Phân vùng ảnh dựa trên ngữ nghĩa bằng mô hình U-Net và kết hợp thêm cơ chế Attention

Lê Nguyễn
21120511
Khoa Công nghệ thông tin
Trường Đại học Khoa học tự nhiên - VNUHCM
Hồ Chí Minh, Việt Nam
0942142707
21120511@student.hcmus.edu.vn

Phan Nguyên Phương
21120312
Khoa Công nghệ thông tin
Trường Đại học Khoa học tự nhiên - VNUHCM
Hồ Chí Minh, Việt Nam
0767583397
21120312@student.hcmus.edu.vn

Vũ Minh Thư
21120143
Khoa Công nghệ thông tin
Trường Đại học Khoa học tự nhiên - VNUHCM
Hồ Chí Minh, Việt Nam
0971444020
21120143@student.hcmus.edu.vn

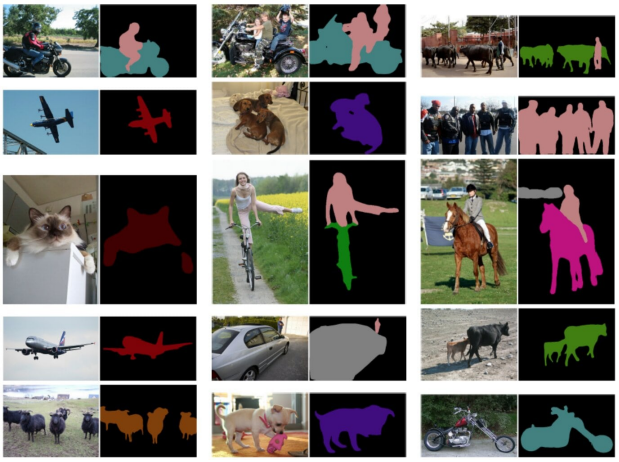
***Tóm tắt nội dung*—Trong bài báo này, nhóm tác giả sẽ sử dụng mô hình U-Net [1] và các bản cải tiến của nó bao gồm U-Net với cơ chế Attention [2]. Nhóm tác giả đánh giá điểm mạnh và yếu giữa hai mô hình U-Net khác nhau này dựa trên tập dữ liệu Cityscapes [3].**

***Từ khoá*—Học sâu, Phân vùng ảnh dựa trên ngữ nghĩa, kiến trúc U-Net, cơ chế Attention.**

I. Giới thiệu

Phân vùng ảnh dựa trên ngữ nghĩa là một lĩnh vực quan trọng trong thị giác máy tính và được ứng dụng rộng rãi trong các lĩnh vực như phân tích ảnh y khoa, hệ thống xe tự hành, giám sát video và thực tế tăng cường. Mục tiêu của bài toán này là gán nhãn ngữ nghĩa cho từng điểm ảnh, từ đó phân biệt và xác định các vật thể và đối tượng. Cụ thể, nó sẽ thực hiện việc gán một tập nhãn gồm các loại đối tượng (ví dụ: con người, xe cộ, cây cối, bầu trời, đường xá...) cho tất cả các điểm ảnh trong ảnh. Điều này giúp nội dung hình ảnh được cung cấp chi tiết hơn so với việc chỉ dự đoán một nhãn cho toàn bộ hình ảnh.

Nhiều phương pháp phân vùng ảnh dựa trên ngữ nghĩa đã được ra đời như phân ngưỡng (thresholding), nhóm dựa trên histogram (histogram-based bundling), lan vùng (region growing), phân cụm K-means (K-means clustering)... Tuy nhiên, nhóm phương pháp truyền thống này thường gặp nhiều hạn chế về độ chính xác khi đối tượng có hình dạng phức tạp hoặc ảnh có độ nhiễu cao (ví dụ: phương pháp phân ngưỡng có thể gặp khó khăn trong việc phân biệt các đối tượng có độ tương phản thấp), khả năng tổng quát hóa kém do thường phụ thuộc nhiều vào đặc điểm thủ công được thiết kế cho từng bài toán cụ thể (ví dụ: phân cụm K-means yêu cầu số lượng cụm được xác định trước) và thời gian tính toán lớn khi phải xử lý ảnh có độ phân giải cao.

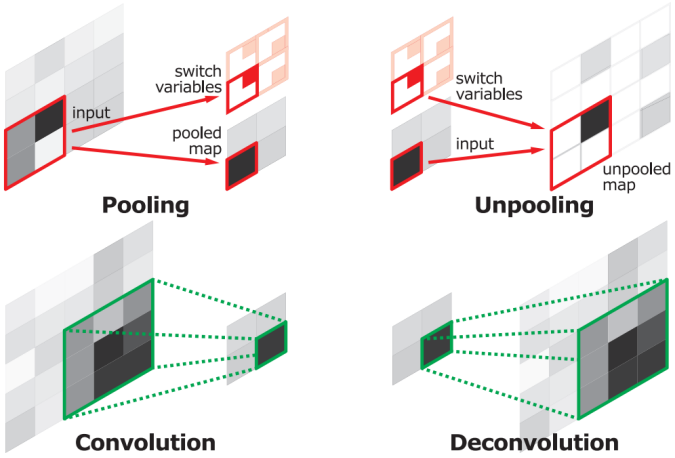


Hình 1: Kết quả phân vùng của DeepLabV3 trên ảnh mẫu [4].

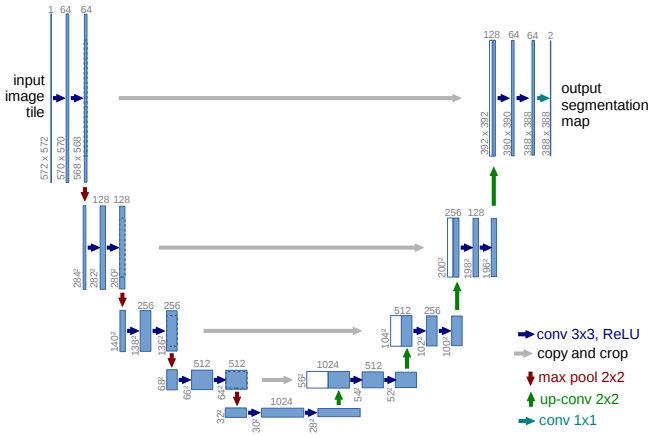
Tuy nhiên, sự ra đời của các mô hình học sâu, đặc biệt là các mạng nơ-ron tích chập (Convolutional Neural Networks - CNN), đã mang lại những cải tiến hiệu suất đáng chú ý, từ đó dẫn đến sự thay đổi mô hình trong lĩnh vực phân đoạn hình ảnh này. Ví dụ, mô hình DeepLabv3 đã đạt được độ chính xác cao và hiệu suất vượt trội so với các phương pháp truyền thống (Hình 1).

Trong bài báo cáo này, nhóm sẽ khảo sát một số phương pháp học sâu được sử dụng cho phân vùng ảnh dựa trên ngữ nghĩa sau đó chọn ra một mô hình để thực nghiệm và đánh giá kết quả.

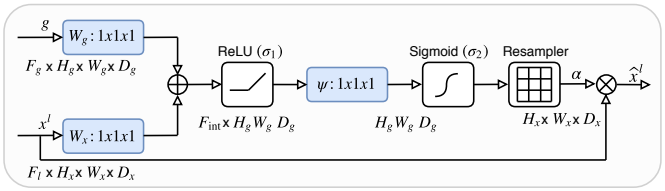
Các nội dung còn lại của báo cáo được tổ chức như sau: Phần II Các công trình liên quan trình bày một số phương pháp học sâu liên quan đến bài toán phân vùng ảnh dựa trên ngữ nghĩa đã đề cập ở trên. Phần III Phương pháp mô tả chi tiết tập dữ liệu,



Hình 2: Ví dụ về Tích chập chuyển vị (Deconvolution) và Mở gộp (Unpool) trong kiến trúc Bộ mã hoá-Bộ giải mã (hình được lấy từ [6]).



Hình 3: Mô hình U-Net gốc của Ronneberger và cộng sự [1]



Hình 4: Cổng attention trong bài báo của Oktay và cộng sự [2].

kiến trúc và các thành phần của mô hình được sử dụng. Phần IV Thực nghiệm và Kết quả trình bày các thử nghiệm và kết quả đánh giá. Phần V Tổng kết tổng kết và đưa ra một số hướng nghiên cứu trong tương lai.

II. Các công trình liên quan

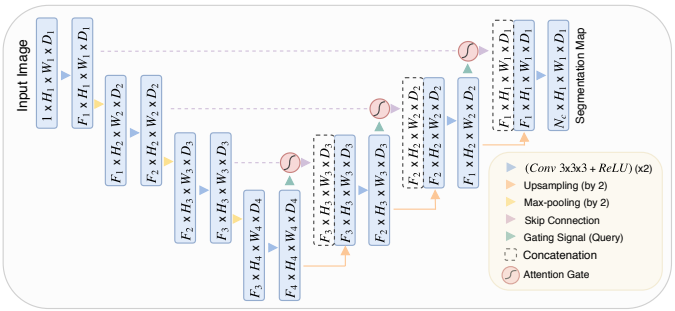
Long và cộng sự [5] đã đặt nền tảng cho việc chỉ sử dụng các lớp Tích chập trong toàn bộ Mạng nơ ron. Mạng tích chập toàn bộ (gọi tắt là FCN) đưa ra một ý tưởng quan trọng "Kết nối Ngắt quãng" (Skip Connection), trong đó ánh xạ đặc trưng (feature map) của một lớp sẽ được mở rộng mẫu (up-sampled) và sau đó kết hợp với ánh xạ đặc trưng của một lớp khác, trong FCN, việc kết hợp này chính là cộng lại giữa hai ánh xạ đặc trưng. Việc dùng kết nối ngắt quãng này sẽ kết hợp được thông tin giữa các lớp với nhau và đưa ra cách phân vùng chính xác. FCN được xem là một cột mốc quan trọng trong tác vụ phân vùng ảnh dựa trên ngữ nghĩa thế nhưng FCN vẫn còn một số hạn chế như không đủ nhanh để có thể suy diễn trong thời gian thực và không xử lý ngữ cảnh toàn cục của ảnh một cách hiệu quả.

Noh và cộng sự [7] đã đưa ra một trong những mô hình đầu tiên sử dụng cơ chế gọi là "Bộ mã hoá-Bộ giải mã" (Encoder-Decoder). Trong đó mạng sẽ được chia ra làm hai phần, phần được gọi là "Bộ mã hoá" và phần được gọi là "Bộ giải mã". Trong nghiên cứu của Noh và cộng sự [7], Bộ mã hoá là một mạng con gồm nhiều lớp tích chập và lớp gộp (pool), còn Bộ giải mã là một mạng con gồm nhiều lớp để giải mã lớp tích chập và lớp gộp ở bộ mã hoá, gọi là Lớp tích chập chuyển vị (transposed convolution hay deconvolution) và Lớp mở gộp (unpool). Để dễ hiểu hơn, có thể xem hình 2 để hiểu rõ hơn.

Nhờ nghiên cứu của Noh [7] và Long [5], Ronneberger và cộng sự [1] đã đưa ra mô hình U-Net, mô hình này kết hợp hai ý tưởng chính trên, "Kết nối ngắt quãng" và "Bộ mã hoá-Bộ giải mã". Ở kết nối ngắt quãng (mũi tên màu xám ở hình 3), mô hình U-Net nối hai ánh xạ đặc trưng lại với nhau và sau đó cho một lớp tích chập cùng với một hàm kích hoạt không tuyến tính.

Mô hình U-Net chia làm hai phần. Phần lấy mẫu xuống (down-sampling) hay được xem như bộ mã hoá (phần bên trái của hình 3) có nhiệm vụ trích xuất các đặc trưng của ảnh, phần này được cấu tạo từ các lớp tích chập có kích thước 3×3. Phần còn lại là phần mở rộng mẫu (up-sampling) hay bộ giải mã (phần bên phải của hình 3) sẽ làm giảm số lượng kênh của các ánh xạ đặc trưng, đồng thời các ánh xạ đặc trưng ở phần mở rộng được nối với ánh xạ bên phần lấy mẫu xuống (kết nối ngắt quãng), bộ giải mã được tạo thành từ nhiều tích chập chuyển vị như của [7] hoặc có thể dùng các lớp nội suy song tuyến (bilinear interpolation). Cuối cùng, một lớp tích chập có kích thước 1×1 sẽ xử lý ánh xạ đặc trưng và đưa ra một phân vùng ảnh mà phân loại từng điểm ảnh vào nhãn hợp lý của nó. Mô hình U-Net là một trong những mô hình nổi tiếng trong việc phân vùng ảnh y khoa, ngoài ra U-Net còn chiến thắng trong cuộc thi ISBI 2015. Trong bài báo này, nhóm tác giả sẽ sử dụng nó trong việc phân vùng ảnh tổng quát và xem hiệu quả của nó.

Để cải tiến U-Net, nhóm tác giả đưa ra cải tiến mà nhóm tác giả thấy phù hợp và dễ hiểu nhất. Đó là thêm cơ chế Cổng Attention (Attention Gate) vào đoạn "Kết nối ngắt quãng" của U-Net gốc. Cải tiến này được dựa trên bài báo của Oktay và cộng sự [2]. Thay vì chỉ nối hai ánh xạ đặc trưng lại với nhau như trong bài báo [1], cải tiến này sẽ kết hợp hai ánh xạ đặc trưng ấy thông qua một cổng Attention (như hình 4) sau đó mới tiến hành nối lại. Toàn bộ mô hình U-Net cộng với cải tiến trong bài báo gốc [2] ở hình 5.



Hình 5: Mô hình Attention U-Net trong bài báo của Oktay và cộng sự [2].

(a) Ảnh ví dụ (b) Nhãn tương ứng

Hình 6: Một ví dụ về ảnh và nhãn tương ứng của nó (đã được gộp lại còn 8 nhãn) trong bộ dữ liệu Cityscapes [3].

III. Phương pháp

A. Tập dữ liệu

Nhóm tác giả sử dụng bộ dữ liệu Cityscapes [3] cho việc huấn luyện, thử nghiệm và đánh giá. Ngoài ra, bộ dữ liệu này còn được dùng để so sánh độ hiệu quả của ba mô hình U-Net khác nhau trên một tác vụ Phân vùng ảnh ngữ nghĩa. Bộ dữ liệu ban đầu sẽ bao gồm 2975 ảnh cho tập huấn luyện, 500 ảnh cho tập thẩm định với nhãn được công khai và 1525 ảnh thử nghiệm với nhãn được giấu đi. Tập dữ liệu ban đầu sẽ gồm 34 nhãn được dùng cho tác vụ phân vùng ảnh dựa trên đối tượng, để làm phù hợp cho tác vụ phân vùng ảnh dựa trên ngữ nghĩa, nhóm tác giả đã gộp lại còn 8 nhãn tất cả. Một ví dụ về bộ dữ liệu này được đưa ra ở hình 6. Do giới hạn về phần cứng, nhóm tác giả đã thay đổi kích thước ảnh từ 1024×2048 xuống còn 224×224. Để tăng số lượng dữ liệu huấn luyện cho tập dữ liệu ban đầu, nhóm tác giả sử dụng thêm các ảnh mới được tạo ra từ các ảnh ban đầu, các ảnh mới này được tạo ra bằng cách xoay ảnh, lật ảnh, hoặc thay đổi màu ảnh ban đầu và cuối cùng nhóm tác giả được 8195 ảnh tất cả cho tập huấn luyện.

B. Mô hình U-Net

Mô hình đầu tiên mà nhóm tác giả chọn để thực hiện tác vụ này là mô hình U-Net [1]. Mô hình được trực quan hoá trong hình 3. Vẫn giữ nguyên so với bài báo gốc, nhóm tác giả chọn số lượng bộ lọc của lớp tích chập ban đầu là 64. Đầu tiên ở lớp Tích chập với nhân kích thước 3×3 (Conv 3x3 trong hình 3), nhóm tác giả chọn tham số padding là 1 thay vì 0 như trong bài báo gốc, điều này giúp tránh trường hợp ảnh có kích thước lẻ khi đi qua lớp Gộp Tối đa với nhân kích thước 2×2 (Max pool 2x2 trong hình 7) cũng như là làm cho ánh xạ đặc trưng đầu ra có cùng kích thước với ảnh đầu vào. Tiếp theo, nhóm tác giả chọn hàm kích hoạt là Hàm chỉnh lưu rò rỉ (LeakyReLU) [8] thay vì Hàm chỉnh lưu thông thường (ReLU) như trong bài báo gốc, bởi vì sau nhiêu lần thử nghiệm nhóm tác giả thấy Hàm chỉnh lưu rò rỉ phù hợp hơn với dữ liệu và cho kết quả tốt hơn. Có thể thấy ở mô hình U-Net (như hình 7) chia làm hai phần, một phần đi xuống và nhóm tác giả sẽ gọi là Bộ mã hoá, một phần đi lên và nhóm tác giả sẽ gọi là Bộ giải mã, đoạn nối giữa hai bộ này sẽ được nhóm tác giả gọi là Cầu. Để cho quá trình huấn luyện nhanh và ổn định hơn, ở mỗi lớp Tích chập, trước khi qua hàm kích hoạt ReLU thì sẽ đi qua một lớp Chuẩn hoá Hàng loạt (BatchNorm trong hình 7) [9]. Cuối cùng, ở đoạn lấy mẫu lên, nhóm tác giả chọn cách dùng nội suy song tuyến (như trong hình 7).

C. Kiến trúc U-Net kèm cơ chế Attention

Để làm đơn giản đi cổng Attention, nhóm tác giả đã bỏ đi Bộ lấy mẫu lại (Resampler) trong bài báo gốc [2] (hoặc trong hình 4). Mô hình U-Net kèm cơ chế Attention được nhóm tác giả đơn giản hoá và trực quan hoá ở hình 9. Ở cổng Attention, hình hộp màu xanh là một lớp tích chập có kích thước 1×1 với bộ tham số của nó (ví dụ như bộ tham số W_g ở hình 8). Khi ánh xạ đặc trưng ở lớp trước đó, gọi là x và ánh xạ đặc trưng ở lớp hiện tại sau khi được lấy mẫu lên, gọi là g cùng đi qua cổng Attention, đầu tiên chúng sẽ đi qua lớp tích chập của tương ứng với mình W_x cho x và W_g cho g, riêng lớp tích chập của g sẽ có bias cộng thêm, gọi là b_g. Sau khi cả hai đã đi qua lớp tích chập của mình, chúng sẽ được cộng lại với nhau và đi qua một hàm kích hoạt σ_1 (ở đây là hàm ReLU), gọi kết quả của đoạn này là q, ta có:

$$q = \sigma_1 \left(W_x^T x + W_g^T g + b_g \right)$$

Hàm kích hoạt σ_1 (ReLU)

Tham số của lớp tích chập mà x đi qua

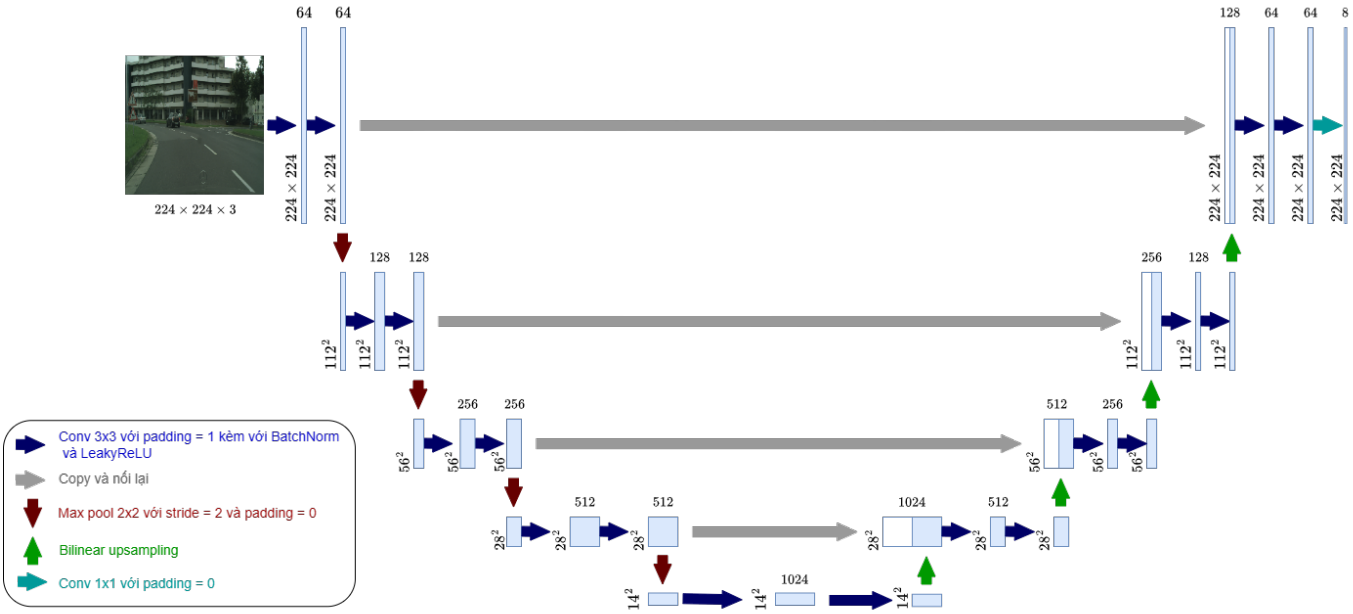
Tham số của lớp tích chập mà g đi qua

Sau khi có được q, ta cho q đi qua một lớp tích chập nữa có kích thước 1×1 và có bộ tham số là ψ, tương tự W_g, lớp này cũng có bias và ta gọi bias đó là b_ψ. Tiếp theo đó cho qua hàm kích hoạt σ_2 (ở đây là hàm Sigmoid), sau đó ta lấy giá trị đã kích hoạt này nhân với từng phần tử của x ban đầu. Cuối cùng, ta được giá trị đầu ra của cổng Attention, gọi là x':

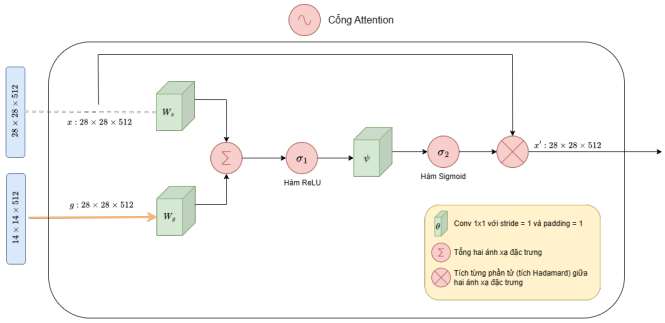
$$x' = \sigma_2 \left(\psi^T q + b_\psi \right) \odot x$$

Hàm kích hoạt σ_2 (Sigmoid)

Tích Hadamard hay tích của từng phần tử với nhau



Hình 7: Mô hình U-Net của nhóm tác giả (đã được thay đổi một chút so với bản gốc [1] và được sử dụng trên ảnh màu RGB có kích thước 224×224). Mỗi hình chữ nhật màu xanh tương ứng với một ánh xạ đặc trưng nhiều kênh. Số kênh của ánh xạ ấy được kí hiệu phía trên hình chữ nhật. Hình chữ nhật màu trắng tương ứng với một bản sao của hình chữ nhật màu xanh phía sau mũi tên.



Hình 8: Cổng Attention của nhóm tác giả (đã thay đổi một chút so với bản gốc [2]).

Khi có được giá trị đầu ra x', ta sẽ tiến hành nối lại với ánh xạ đặc trưng của lớp hiện tại tương tự như trong bài báo U-Net gốc [1] (trực quan hoá ở hình 9).

D. Huấn luyện

Cả ba mô hình đều sử dụng chung một hàm mất mát là hàm Entropy Chéo (Cross Entropy). Sau đó hàm mất mát này sẽ được tối ưu bằng thuật toán AdamW [10] trong đó tỉ lệ học (learning rate) được thiết lập là 0.01 và độ phân rã trọng số (weight decay) được thiết lập là 1×10^{-5}. Ngoài ra mỗi mô hình được huấn luyện trên 25 epoch (do giới hạn về phần cứng) và có kích thước batch là 32. Để đánh giá được độ tốt của mô hình, nhóm tác giả sử dụng độ đo MeanIoU, với IoU được chọn là độ đo Jaccard (hay chỉ số Jaccard) cho nhiều lớp [11]. Ngoài ra, việc khởi tạo giá trị ban đầu cho các trọng số cũng cực kì quan trọng, trong nghiên cứu này, nhóm tác giả chọn khởi tạo các giá trị ấy từ một phân phối chuẩn có trung bình là 0 và phương sai là $2/(N \times M)$ với N là kích thước hiện tại của lớp Tích chập và M là số kênh đầu ra của lớp Tích chập ấy. Ví dụ, một lớp Tích chập có kích thước 3×3 và số kênh đầu ra là 64 thì $N = 9$ và $M = 64$, riêng giá trị bias sẽ được khởi tạo giá trị ban đầu là 0.

IV. Thực nghiệm và Kết quả

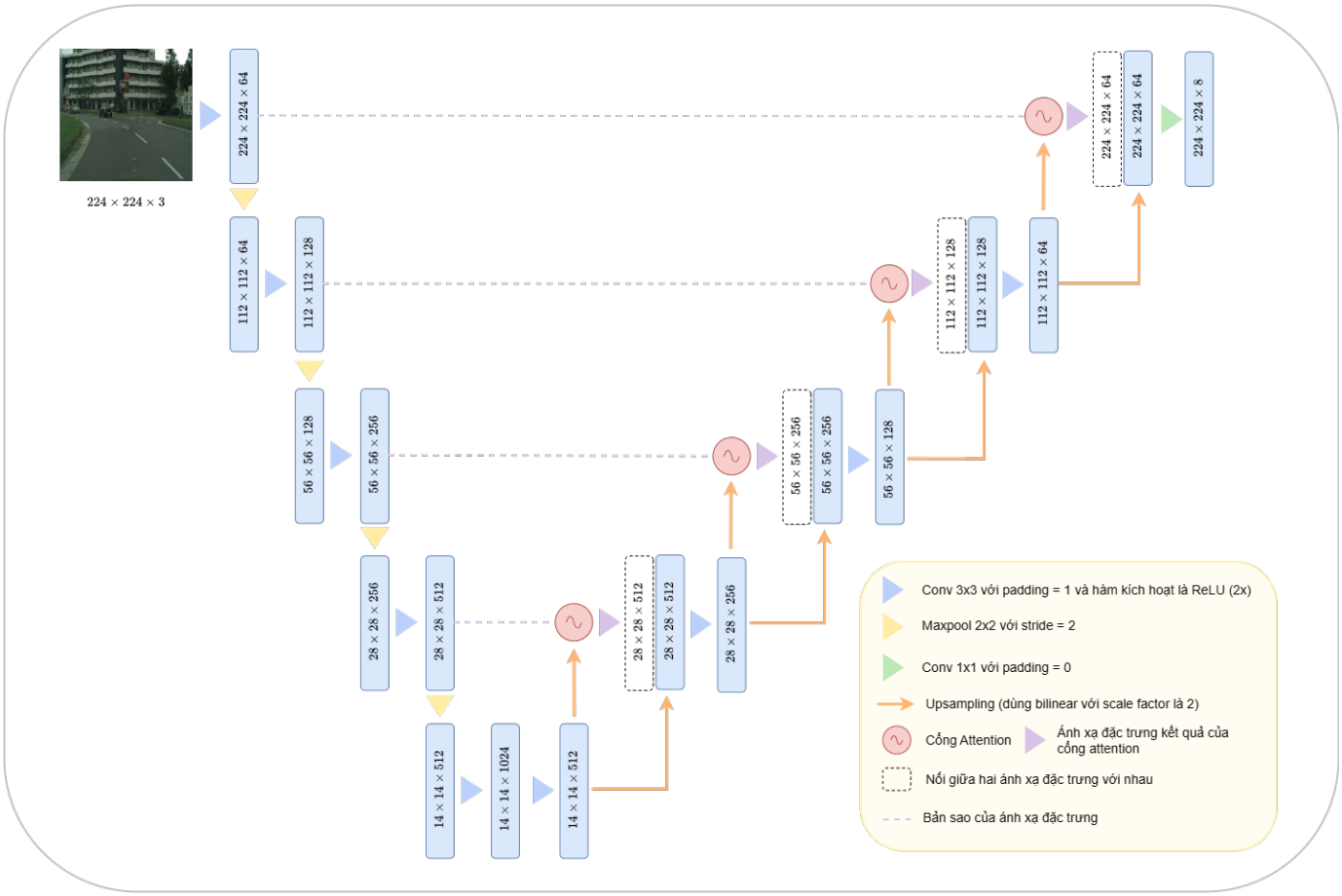
Bảng I: So sánh giữa hai mô hình U-Net và Attention U-Net của nhóm tác giả bằng MeanIoU cao nhất và thời gian huấn luyện

Mô hình	MeanIoU cao nhất	Thời gian huấn luyện
U-Net (hình 7)	0.714	2 giờ
Attention U-Net (hình 9)	0.74	2.5 giờ

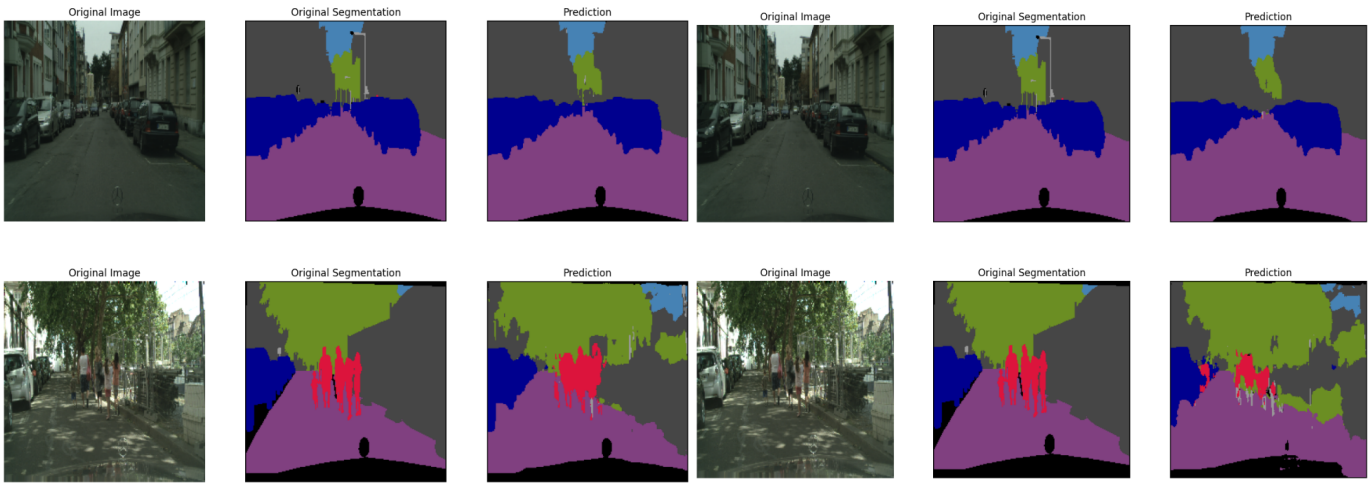
Có thể thấy, việc sử dụng thêm cổng Attention cho phần Kết nối ngắt quãng đã cho kết quả tốt hơn rất nhiều khi MeanIoU cao nhất tăng thêm gần 0.03, thế nhưng thời gian huấn luyện sẽ tăng lên thêm 0.5 giờ (theo bảng I), ngoài ra theo hình 10 thì dự đoán của Attention U-Net phần nào đó chuẩn xác và gần với nhãn gốc hơn so với U-Net.

V. Tổng kết

Có thể thấy nhóm tác giả đã áp dụng thành công hai mô hình U-Net [1] và cải tiến của nó là Attention U-Net [2] vào bài toán phân vùng ảnh bằng ngữ nghĩa trên tập dữ liệu Cityscapes [3], kết quả cho ra ngoài sức mong đợi của nhóm tác giả, việc áp dụng tuân thủ các quy tắc của bài báo gốc và thêm sự thay đổi của nhóm tác giả đã góp phần không nhỏ cho kết quả này. Thế nhưng, trong tương lai, nhóm tác giả sẽ sử dụng các mô hình khác mạnh mẽ và hiệu quả hơn ví dụ như U-Net++ [12], đây là



Hình 9: Mô hình U-Net kèm cơ chế Attention của nhóm tác giả (đã được thay một chút so với bản gốc [2] và được sử dụng trên ảnh màu RGB có kích thước 224×224).



(a) Dự đoán của Attention U-Net

(b) Dự đoán của U-Net

Hình 10: Dự đoán của hai mô hình U-Net và Attention U-Net. Original Segmentation là nhãn còn Prediction là dự đoán của mô hình.

mô hình mà nhóm tác giả đã xem xét sử dụng, nhưng vì kiến trúc phức tạp của nó nên nhóm tác giả đã không chọn, ngoài các mô hình dựa vào U-Net, các mô hình khác mới và hiện đại hơn cũng có thể được xem xét như SegFormer [13] hay VLTSeg [14].

TÀI LIỆU

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[2] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *CoRR*, vol. abs/1804.03999, 2018.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016.

[4] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *CoRR*, vol. abs/2001.05566, 2020.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[6] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *CoRR*, vol. abs/1711.10684, 2017.

[7] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," *CoRR*, vol. abs/1505.04366, 2015.

[8] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *CoRR*, vol. abs/1505.00853, 2015.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[10] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.

[11] L. Li, Y. Wu, and M. Ye, "Experimental comparisons of multi-class classifiers," *Informatica*, vol. 39, no. 1, 2015.

[12] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *CoRR*, vol. abs/1807.10165, 2018.

[13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 12077–12090, Curran Associates, Inc., 2021.

[14] C. Hümmer, M. Schwonberg, L. Zhou, H. Cao, A. Knoll, and H. Gottschalk, "Vltseg: Simple transfer of clip-based vision-language representations for domain generalized semantic segmentation," 2023.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.