TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
o0o

Tên báo cáo

BÁO CÁO
PROJECT 3

Giáo viên hướng dẫn: Nguyễn Khánh Phương
Sinh viên thực hiện : Vũ Long Giang - 20183519
Lớp : Khoa học máy tính 04

Hà Nội - 2022

Mục lục

Mở đầu

Với sự phát triển của khoa học kĩ thuật nói chung, ngành AI nói riêng, các ứng dụng áp dụng trí tuệ nhân tạo ngày càng nhiều và đi sâu vào đời sống của con người. Trong trí tuệ nhân tạo, thị giác máy tính và xử lý ngôn ngữ tự nhiên là 2 chuyên ngành nhỏ nhận được rất nhiều sự quan tâm của các nhà nghiên cứu cũng như các nhà phát triển dịch vụ. Mặc dù các bài toán trong thị giác máy tính và xử lý ngôn ngữ tự nhiên đã được đưa ra nghiên cứu và đạt được nhiều thành tựu đáng kể, lớp các bài toán kết nối giữa hai lĩnh vực này (cross-modal) còn rất mới và đang trong quá trình phát triển mạnh. Điển hình trong lớp bài toán này có thể kể đến các bài toán về Visual Question Answering (VQA), Image Text Retrieval (ITR), Image Captioning (IC). Bài toán IC có nhiều ứng dụng trong thực tế:

- Hỗ trợ người khuyết tật, người mắt kém hiểu được hình ảnh xung quanh.
- Hỗ trợ các hệ thống tìm kiếm ảnh theo nội dung cho trước hoặc tìm kiếm thông tin, sự kiện theo ảnh cho trước.

Trong những năm gần đây, IC là bài toán thu hút được nhiều sự chú ý của nhiều nhà khoa học cũng như các trung tâm nghiên cứu lớn trên thế giới. Đáng kể nhất là sự ra đời của mô hình OSCAR [6] và OSCAR+ [14] của các nhà khoa học tại Microsoft. OSCAR và OSCAR+ thể hiện hiệu năng vượt trội so với các mô hình trước đó ở nhiều bài toán cross-model, và đang là thuật toán cho kết quả tốt nhất đối với tác vụ Image Captioning (SOTA). Ý tưởng chung của OSCAR và OSCAR+ khác với các mô hình trước đó là đưa thêm các đối tượng trong được phát hiện trong ảnh thành thuộc tính để sinh tiêu đề, thay vì chỉ dùng các thuộc tính vùng (region feature) trong ảnh.

Báo cáo này tập trung vào việc đưa ra tổng quan, quá trình phát triển và các hướng tiếp cận cho bài toán, cài đặt hai mô hình thuật toán là Show, Attend and Tell [9] và finetuned trên mô hình OSCAR+ [14], xây dựng kịch bản thực nghiệm để đưa ra đánh giá.

Chương 1

Đặt vấn đề

1.1 Đặt vấn đề

Các bài toán kết hợp giữa xử lý ảnh và ngôn ngữ tự nhiên đang tiếp tục nhận được sự quan tâm của các nhà nghiên cứu. Lớp bài toán này yêu cầu cả các kĩ thuật trong ảnh cũng như trong ngôn ngữ để sinh ra được kết quả tốt. Cụ thể đối với IC, từ những ý tưởng đơn thuần đầu tiên như sinh và điền mẫu để tạo ra tiêu đề hay áp dụng những kĩ thuật sinh ngôn ngữ thủ công [1][11][12] cho đến ngày nay, rất nhiều kĩ thuật phức tạp đã được áp dụng để đạt hiệu quả cao hơn trong chất lượng của tiêu đề [2][4][13][10][6][14], trong đó cơ chế Attention luôn là ý tưởng chủ đạo. Báo cáo này sẽ đưa ra những nhận xét cụ thể về từng hướng tiếp cận và thực hiện đánh giá trên 2 mô hình cụ thể. Cụ thể, những công việc được thực hiện:

- Thống kê, đánh giá các hướng tiếp cận cho bài toán IC.
- Xử lý một bộ dữ liệu nhỏ phục vụ quá trình thực nghiệm.
- Xây dựng mô hình IC dựa trên mô hình Show, Attend and Tell [9]
- Xây dựng mô hình IC dựa trên mô hình pretrained OSCAR+ (VINVL) [14].
- Đề xuất cải tiến.
- Đánh giá kết quả trên các độ đo BLEU, CIDEr, SPICE.

1.2 Các hướng tiếp cận

Chương 2

Các nghiên cứu liên quan

Trong mục này, các nghiên cứu hiện đại (với hướng tiếp cận của mô hình seq2seq) sẽ được trình bày chi tiết, phục vụ quá trình cài đặt và cải tiến phía sau.

2.1 Show, Attend and Tell

Giống với một vài mô hình được đề xuất trước đó, Show, Attend and Tell [9] sử dụng kiến trúc mô hình dạng seq2seq, mạng *Encoder* dùng để mã hóa ảnh đầu vào và *Decoder* dùng để giải mã tạo thành một tiêu đề hoàn chỉnh. Tuy nhiên, điểm khác biệt của Show, Attend and Tell với các mô hình trước đó là việc sử dụng cơ chế Attention để nâng cao chất lượng của quá trình mã hóa. Kiến trúc tổng quan của mô hình được thể hiện ở hình 2.1. Mô hình đề xuất nhận đâu vào là một ảnh I và đầu ra là một tiêu đề y. Ảnh đầu vào được mã hóa bằng một mạng nơ-ron tích chập tạo thành tập hợp các thuộc tính vùng a, mạng này mã hóa I thành L vecto D chiều. Cụ thể, y và a được biểu diễn như sau:

$$y = \{y_1, ..., y_C\}, y_i \in \mathbb{R}^K$$

$$a = \{a_1, ..., a_L\}, a_i \in \mathbb{R}^D$$

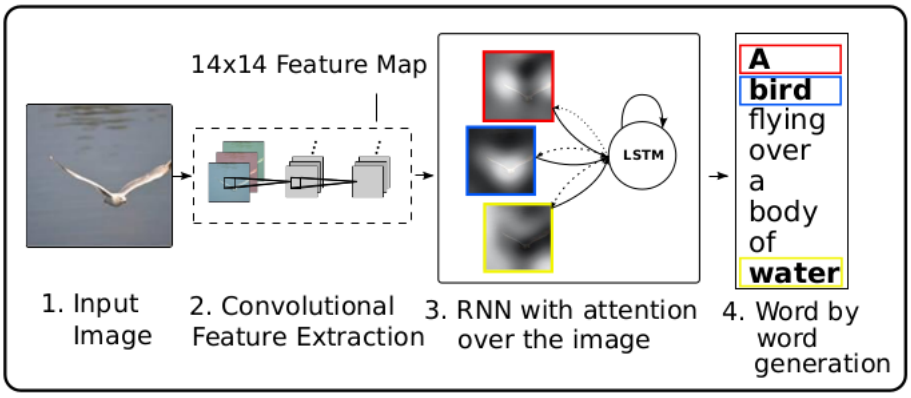
Cơ chế Attention được áp dụng trên tập các thuộc tính vùng:

$$\alpha_{ti} = softmax(f_{att}(a_i, h_{t-1})) \tag{2.1}$$

Trong đó, α_{ti} là trọng số của vùng i tại thời điểm sinh từ thứ t, f_{att} là một mô hình attention, h_{t-1} là *hidden state* của *Decoder* sau khi sinh từ thứ *t-1*. Trong [9], tác giả đề xuất 2 cách sử dụng đối với α:

- Sử dụng α_{ti} như xác suất vùng i là vùng được dùng để sinh từ thứ t. Tại mỗi bước sinh từ, vecto ngữ cảnh $z_t = a_i$ với α_{ti} là lớn nhất. Cơ chế này gọi là *"hard" attention.*
- Sử dụng α_{ti} như là trọng số của vùng i tại thời điểm sinh từ thứ t. Tại mỗi bước sinh từ, vecto ngữ cảnh $z_t = \sum_{i=1}^{L} \alpha_{ti} a_i$. Cơ chế này gọi là *"soft" attention.*

Trong báo cáo này, mô hình Show, Attend and Tell được cài đặt với *"soft" attention. Encoder* được xây dựng trên kiến trúc LSTM với đầu vào tại mỗi bước sinh là vecto ngữ cảnh z_t, *hidden state* h_{t-1} và embedding của từ thứ *t-1* Ey_{t-1}.



Hình 2.1: Kiến trúc mô hình Show, Attend and Tell

Hàm mục tiêu được sử dụng có dạng:

$$L_d = -\lg(P(y|x)) + \lambda \sum_{i}^{L}(1 - \sum_{1}^{C} \alpha_{ti})^2 \tag{2.2}$$

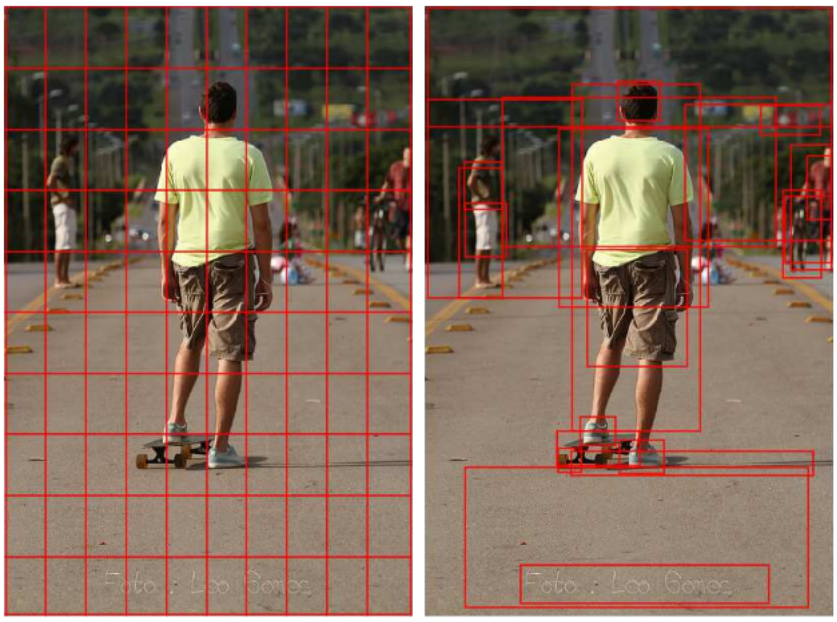
Trong đó thành phần $\lambda \sum_{i}^{L}(1 - \sum_{1}^{C} \alpha_{ti})^2$ được sử dụng để cố gắng đưa $\sum_t \alpha_{ti} \approx 1$, nghĩa là đưa độ quan trọng của các vùng trong ảnh trong cả quá trình sinh tiêu đề bằng nhau.

Nhận xét: Mô hình được đề xuất có sự cải thiện với việc sử dụng cơ chế Attention, tuy nhiên việc áp dụng ở mức cơ bản nhất, một mô hình Attention được áp dụng hỗ trợ quá trình mã hóa ảnh. Mô hình này mở ra hướng áp dụng cơ chế Attention tiềm năng cho bài toán IC.

2.2 BUTD

Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering [2] đề xuất một mô hình giải quyết hai bài toán IC và VQA dựa trên Attention. Ý tưởng và điểm khác biệt của bài bào so với các mô hình trước đó là cải tiến quá trình trích xuất thuộc tính của ảnh. Đầu tiên, *"bottom-up"* attention được sử dụng để lấy ra các vùng trong ảnh, đây là điểm tạo nên sự cải thiện rõ rệt so với các mô hình đã có, sử dụng cơ chế này giúp mô hình lấy ra được các vùng có ý nghĩa hơn trong ảnh, thay vì lấy các vùng bằng nhau với việc sử dụng mạng tích chập đơn thuần (hình ảnh trực quan được thể hiện ở ảnh 2.2). Tiếp sau đó, *"top-down"* attention được sử dụng để tính các trọng số cho các vùng vừa được lấy ra, từ đó đưa ra biểu diễn cho ảnh và thực hiện giải mã. Chi tiết về *"bottom-up"* và *"top-down"* attention được trình bày tiếp theo.

***"Bottom-up"* attention:** Trong mô hình được đề xuất, tác giả sử dụng Faster R-CNN [7] như một mô hình *"hard"* attention. Mô hình Faster R-CNN xác định các đối tượng và các vùng chứa đối tượng đó trong một ảnh, như vậy chỉ các vùng chứa đối tượng mới tham gia vào quá biểu diễn ảnh trong không gian mã hóa, giống với cơ chế *"hard"* attention đã trình bày ở trên. Với mỗi vùng đối tượng được xác định, vecto mean-pooled từ lớp tích chập cuối của mạng được sử dụng như là vecto thuộc tính cho vùng đó. Mô hình được thực nghiệm trong bài sử dụng Faster R-CNN pretrained trên bộ dữ liệu ImageNet[8] và tiếp tục train trên bộ dữ liệu Visual Genome[5]. Ví dụ về kết quả của *"bottom-up"* attention được biểu diễn ở hình 2.3a.



Hình 2.2: Các mô hình cũ chia ảnh thành các vùng bằng nhau để lấy ra các thuộc tính (bên trái), điều này làm cho mô hình không hiểu hết được nội dung của ảnh, phương pháp đề xuất (bên phải) dựa trên attention để lấy được thông tin có ý nghĩa hơn.

***"Top-down"* attention:** Cơ bản giống với cơ chế *"soft"* attention trong [9]. Trong mô hình đề xuất, tác giả sử dụng 2 lớp LSTM. Trong đó, lớp thứ nhất được coi là *"top-down"* attention LSTM và lớp thứ hai là Language LSTM (hình 2.3b).

- *"Top-down"* attention LSTM: Đầu vào x_t được tính toán như sau:

$$x_t^1 = [h_{t-1}^2, \bar{v}, W_e \Pi_t] \tag{2.3}$$

$$\bar{v} = \frac{1}{k} \sum_{i}^{k} v_i \tag{2.4}$$

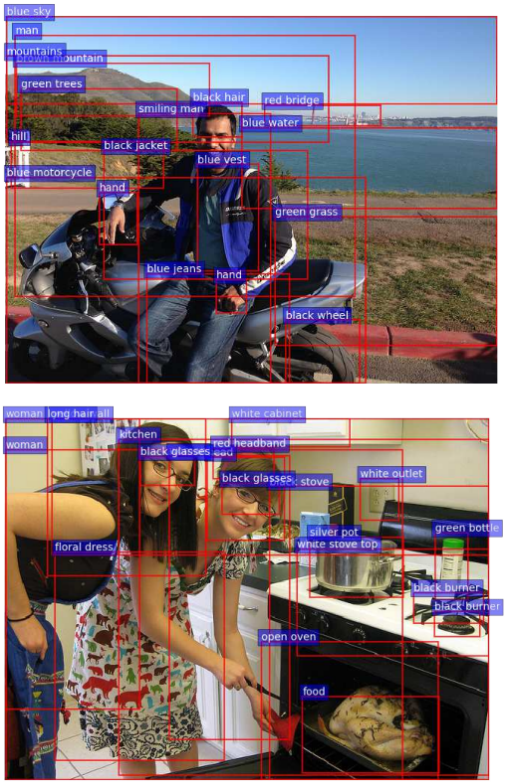
  Trong đó,h_{t-1} là *hidden state* của LSTM tại bước *t-1* ,v_i là các thuộc tính vùng của ảnh, W_e là ma trận embedding, Π là vecto one-hot của từ đầu vào tại bước *t*. *Hidden state* của lớp LSTM này là đầu vào cho một mạng Attend, từ đó tạo ra vecto ngữ cảnh $\hat{v}_t$ là đầu vào của lớp Language LSTM.

- Language LSTM: Đầu vào lớp này là *"hidden state"* từ lớp *"Top-down"* attention LSTM h_t^1, vecto ngữ cảnh $\hat{v}_t$ và *"hidden state"* từ bước sinh liền trước h_{t-1}^2.

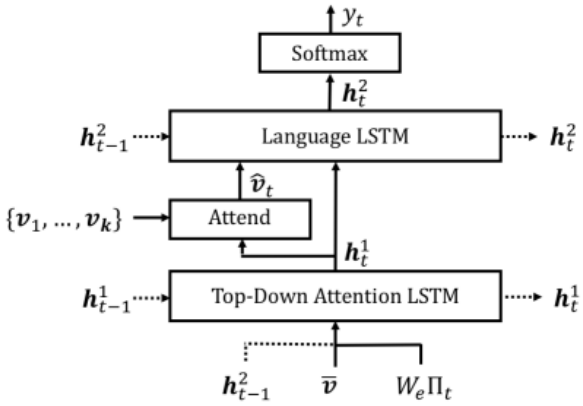
Hàm mục tiêu được sử dụng:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \lg(p_\theta(y_t^* | y_{1:t-1}^*))$$

Nhận xét:Mô hình được đề xuất tạo ra sự khác biệt nhờ vào các thuộc tính vùng. Việc sử dụng các đối tượng phát hiện được trong ảnh làm thuộc tính cũng là ý tưởng của nhiều mô hình sau này, điển hình là OSCAR [6] và OSCAR+[14]. Đây là bài báo chìa khóa cho rất nhiều nghiên cứu phía sau.



(a) Ví dụ về kết quả của Faster R-CNN



(b) Kiến trúc mạng *Decoder*

Hình 2.3: BUTD

2.3 AoANet

2.4 OSCAR and OSCAR+

Chương 3

Đề xuất cải tiến

3.1 Adversarial Loss

Chương 4

Thực nghiệm

4.1 Bộ dữ liệu

Để thực hiện đánh giá mới kết quả, bộ dữ liệu nhỏ được trích từ bộ dữ liệu lớn VizWiz [3] được sử dụng. VizWiz là bộ dữ liệu về ảnh do người mù ghi lại, được xây dựng nhằm mục đích đánh giá các thuật toán cross-modal, hỗ trợ người khuyết tật trong việc nhận diện thế giới xung quanh. Do điều kiện về hạ tầng, sử dụng bộ dữ liệu lớn VizWiz là không khả thi, một bộ dữ liệu nhỏ gồm 10000 ảnh được lấy ra, tương ứng với mỗi ảnh là 5 tiêu đề, ví dụ về bộ dữ liệu có thể nhìn ở ảnh 4.1.

Hình 4.1: Ví dụ về ảnh và tiêu đề trong bộ dữ liệu VizWiz nhỏ

Bộ dữ liệu được chia thành 3 tập train bao gồm 7000 ảnh, tập val gồm 1000 ảnh và tập test bao gồm 2000 ảnh.

4.2 Cấu hình và kịch bản thực nghiệm

4.2.1 Phần cứng

Google colab pro
CPU Intel Xenon GPU Tesla T40 RAM

4.2.2 Độ đo

4.2.3 Show, Attend and Tell

4.2.4 Finetuned OSCAR+

4.3 Kết quả

Chương 5

Tổng kết

Tài liệu tham khảo

[1] Ahmet Aker and Robert Gaizauskas, *Generating image descriptions using dependency relational patterns*, Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, pp. 1250–1258.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, *Bottom-up and top-down attention for image captioning and visual question answering*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086.

[3] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham, *Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 939–948.

[4] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei, *Attention on attention for image captioning*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4634–4643.

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., *Visual genome: Connecting language and vision using crowdsourced dense image annotations*, International journal of computer vision **123** (2017), no. 1, 32–73.

[6] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al., *Oscar: Object-semantics aligned pre-training for vision-language tasks*, European Conference on Computer Vision, Springer, 2020, pp. 121–137.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, Advances in neural information processing systems **28** (2015), 91–99.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., *Imagenet large scale visual recognition challenge*, International journal of computer vision **115** (2015), no. 3, 211–252.

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, *Show, attend and tell: Neural image caption generation with visual attention*, International conference on machine learning, PMLR, 2015, pp. 2048–2057.

[10] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai, *Auto-encoding scene graphs for image captioning*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10685–10694.

[11] Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos, *Corpus-guided sentence generation of natural images*, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 444–454.

[12] Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu, *I2t: Image parsing to text description*, Proceedings of the IEEE **98** (2010), no. 8, 1485–1508.

[13] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, *Exploring visual relationship for image captioning*, Proceedings of the European conference on computer vision (ECCV), 2018, pp. 684–699.

[14] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao, *Vinvl: Revisiting visual representations in vision-language models*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.