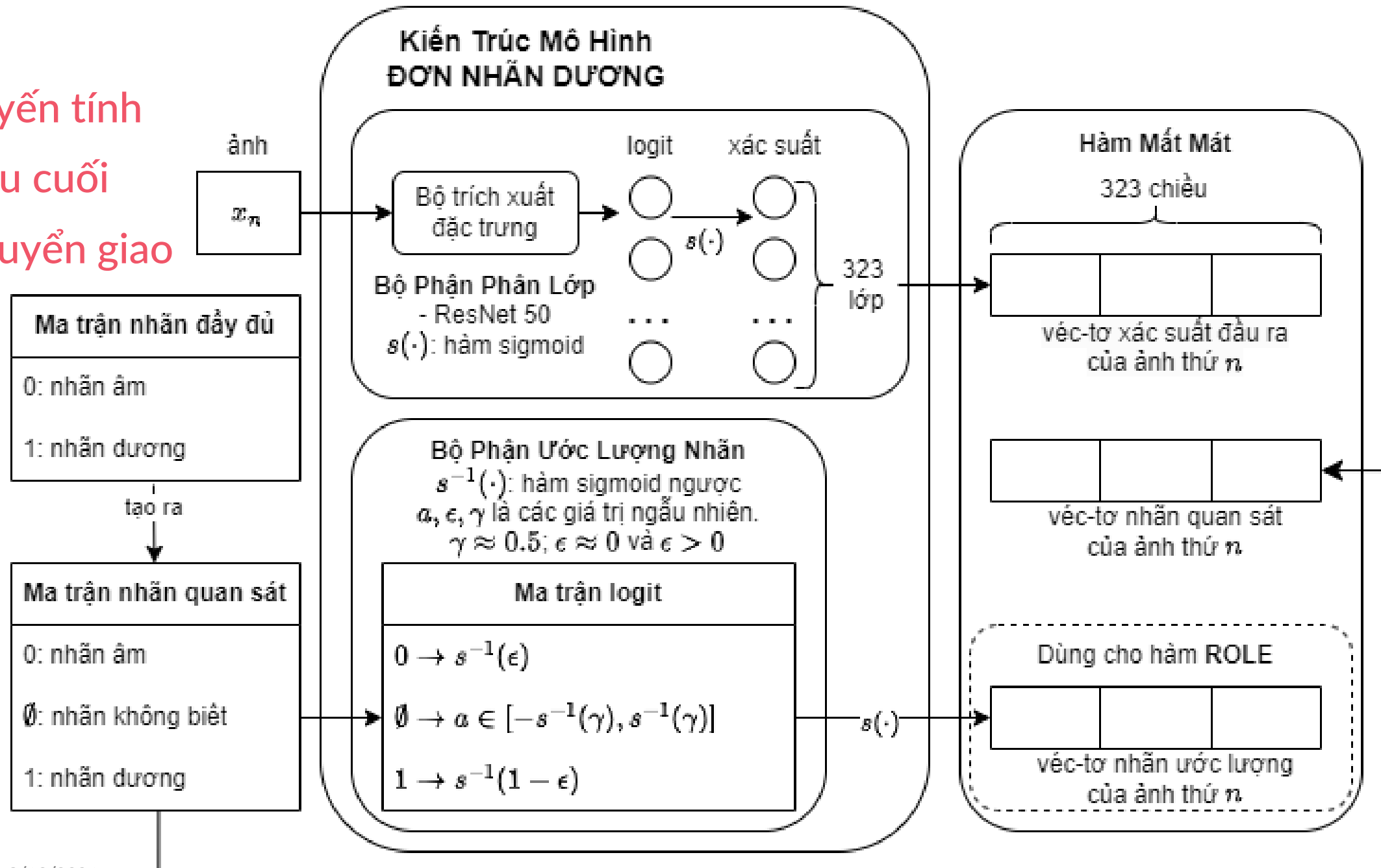
$k \neq \hat{k}(F_B)$

$$\mathcal{L}'(F_B | \tilde{Y}_B) = \frac{1}{|B|} \sum_{n \in B} \mathcal{L}_{BCE}(\mathbb{f}_n, sg(\tilde{y}_n)) + \mathcal{L}_{EPR}(\mathbf{F}_B, \mathbf{Z}_B)$$

Ước lượng nhãn trực tuyến điều chuẩn:

$$\mathcal{L}_{ROLE}(F_B, \tilde{Y}_B) = \frac{\mathcal{L}'(F_B | \tilde{Y}_B) + \mathcal{L}'(\tilde{Y}_B | F_B)}{2}$$

Tuyến tính
Đầu cuối
Chuyển giao





Cải Tiến Đề Xuất

Đơn nhận dưỡng



Chuyển vector nhãn z sang tập xác suất tương ứng.
Nhãn không biết xem như nhãn âm.

1	0	0	...	0
---	---	---	-----	---

Mục tiêu: $x_i = \sigma^{-1}(z_i)$

Áp dụng kĩ thuật làm mịn nhãn.

$1 - \epsilon$	$\frac{\epsilon}{L - 1}$	$\frac{\epsilon}{L - 1}$...	$\frac{\epsilon}{L - 1}$
----------------	--------------------------	--------------------------	-----	--------------------------

\parallel
 $\sigma(x^+)$ \parallel
 $\sigma(x^-) \rightarrow x^+$ phụ thuộc x^- .



Huber: $\mathcal{L}_{HU}(\mathbf{f}_n, \mathbf{z}_n)$ $\begin{cases} \text{MSE} \\ \text{MAE} \end{cases}$

Focal: $\mathcal{L}_{FO}(\mathbf{f}_n, \mathbf{z}_n) = \mathcal{L}_{AN}(\mathbf{f}_n, \mathbf{z}_n)$ $\begin{cases} \alpha_i \\ (1 - f_{ni})^\gamma \end{cases}$



Thực Nghiệm và Đánh Giá

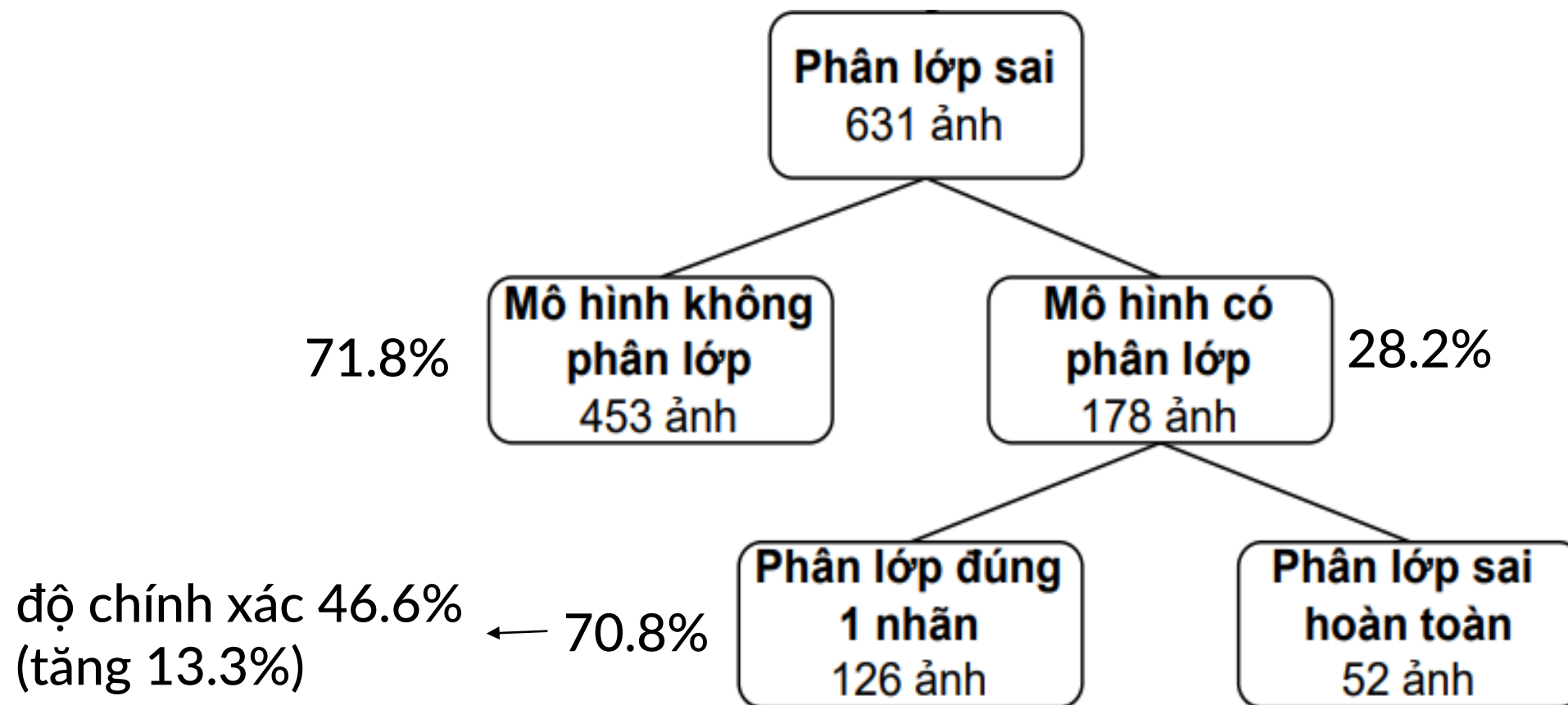
Đơn nhân dươg



Các kết quả tốt nhất với **ResNet 50** giữa:

- Các tốc độ học (10^{-3} , 10^{-4} , 10^{-5}).
- Các kích thước lô (8, 16, 32).
- Trên tập dữ liệu gốc (chưa cắt).

Hàm mất mất	Chế độ huấn luyện	Hàm kích hoạt	mAP tập đánh giá	mAP tập kiểm tra
FO	đầu cuối	sigmoid	0.3724	0.7938
ROLE	đầu cuối	sigmoid	19.0979	26.5441
HU	đầu cuối	sigmoid	21.0043	32.6947
AN-LS	đầu cuối	sigmoid	24.2951	34.6173
AN-LS	chuyển giao	sigmoid	24.4427	34.8327
HU	chuyển giao	softmax	0.4554	1.5206





Số lượng	Các cặp (nhãn đúng, nhãn đoán)
4	(water, soft_drink_with_a_taste), (espresso, coffee)
3	(water, glucose_drink_50g), (espresso, ristretto_with_caffeine)
2	(water, water_with_lemon_juice), (bread_wholemeal, bread_whole_wheat), (mixed_salad_chopped_without_sauce, salad_leaf_salad_green), (coffee, white_coffee), (coffee, ristretto_with_caffeine)

Các **nhóm nhãn dễ nhầm lẫn** với nhau chiếm **46.2%**.

Các nhãn **tương đồng về nghĩa**. → **Ảnh hưởng đến độ tin cậy về kết quả** của mô hình.



water



soft_drink_with_a_taste



glucose_drink_50g



water_with_lemon_juice

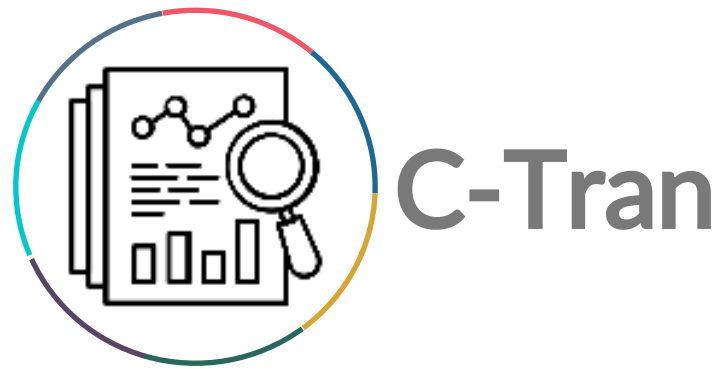


Ưu điểm: làm tốt trong vấn đề trích xuất và nhận diện các đặc trưng ảnh với số lượng nhãn cần đánh thấp.

Hạn chế: các vấn đề về ý nghĩa nhãn làm ảnh hưởng lớn đến kết quả phân lớp.

Giải pháp: cần một mô hình có thể học được mối liên hệ giữa các nhãn.

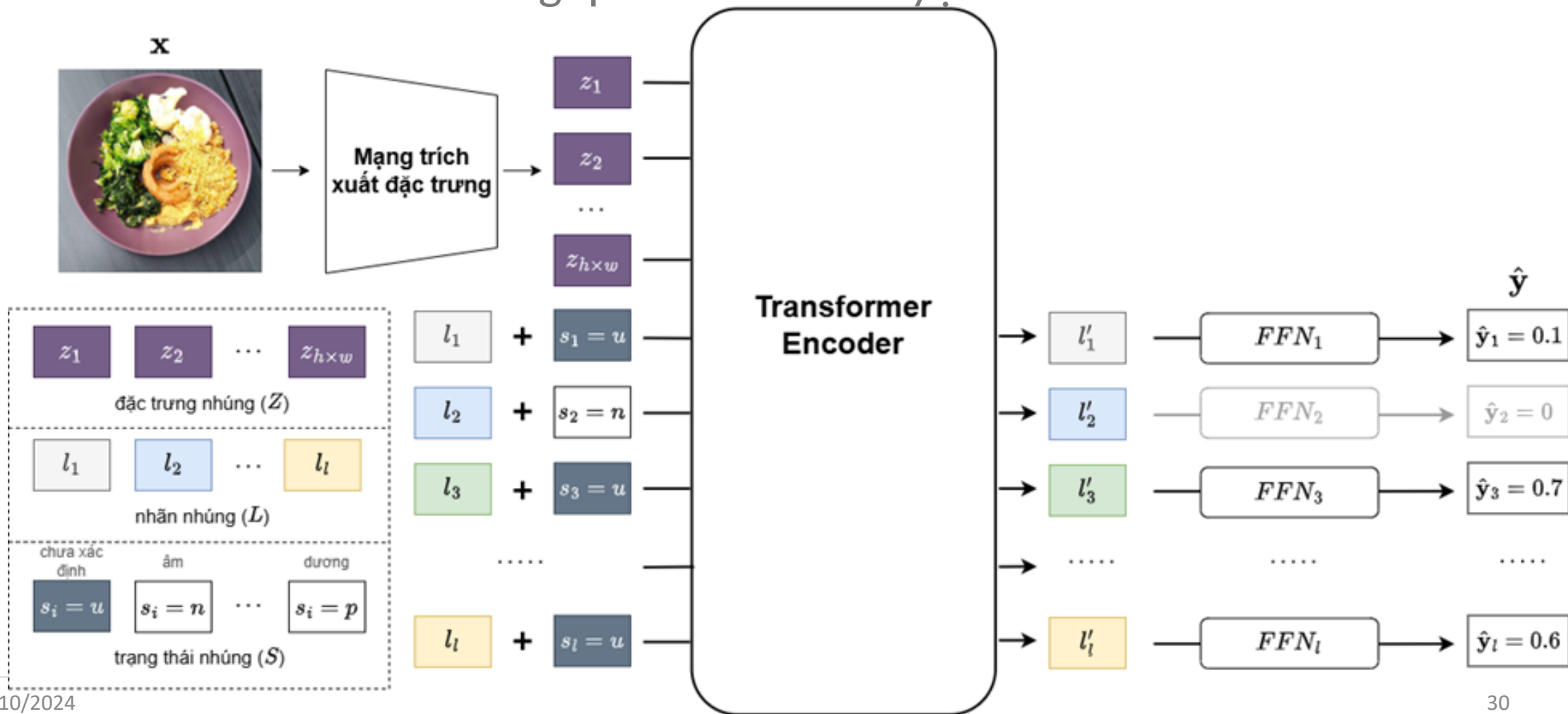
→ C-Tran.





Tổng quan

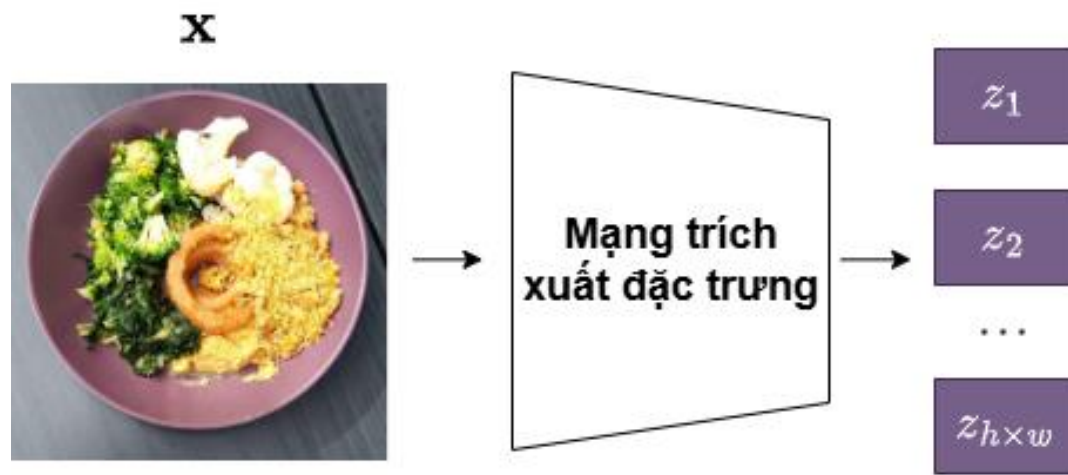
- Khai thác sự phụ thuộc giữa các đặc trưng và các nhãn.
- Che nhãn hình ảnh trong quá trình huấn luyện.





Các thành phần chính

1. Nhúng đặc trưng Z



Các vector $z_i \in Z = \mathbb{R}^d$, đại diện cho các vùng được ánh xạ từ các mảng không gian gốc của ảnh thông qua mạng trích xuất đặc trưng.



Các thành phần chính

2. Nhúng nhãn L



$L = \{l_1, l_2, \dots, l_l\}, l_i \in \mathbb{R}^d$, đại diện cho các nhãn l có thể có trong tập dữ liệu.

$$l_1 + s_1 = u$$

$$l_2 + s_2 = n$$

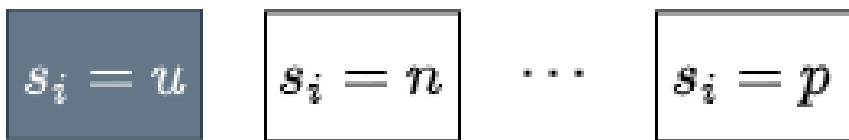
$$l_3 + s_3 = u$$

.....

$$l_l + s_l = u$$

3. Thêm thông tin về nhãn thông qua nhúng trạng thái S

$$\tilde{l}_i = l_i + s_i$$

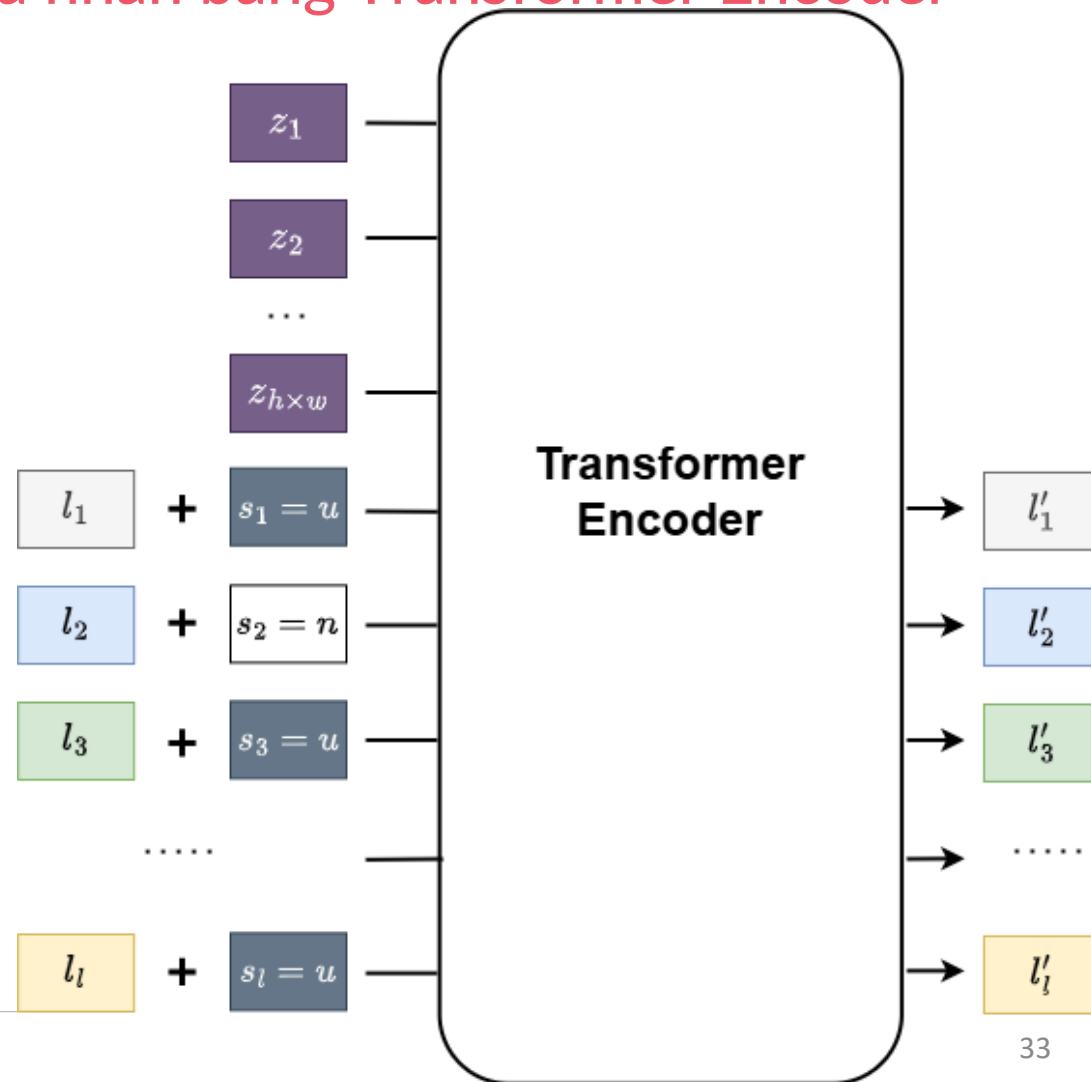


$s_i \in \{U, N, P\}$: không xác định (U), âm (N), dương (P).



4. Mô hình hóa sự tương tác giữa đặc trưng và nhãn bằng Transformer Encoder

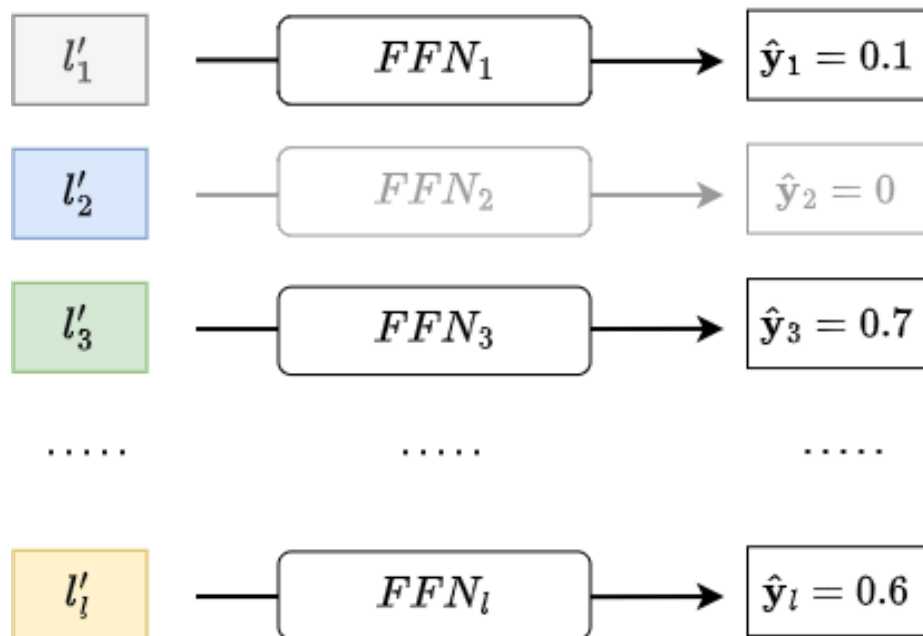
- $H = \{z_1, \dots, z_{h \times w}, \tilde{l}_1, \dots, \tilde{l}_l\}$ là đầu vào của Transformer Encoder.
- Đầu ra là $H' = \{z'_1, \dots, z'_{h \times w}, l'_1, \dots, l'_l\}$





Các thành phần chính

5. Quá trình suy luận để phân loại nhãn



6. Hàm mất mát

L_{BCE} : Binary Cross Entropy

Mạng chuyển tiếp độc lập FFN_i cho l'_i gồm 1 lớp tuyến tính.
Sau đó, dùng hàm **sigmoid** để tính giá trị xác suất cho các nhãn l'_i .

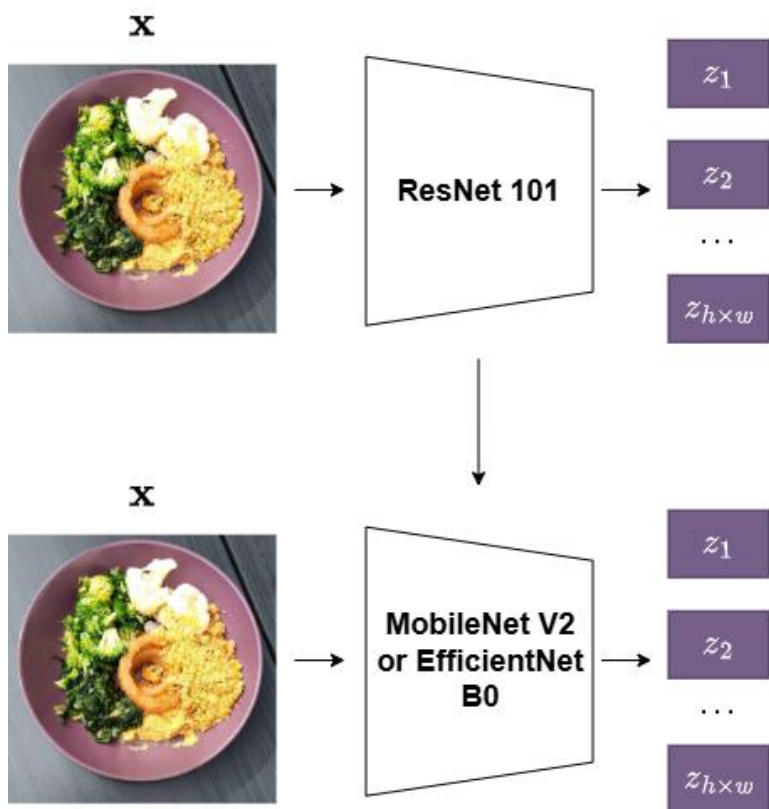


Cải Tiến Đề Xuất

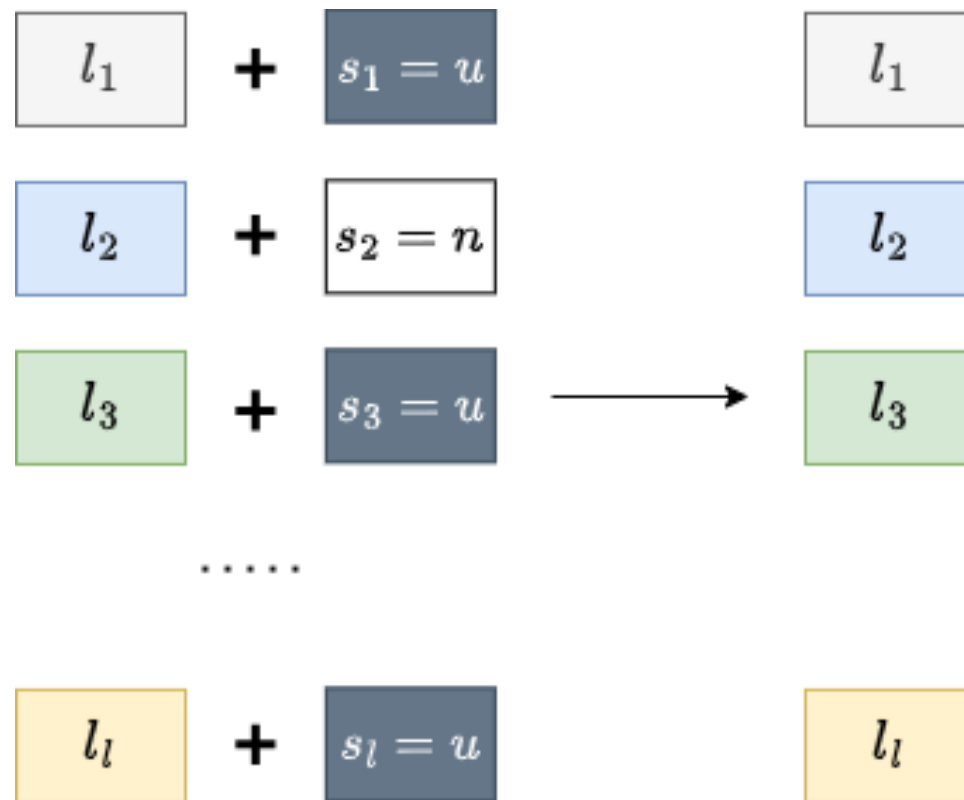
C-Tran



Thay đổi mạng trích xuất đặc trưng

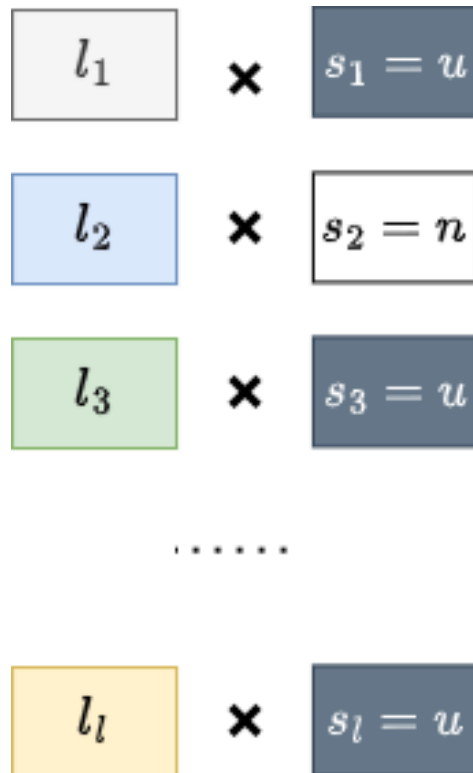


Loại bỏ thông tin trạng thái nhận.

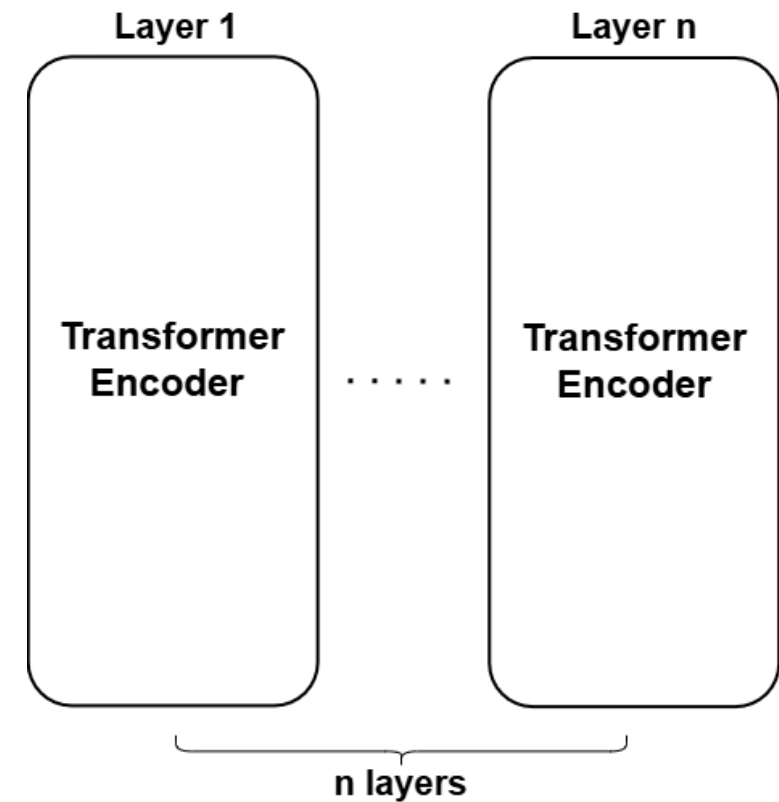


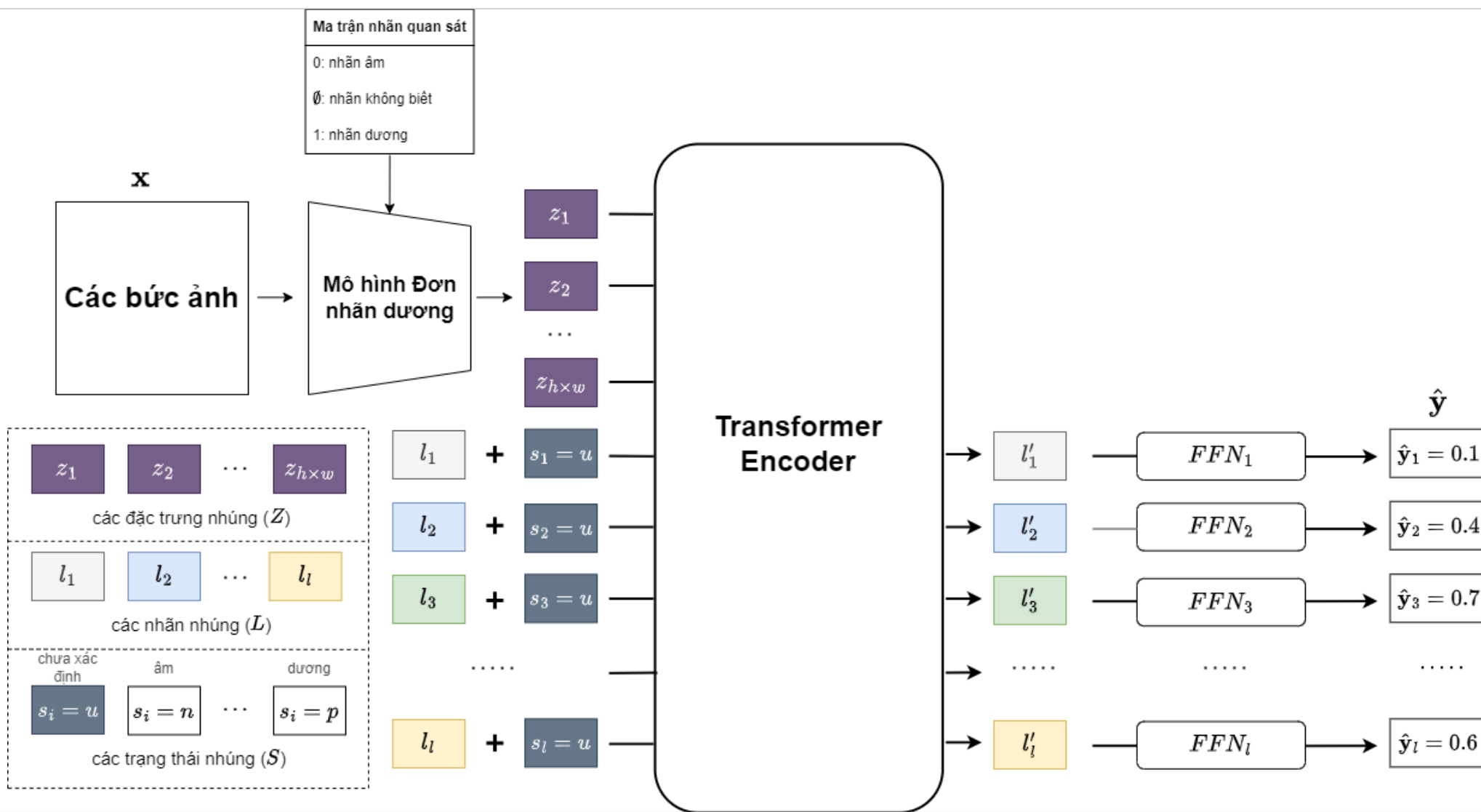


Thay đổi phương thức thêm trạng thái cho nhãn.



Thay đổi số lượng lớp Encoder.





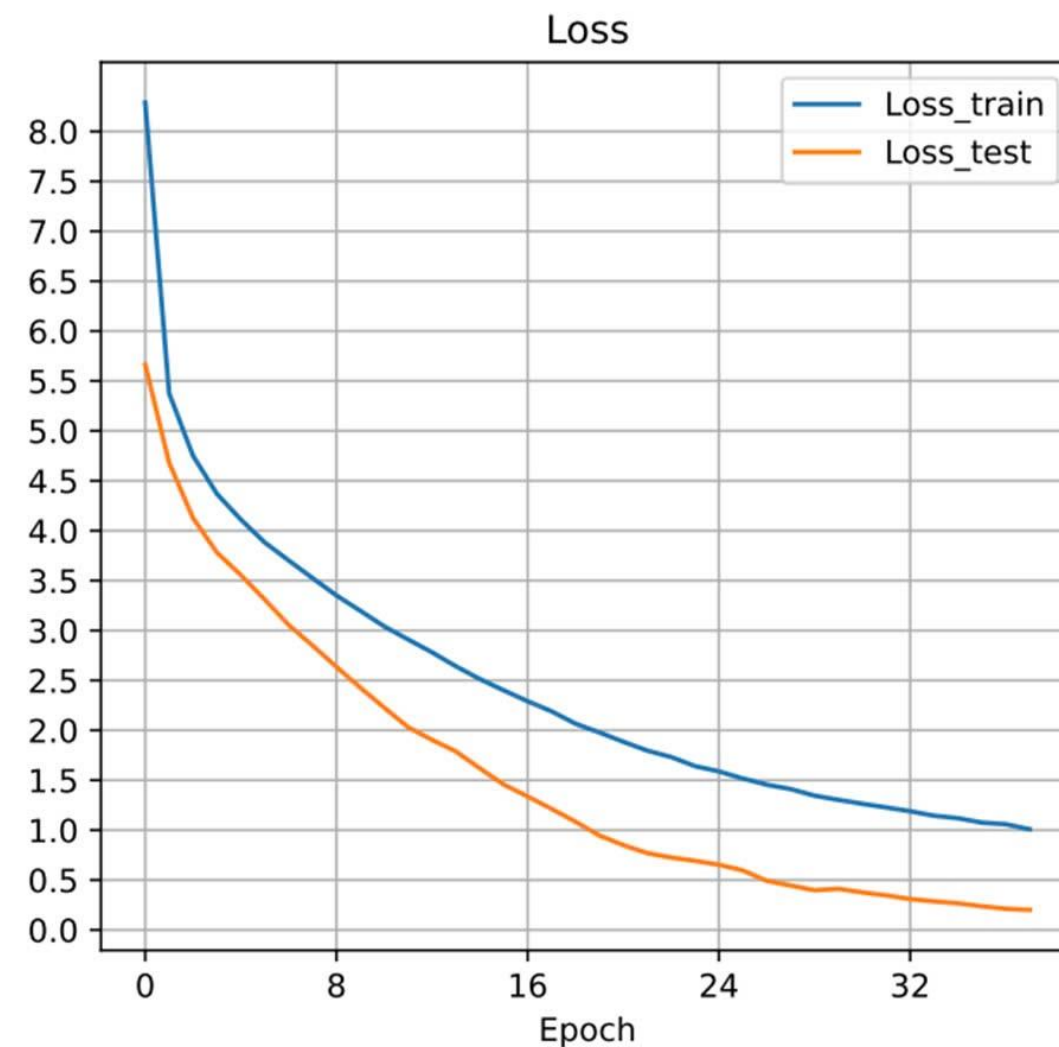
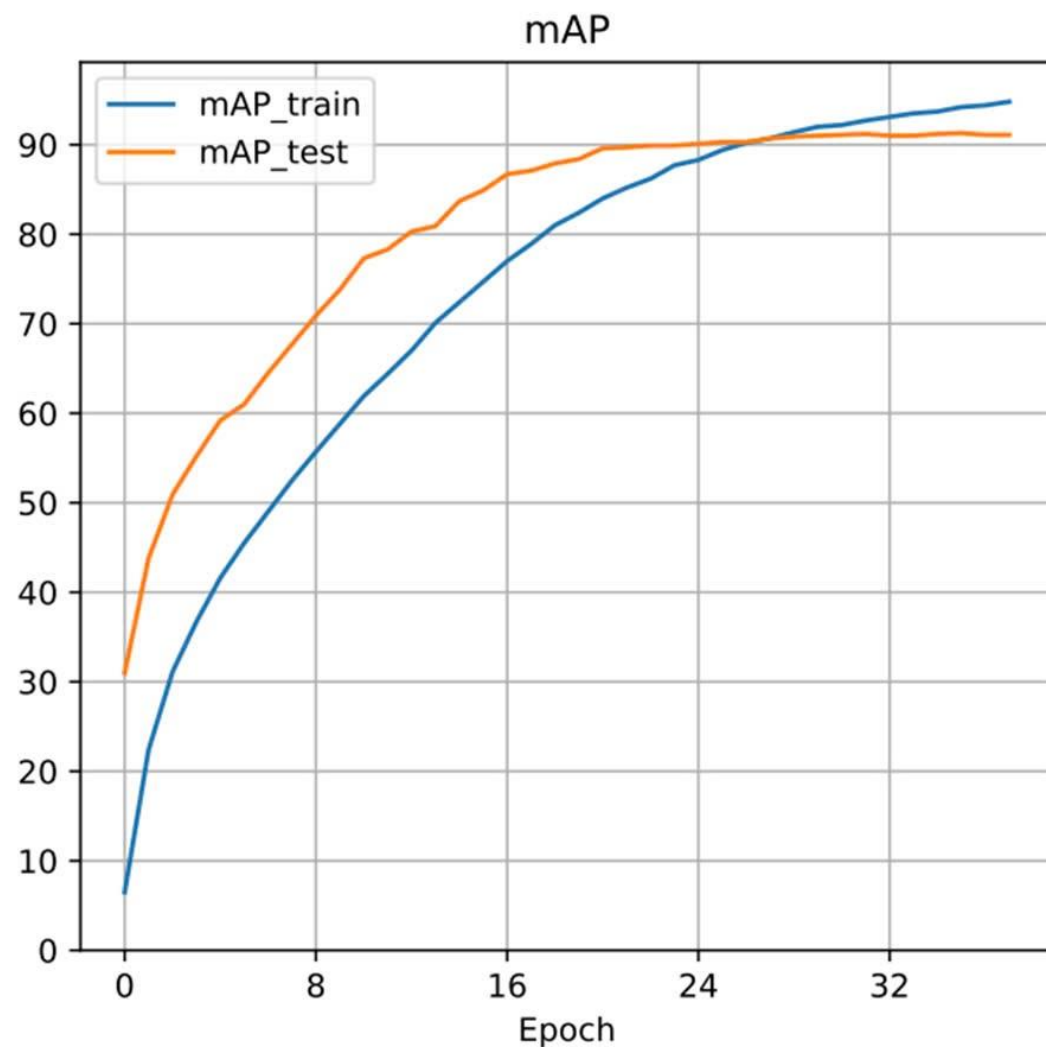


Thực Nghiệm và Đánh Giá

C-Tran

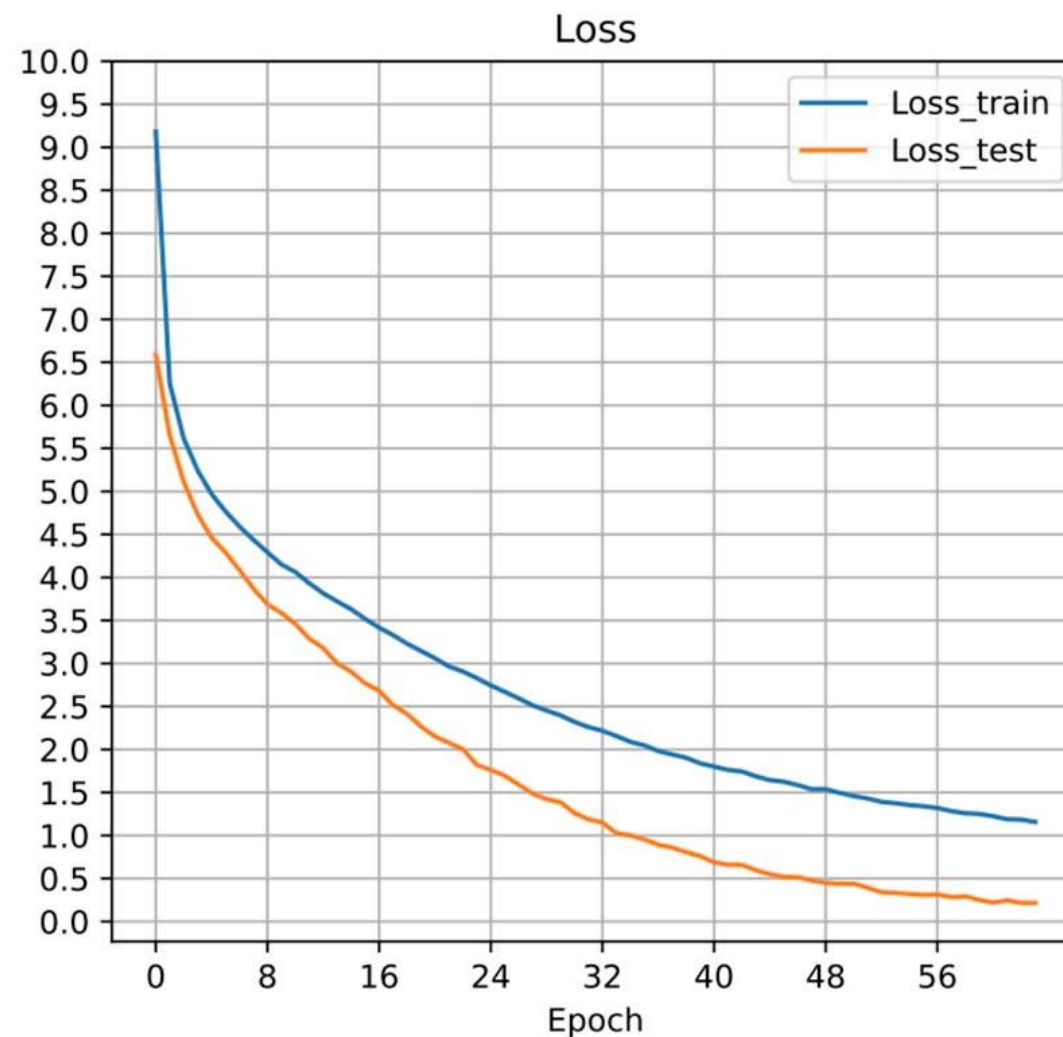
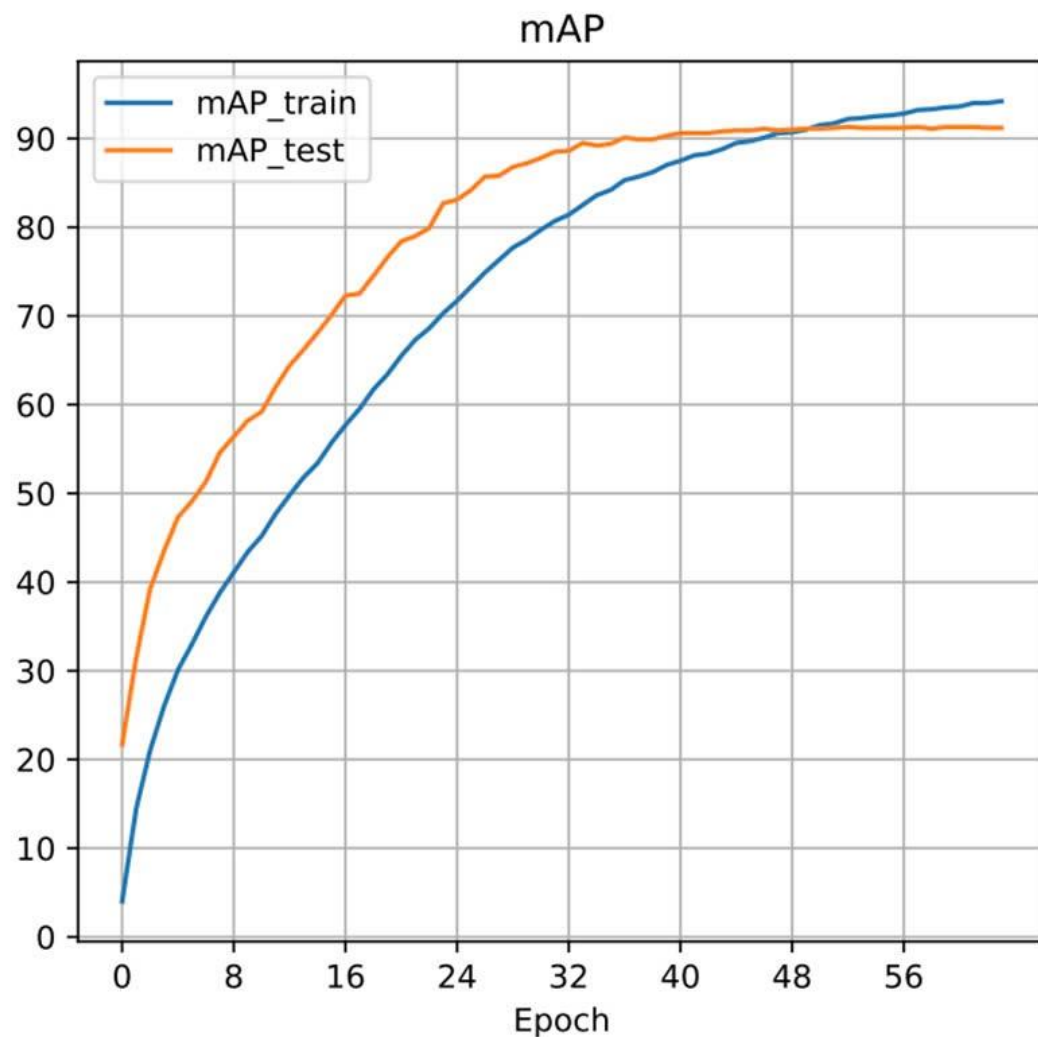


Đồ thị kết quả ResNet 101





Đồ thị kết quả MobileNet V2





Bảng kết quả



Mô hình gốc

Mô hình
đề xuất

Mạng trích xuất đặc trưng	Hàm kích hoạt	Số lớp En- coder	Che nhân huấn luyện	Lượng nhân biết trước	Trạng thái nhân	Kết quả kiểm tra
ResNet 101	sigmoid	3	có	0	tổng	91.3
ResNet 101	softmax	3	có	0	tổng	90.1
ResNet 101	sigmoid	3	có	243	tổng	91.3
EfficientNet B0	sigmoid	3	có	0	tổng	91.3
MobileNet V2	sigmoid	3	có	0	tổng	91.3
MobileNet V2	sigmoid	3	có	243	tổng	91.3
MobileNet V2	sigmoid	3	có	0	tích	90.6
MobileNet V2	softmax	3	có	0	tổng	89.8
MobileNet V2	sigmoid	4	có	0	tổng	91.3
MobileNet V2	sigmoid	2	không	0	tổng	91.3
MobileNet V2	sigmoid	2	có	0	tổng	91.3

Kết quả kiểm tra là
độ chính xác **mAP**.



KẾT LUẬN CHUNG



Kết Luận Chung

Cắt vùng dư thừa trong ảnh,
nâng cao chất lượng phân loại.

Cải tiến các mô hình và
thu được kết quả tốt với
kích thước nhỏ



Đã giải
quyết

Tối ưu vấn đề ngữ nghĩa
nhấn ở Đơn nhận dương.

Cải thiện ảnh đầu vào
có độ phân giải thấp



Chưa giải
quyết



HƯỚNG PHÁT TRIỂN



- Cắt vùng dư thừa đối với ảnh chưa có thông tin về phân vùng, khung chứa.
- Cải tiến các hàm mất mát và độ đo phù hợp với sự mất cân bằng nhãn.

XIN CẢM ƠN
QUÝ THẦY CÔ
ĐÃ LẮNG NGHE