

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



Project 3

CÁC MÔ HÌNH DEEP LEARNING SỬ DỤNG
CHO BÀI TOÁN SINH TIÊU ĐỀ CHO ẢNH VÀ CẢI TIẾN

Giảng viên hướng dẫn: TS.Nguyễn Khánh Phương

Sinh viên thực hiện : Vũ Long Giang - 20183519

Lớp : Khoa học máy tính 04

Hà Nội - 2022

Mục lục

1	Đặt vấn đề	4
1.1	Dặt vấn đề	4
1.2	Các hướng tiếp cận	4
1.2.1	Phương pháp thủ công cho bài toán sinh tiêu đề cho ảnh	4
1.2.2	Phương pháp áp dụng deep learning cho bài toán sinh tiêu đề cho ảnh	5
2	Các nghiên cứu liên quan	8
2.1	Show, Attend and Tell	8
2.2	BUTD	9
2.3	AoANet	11
2.4	OSCAR and OSCAR+	13
2.4.1	OSCAR	13
2.4.2	OSCAR+	15
3	Đề xuất cải tiến	16
3.1	Adversarial Loss	16
4	Thực nghiệm	18
4.1	Bộ dữ liệu	18
4.2	Cấu hình và kịch bản thực nghiệm	18
4.2.1	Phần cứng	18
4.2.2	Độ đo	19
4.2.3	Show, Attend and Tell	20
4.2.4	Finetuned OSCAR+	21
4.3	Kết quả	21
5	Tổng kết	25

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. Diễn 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 [9] và OSCAR+ [22] 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 [16] và finetuned trên mô hình OSCAR+ [22], 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][18][20] 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 [3][7][21][17][9][22], 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:

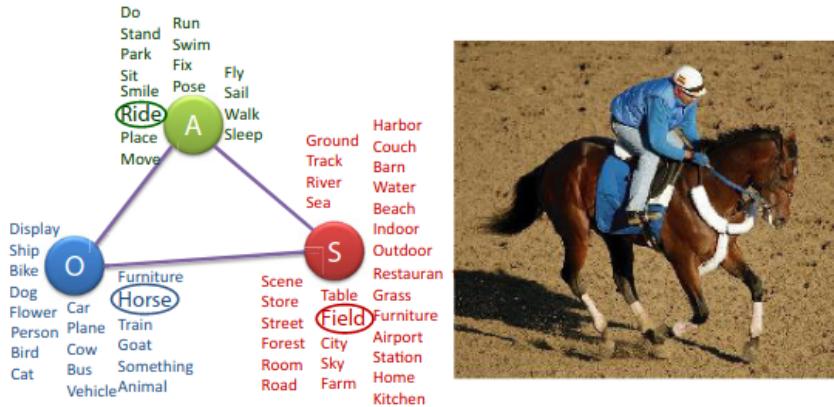
- Thống kê, đánh giá các hướng tiếp cận cho bài toán IC.
- Xử lý một bộ dữ liệu nhỏ phục vụ quá trình thực nghiệm.
- Xây dựng mô hình IC dựa trên mô hình Show, Attend and Tell [16]
- Xây dựng mô hình IC dựa trên mô hình pretrained OSCAR+ (VINVL) [22].
- Đề xuất cải tiến.
- Đánh giá kết quả trên các độ đo BLEU, CIDEr, SPICE.

1.2 Các hướng tiếp cận

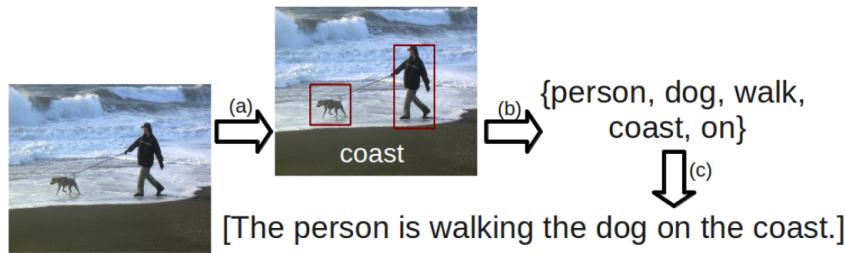
1.2.1 Phương pháp thủ công cho bài toán sinh tiêu đề cho ảnh

Trước khi các mô hình deep learning được áp dụng rộng rãi, có hai phương pháp phổ biến được áp dụng cho bài toán sinh tiêu đề cho ảnh: Retrieval based captioning và Template based captioning:

- **Retrieval based captioning:** Với phương pháp này, một tập các tiêu đề được xây dựng trước. Với một ảnh đầu vào, tiêu đề được sinh ra bằng cách lấy một hoặc một vài câu được cho là tốt nhất trong kho tiêu đề ban đầu. Tiêu đề được sinh có thể là một câu riêng lẻ hoặc là một đoạn gồm nhiều câu. Tiêu biểu cho phương pháp này, [5]



Hình 1.1: Ví dụ về việc sử dụng không gian biểu diễn (object, action, scene).



Hình 1.2: Tổng quát quá trình sinh tiêu đề của Corpus-guided Sentence Generation.

đề xuất một không gian biểu diễn (object, action, scene) để kết nối giữa ảnh và tiêu đề (ví dụ xem ở hình 1.1). Các nút và cạnh được xác định bằng mô hình các mô hình tuyến tính. Trong [11], tập 1 triệu ảnh đã có tiêu đề từ các trang web, nội dung của ảnh đầu vào được dùng để so sánh và sắp xếp tập 1 triệu ảnh. Tiêu đề của một vài ảnh tốt nhất được sử dụng làm tiêu đề cho ảnh đầu vào.

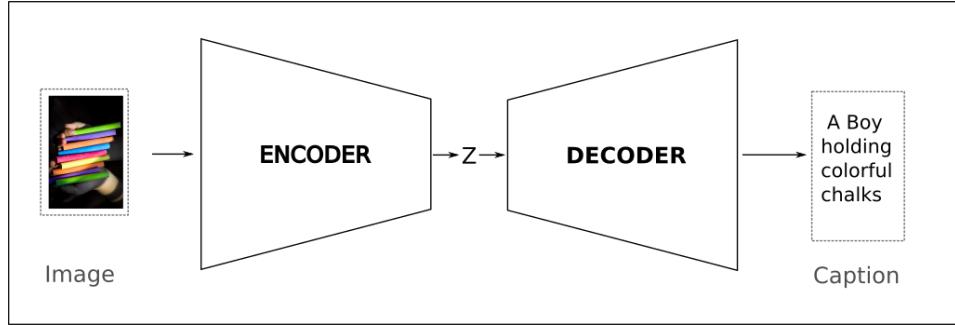
- **Template based captioning:** Trong phương pháp này, tiêu đề được sinh dựa trên một cú pháp hoặc một ràng buộc về ngữ nghĩa cho trước, mỗi ảnh phải được trích xuất thông tin về nội dung để kết nối tới đúng mẫu cho trước. Trong [19], một mẫu bao gồm 4 thành phần (Nouns-Verbs-Scenes-Prepositions) được sử dụng, ảnh đầu vào được cho qua các mô hình phân loại để lấy ra đối tượng và ngữ cảnh của bức ảnh, các thông tin đó sẽ là đầu vào cho một mô hình sinh tiêu đề, lần lượt từng thành phần trong cấu trúc (Nouns-Verbs-Scenes-Prepositions) được sinh ra tạo thành tiêu đề hoàn chỉnh (ví dụ ở hình 1.2).

Các phương pháp cổ điển có cách tiếp cận khá trực quan, tuy nhiên gặp nhiều hạn chế trong việc sinh ra đa dạng câu cũng như việc hiểu thông tin trong ảnh.

1.2.2 Phương pháp áp dụng deep learning cho bài toán sinh tiêu đề cho ảnh

Với sự phát triển của AI nói chung và deep learning nói riêng, các phương pháp deep learning ngày càng thể hiện sự hiệu quả trong các bài toán hiện đại ở nhiều lĩnh vực. Bài

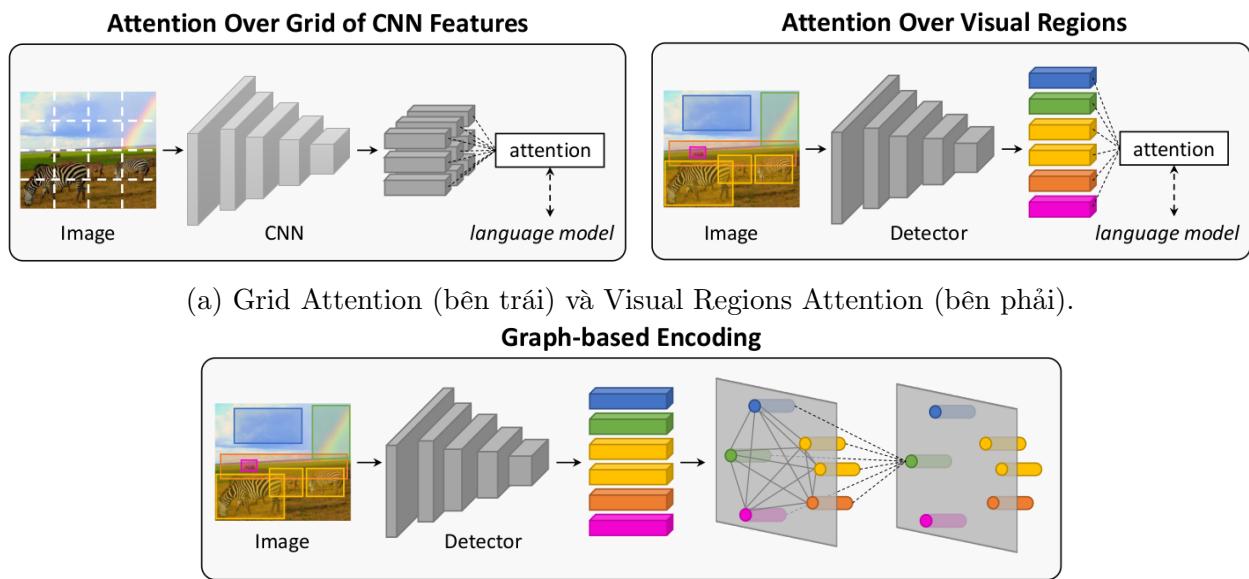
toán sinh tiêu đề cho ảnh cũng là một bài toán được cải thiện rất nhiều về hiệu năng khi áp dụng các mô hình học sâu. Một trong những mô hình đầu tiên đưa học sâu vào bài toán sinh tiêu đề cho ảnh là Show and Tell [15], sau đó không lâu Show, Attend and Tell [16] cho thấy cải thiện rõ rệt so với mô hình Show and Tell trước đó nhờ vào cơ chế Attention. Mặc dù có nhiều cải tiến ở các phần khác nhau, kiến trúc *Encoder-Decoder* vẫn là kiến trúc chuẩn được áp dụng cho bài toán sinh tiêu đề ảnh (hình 1.3).



Hình 1.3: Kiến trúc Encoder-Decoder.

Trong sơ đồ trên, *Encoder* đóng vai trò như một mạng mã hóa, lấy ra các vecto thuộc tính biểu diễn cho ảnh đầu vào, thường áp dụng các mô hình xử lý ảnh. *Decoder* là mạng giải mã, sinh tiêu đề từ biểu diễn ẩn Z của ảnh đầu vào, thường sử dụng các mô hình của lĩnh vực xử lý ngôn ngữ tự nhiên. Trong từng phần có những sự cải tiến riêng:

- **Encoder:** CNN là mô hình đơn giản và phổ biến nhất được sử dụng cho một bài toán xử lý ảnh. Tiêu biểu cho cách xử lý này là mô hình Show and Tell, khi một mạng Deep CNN được sử dụng làm Encoder. Cải tiến phổ biến sau đó là áp dụng cơ chế Attention cho việc mã hóa ảnh. Đối với dữ liệu ảnh, có nhiều cách để áp dụng cơ chế Attention, hai cách phổ biến và được dùng nhiều nhất là Grid Attention (chia ảnh ra thành các vùng nhỏ bằng nhau, coi mỗi vùng là một token và áp dụng cơ chế Attention trên đó) và Visual Regions Attention (các vùng quan trọng trong ảnh được trích xuất trước và cơ chế Attention được áp dụng trên các vùng đó), một cách trực quan, hai cơ chế được biểu diễn trong hình 1.4a. Tiêu biểu cho cách áp dụng Grid Attention có thể kể đến mô hình Show, Attend and Tell, BUTD[3] là ví dụ tương tự đối với Visual Regions Attention. Những hướng tiếp cận mới hơn cho mã hóa ảnh là những mô hình đồ thị (Graph-based Encoding, hình 1.4b), tiêu biểu trong đó là mô hình GCN-LSTM[21] và SGAE[17].
- **Decoder:** Tương tự như mạng *Encoder*, mạng giải mã có những cải thiện rõ rệt nhờ vào cơ chế Attention. Mô hình cơ bản nhất cho mạng giải mã là RNN hoặc LSTM, tiêu biểu vẫn là mô hình Show and Tell. Nhờ vào Attention, các mô hình sau có sự cải thiện rõ rệt về chất lượng của *Decoder*, trong đó có BUTD và AoANet[7]. Cuối cùng, mô hình hiện đại nhất được sử dụng cho mạng giải mã là những mô hình dựa trên BERT. Kết hợp với chiến lược pre-train, các mạng giải mã sử dụng BERT cho kết quả tốt hơn nhiều so với các mô hình trước đó, tiêu biểu là OSCAR và OSCAR+ (hiện đang là SOTA của bài toán sinh tiêu đề ảnh).



(a) Grid Attention (bên trái) và Visual Regions Attention (bên phải).

Graph-based Encoding

(b) Mô hình mã hóa ảnh sử dụng đồ thị.

Hình 1.4: Các hướng tiếp cận cho mã hóa ảnh.

Chương 2

Các nghiên cứu liên quan

Trong mục này, các nghiên cứu hiện đại (với hướng tiếp cận của kiến trúc *Encoder-Decoder*) 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 [16] 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 đầu 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:

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

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

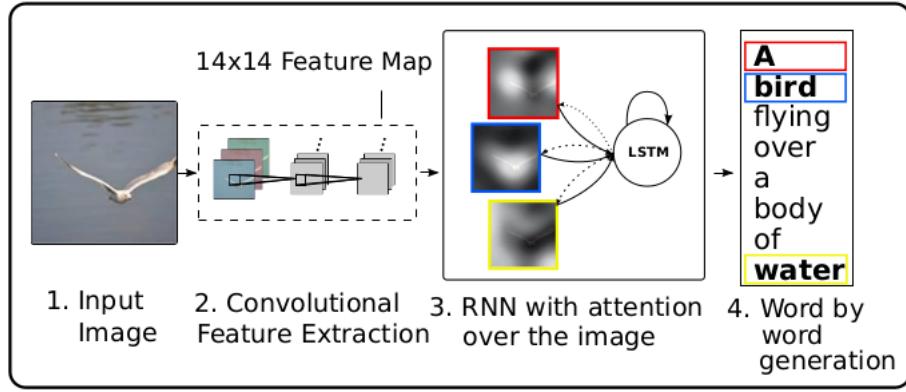
Cơ chế Attention được áp dụng trên tập các thuộc tính vùng:

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

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 [16], 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:

$$L_d = -\lg(P(y|x)) + \lambda \sum_i^L (1 - \sum_1^C \alpha_{ti})^2$$

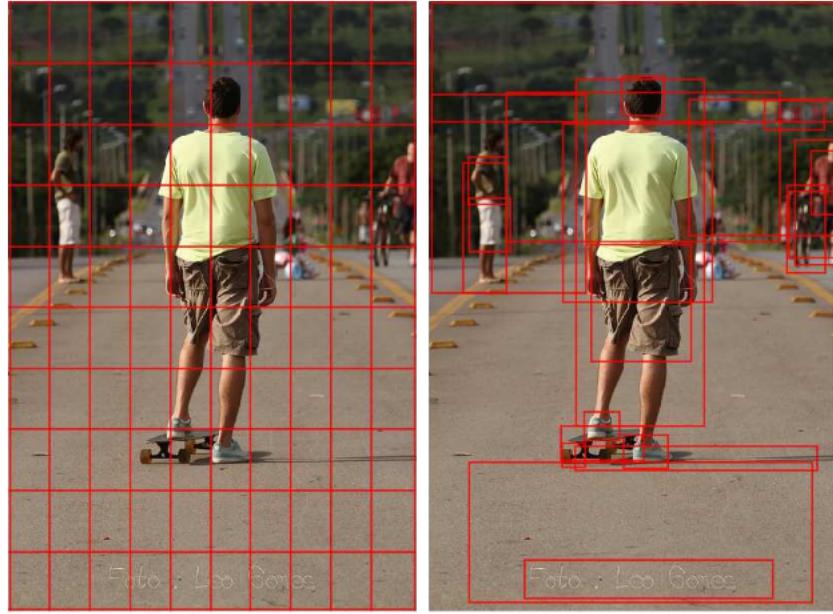
Trong đó thành phần $\lambda \sum_i^L (1 - \sum_1^C \alpha_{ti})^2$ được sử dụng để cố gắng đưa $\sum_t \alpha_{ti} \approx 1$, nghĩa là đưa độ quan trọng của các vùng trong ảnh trong cả quá trình sinh tiêu đề bằng nhau.

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 [3] đề xuất một mô hình giải quyết hai bài toán IC và VQA dựa trên Attention. Ý tưởng và điểm khác biệt của bài báo 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 [13] 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 trình 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 cùng 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[14] và tiếp tục train trên bộ dữ liệu Visual Genome[8]. 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 [16]. 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:

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

$$\bar{v} = \frac{1}{k} \sum_i^k v_i$$

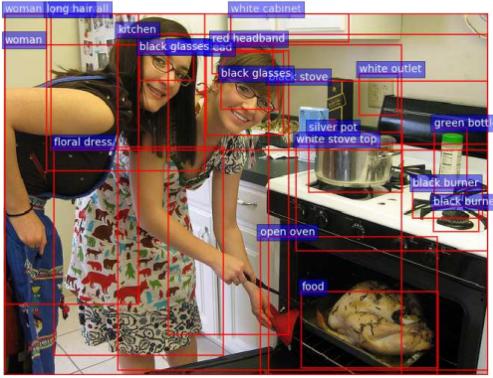
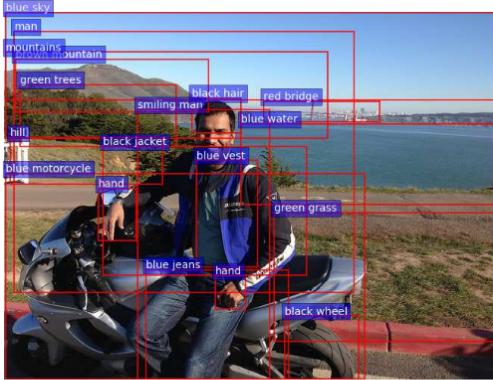
Trong đó, h_{t-1} là *hidden state* của LSTM tại bước $t-1$, v_i là các thuộc tính vùng của ảnh, W_e là ma trận embedding, Π là vecto one-hot của từ đầu vào tại bước t . *Hidden state* của lớp LSTM này là đầu vào cho một mạng Attend, từ đó tạo ra vecto ngữ cảnh \hat{v}_t là đầu vào của lớp Language LSTM.

- 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 .

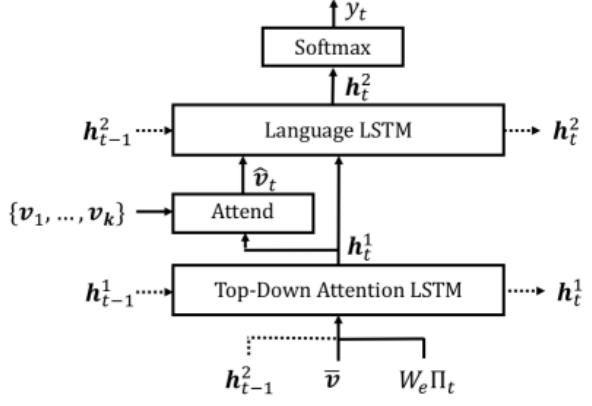
Hàm mục tiêu được sử dụng:

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

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 [9] và OSCAR+[22]. 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

Attention on Attention for Image Captioning[7] đề xuất một mô hình chung áp dụng không chỉ cho bài toán IC mà còn áp dụng cho bất kì bài toán nào có sử dụng cơ chế Attention. Ý tưởng chính của AoANet là sử dụng thêm một mô hình Attention bên trên đầu ra của mô hình Attention cho trước. Điều này được cho là có tác dụng trong việc đánh giá kết quả của mô hình Attention trước đó. AoANet (kiến trúc chi tiết ở hình 2.4) được đề xuất như là một mô hình độc lập được sử dụng trong nhiều trường hợp. Kiến trúc AoANet được áp dụng cho bài toán IC được mô tả ngay sau đây.

Mô hình IC được đề xuất áp dụng AoANet cho cả *Encoder* và *Decoder*:

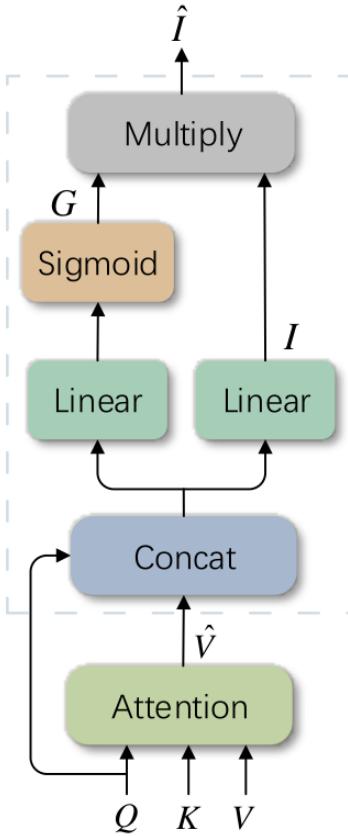
- **Encoder:** Với 1 ảnh đầu vào, một tập các vecto thuộc tính $A = \{a_1, a_2, \dots, a_k\}$ được trích xuất sử dụng mô hình CNN hoặc R-CNN. Thay vì lấy A làm đầu vào cho *Decoder*, một mạng nơ-ron sử dụng AoANet để biến đổi A thành A' (chi tiết hình 2.5a):

$$A' = \text{LayerNorm}(A + \text{AoA}^E(f_{mh-att}, W^{Q_e} A, W^{K_e} A, W^{V_e} A))$$

Trong đó, $W^{Q_e}, W^{K_e}, W^{V_e}$ là các ban trận biến đổi tuyến tính, f_{mh-att} là hàm multi-head attention sử dụng để xử lý đầu vào. Từ đó, tập vecto thuộc tính A' được sử dụng làm đầu vào cho mạng *Decoder*.

- **Decoder:** Vecto c_t được định nghĩa để tính toán xác suất cho mô hình sinh tiêu đề:

$$p(y_t | y_{1:t-1}, I) = \text{softmax}(W_p c_t)$$



Hình 2.4: AoANet architecture

Trong đó, W_p là ma trận trọng số sẽ được tối ưu trong quá trình huấn luyện mô hình, I là ảnh đầu vào. Vecto c_t lưu giữ trạng thái của quá trình sinh tiêu đề cũng như được cập nhật thêm các thông tin mới từ vecto \hat{a} (là kết quả của một mô hình attention với đầu vào là *hidden state* qua mỗi bước sinh, chi tiết xem hình 2.5b).

Các lớp LSTM được sử dụng cho quá trình sinh tiêu đề, với đầu vào bao gồm vecto c_t , $\bar{a} = \frac{1}{k} \sum_i a_i$ (trong đó, a_i là các vecto thuộc tính của A') và thành phần w_t là từ đầu vào (trong quá trình huấn luyện w_t là các từ trong tiêu đề thật, trong quá trình đánh giá, w_t là các từ được sinh ở bước liền trước). Vecto ngữ cảnh c_t được xác định như sau:

$$c_t = AoA^D(f_{mh-att}, W^{Q_d}[h_t], W^{K_d}A, W^{V_d}A)$$

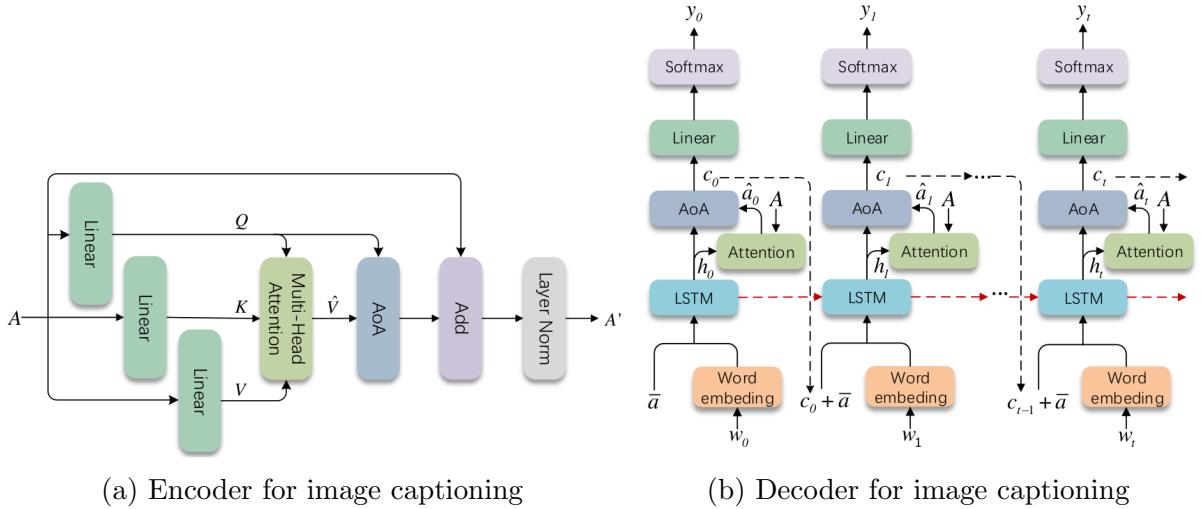
Trong đó, $W^{Q_d}, W^{K_d}, W^{V_d}$ là các ma trận trọng số được tối ưu trong quá trình huấn luyện.

Hai hàm mục tiêu tương ứng với hai chiến thuật huấn luyện được sử dụng đối với mô hình đề xuất:

- **Cross-Entropy Loss:**

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

- **CIDEr-D Score Optimization:** Được sử dụng để tối ưu trực tiếp các độ đo không



Hình 2.5: AoANet for image captioning

khả vi.

$$L_{RL}(\theta) = -E_{y_{1:T} \sim p_\theta}[r(y_{1:T})]$$

Nhận xét: Mô hình AoANet thể hiện ưu điểm so với các mô hình trước đó nhờ vào việc sử dụng thêm một mô hình attention bên trên kiến trúc thông thường. Mô hình cho kết quả tốt và là SOTA tại thời điểm công bố. Việc áp dụng AoANet không chỉ dừng lại trong bài toán IC mà còn mở rộng ra các hướng khác như dịch máy,...

2.4 OSCAR and OSCAR+

Các mô hình pre-trained đang dần trở thành xu hướng các bài toán về xử lý ngôn ngữ tự nhiên và thị giác máy tính. Gần đây, các mô hình pre-trained bắt đầu được sử dụng trong lớp bài toán cross-modal tạo nên hiệu quả rõ rệt và cho thấy ưu thế so với các phương pháp trước đó, đáng kể nhất trong các mô hình này có thể kể đến OSCAR và OSCAR+ (VinVL) được phát triển bởi Microsoft. Ý tưởng chính của OSCAR:

- Sử dụng kết quả của mô hình phát hiện đối tượng là điểm neo giúp cho quá trình học dễ dàng hơn. Một cách dễ hiểu, OSCAR miêu tả các đối tượng trong ảnh bằng vecto embedding của tập từ điển *Decoder*, điều này làm cho việc biểu diễn các đối tượng trong không gian ảnh sang không gian ngôn ngữ gần nhau hơn.
- Sử dụng các mô hình pre-trained trong từng thành phần: Đối với *Encoder*, mô hình pre-trained OD được sử dụng (khác nhau ở OSCAR và OSCAR+) và được huấn luyện trên bộ dữ liệu ảnh lớn, đối với *Decoder* BERT được áp dụng như một mô hình sinh.

Phần tiếp theo trình bày chi tiết về cấu hình và cách huấn luyện OSCAR, điểm khác biệt và cải tiến của OSCAR+ so với OSCAR.

2.4.1 OSCAR

Bộ dữ liệu với N ví dụ được ký hiệu $D = \{(I_i, w_i)\}_{i=1}^N$, với I là ảnh đầu vào và w là chuỗi. OSCAR nhận mỗi ví dụ đầu vào với dạng (w, q, v) , trong đó w là chuỗi, q là chuỗi word

$$\mathbf{x} \triangleq [\underbrace{\mathbf{w}}_{\text{language}}, \underbrace{\mathbf{q}, \mathbf{v}}_{\text{image}}] = [\underbrace{\mathbf{w}, \mathbf{q}}_{\text{language}}, \underbrace{\mathbf{v}}_{\text{image}}] \triangleq \mathbf{x}'$$

Hình 2.6: Hai góc nhìn đầu vào của BERT.

embedding của các đối tượng được xác định trong ảnh, v là tập vecto thuộc tính của ảnh. Chi tiết, q và v được xác định như sau. Ảnh đầu vào I được đưa qua một mô hình Faster R-CNN để lấy ra tập (v', z) , trong đó v' là vecto thuộc tính, z là vecto vùng (thể hiện cho vị trí của vùng trên ảnh, thường có 4 hoặc 6 chiều). v' và z sau đó được nối lại để tạo ra những vecto mang cả ý nghĩa về vị trí và về các thuộc tính của vùng, vecto sau khi được nối đưa qua một phép chiếu tạo thành vecto thuộc tính v , nhằm đảm bảo v có số chiều phù hợp với q và w (nghĩa là cùng số chiều với vecto embedding). Mô hình Faster R-CNN cũng được dùng để phát hiện các đối tượng trong ảnh, các đối tượng này được đưa qua lớp embedding tạo ra chuỗi q . Việc còn lại là huấn luyện một mô hình BERT với đầu vào có dạng (w, q, v) .

Do q vừa mang ý nghĩa của ảnh đầu vào, vừa mang tính chất của chuỗi được sinh (q là chuỗi embedding, nên nằm trong không gian biểu diễn của chuỗi cần sinh), hai góc nhìn được áp dụng đối với đầu vào của BERT (hình 2.6).

- **Dictionary View:** là góc nhìn coi w và q thuộc cùng một thành phần và v đại diện cho ảnh đầu vào. Một chuỗi các token được định nghĩa $h = [w, q]$ và Masked Token Loss (MTL) được áp dụng cho quá trình huấn luyện. Tại mỗi bước, mỗi token trong chuỗi h được thay thế với xác suất 15% bằng token [MASK]. Mục tiêu của quá trình huấn luyện là tìm ra các token bị thay thế đó dựa trên các token xung quanh và tập vecto v :

$$L_{MTK} = -E_{(v,h) \sim D} \log p(h_i | h_{\setminus i}, v)$$

- **Modality View:** là góc nhìn coi q và v là hai thành phần biểu diễn cho ảnh đầu vào, w là chuỗi đầu ra. Tập vecto $h' = [q, v]$ được định nghĩa và hàm Contrastive Loss được sử dụng đối với góc nhìn này. Một tập các biểu diễn ảnh sẽ được lấy mẫu và thay thế q bằng một chuỗi khác với xác suất 50%. Mục tiêu của việc huấn luyện với góc nhìn này là phát hiện xem đâu là một biểu diễn sai (đã bị thay thế q). Việc này được thực hiện bằng cách, thêm một lớp fully-connected để phân loại token [CLS] ở đầu ra của BERT. Hàm mất mát được sử dụng:

$$L_C = -E_{(h', w) \sim D} \log p(y | f(h', w))$$

Trong đó, y là đầu ra của lớp fully-connected, $y = 1$ nếu biểu diễn ảnh là đúng, ngược lại $y = 0$ nếu q đã được thay đổi.

Hàm mất mát được sử dụng trong cả quá trình huấn luyện OSCAR:

$$L_{pre-training} = L_{MTK} + L_C$$

Bộ dữ liệu sử dụng trong quá trình huấn luyện OSCAR gồm 6.5 triệu cặp ảnh-chuỗi, được xây dựng từ 5 bộ dữ liệu: COCO, Conceptual Captions, SBU captions, flicker30k và GQA.

Nhận xét: Kiến trúc OSCAR có nét tương đồng với BUTD, điểm khác biệt lớn nhất là OSCAR sử dụng thêm q là chuỗi embedding của các đối tượng trong ảnh, điều này góp

phần lớn trong việc đưa không gian biểu diễn của ảnh và của chuỗi gần nhau hơn trong quá trình huấn luyện. OSCAR cho kết quả tốt trong các tác vụ phía sau như Image Captioning (IC), Visual Question Answering (VQA) hay Image Text Retrieval. Kết quả của OSCAR là SOTA trong 6 tác vụ vào thời điểm công bố, bao gồm cả IC.

2.4.2 OSCAR+

Về cơ bản OSCAR+ sử dụng kiến trúc mô hình giống như OSCAR, có hai điểm cải tiến được sử dụng giúp cho kết quả của OSCAR+ tốt hơn mô hình OSCAR trước đó:

- Cải thiện mô hình OD cho việc hiểu và biểu diễn ảnh. Ngoài tên đối tượng và các thuộc tính vùng, mô hình trích xuất thêm các thuộc tính, hành vi của đối tượng (ví dụ như màu sắc, biểu cảm hoặc tương tác giữa các đối tượng). Cụ thể, đối với mô hình OD, 4 bộ dữ liệu được sử dụng bao gồm Visual Genome, COCO, Object365 và OpenImagesV5. Do đa phần các bộ dữ liệu không thông tin về thuộc tính của đối tượng (chỉ bao gồm các đối tượng có trong ảnh) nên chiến lược pre-train, finetuning được áp dụng. Mô hình OD ban đầu được huấn luyện với 4 bộ dữ liệu, sau đó finetune với bộ dữ liệu Visual Genome để lấy ra thông tin về thuộc tính của các đối tượng trong ảnh.
- Thay đổi về hàm mất mát được sử dụng trong quá trình huấn luyện. Hàm mất mát được sử dụng vẫn bao gồm 2 thành phần Masked Token Loss và Contrastive Loss, tuy nhiên khác với OSCAR, OSCAR+ đề xuất hàm Contrastive Loss với 3 đối tượng:

$$L_{CL3} = -E_{(w,q,v,c) \sim \tilde{D}} \log p(c|f(w, q, v))$$

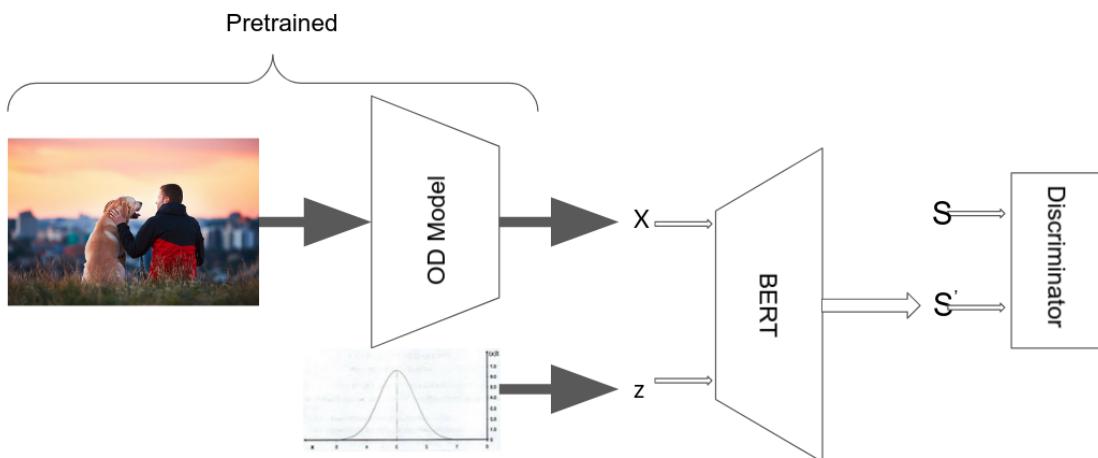
Các ví dụ "*lỗi*" được tạo ra theo 2 cách, thay w bởi w' hoặc thay q bởi q' . Tại đầu ra của BERT, một mạng fully connected được sử dụng để phân loại token [CLS] theo 3 lớp: cặp ví dụ đúng ($c = 0$), sai w ($c = 1$) và sai q ($c = 2$). Tập ví dụ được tạo với tỷ lệ 50% là ví dụ đúng, 25% ví dụ được thay đổi w và 25% ví dụ được thay đổi q .

Nhận xét: Dưa thêm thông tin về thuộc tính của đối tượng ở đầu vào cho BERT cũng như thay đổi hàm mất mát cho thấy hiệu quả tốt đối với mô hình. Thay đổi hàm mất mát làm cho mô hình học tốt hơn đối với những tác vụ như Visual Question Answering (VQA). Kết quả của OSCAR+ đạt được SOTA trên 7 tác vụ và thường tạo ra chênh lệnh lớn so với các mô hình SOTA trước đó.

Chương 3

Đề xuất cải tiến

3.1 Adversarial Loss



Hình 3.1: Kiến trúc đề xuất.

Hàm mất mát đối kháng được sử dụng phổ biến từ khi mạng GAN ra đời năm 2014. Từ đó đến nay, các mô hình cải tiến của GAN, sử dụng hàm mất mát đối kháng cho những kết quả rất tích cực, đã có ứng dụng thực tế cho bài toán sinh ảnh. Ý tưởng sử dụng hàm mất mát đối kháng khác phát từ việc, mô hình sinh một văn bản từ ảnh có nét tương đồng với mô hình sinh một ảnh từ ảnh cho trước, với các mô hình đề xuất trước, các câu sinh ra thường chỉ có một dạng cố định, với hàm mất mát đối kháng và với cơ chế sử dụng các đầu vào nhiều, cải tiến này mong muốn từ một ảnh đầu vào có thể sinh ra những tiêu đề đầu ra có sắc thái khác nhau. Trong [4], tác giả đề xuất một mô hình cGAN cho bài toán sinh tiêu đề, tuy nhiên Generator trong bài là một mạng dạng CNN-RNN đơn thuần, CNN được pretrain trên tập đầu với với hàm mất mát là MLE. Đề xuất cải tiến (hình 3.1):

- Sử dụng một mô hình nhận diện vật thể để trích xuất thông tin từ ảnh (tương tự như trong mô hình OSCAR và OSCAR+). Các thông tin này bao gồm đối tượng, thuộc tính của đối tượng và các thuộc tính vùng sẽ được dùng làm đầu vào cho mạng sinh Generator.

-
- Sử dụng một mô hình BERT làm Generator. Khác với OSCAR và OSCAR+, tiêu đề mục tiêu không được đưa vào quá trình huấn luyện cho BERT mà được dùng như nhãn của mạng Discriminator. Ngoài đầu vào là các thuộc tính của ảnh, một vector nhiễu z được sử dụng với mong muốn tăng độ đa dạng của câu tiêu đề được sinh ra.
 - Tại mạng Discriminator, câu được sinh và câu mục tiêu được ghép với nhau, phân cách bởi token [SEP] làm đầu vào cho một mạng phân loại với 2 nhãn: 0 nếu độ tương đồng giữa hai câu không cao, 1 nếu hai câu có độ tương đồng cao. Discriminator này cũng được pretrain như một mô hình ngôn ngữ.

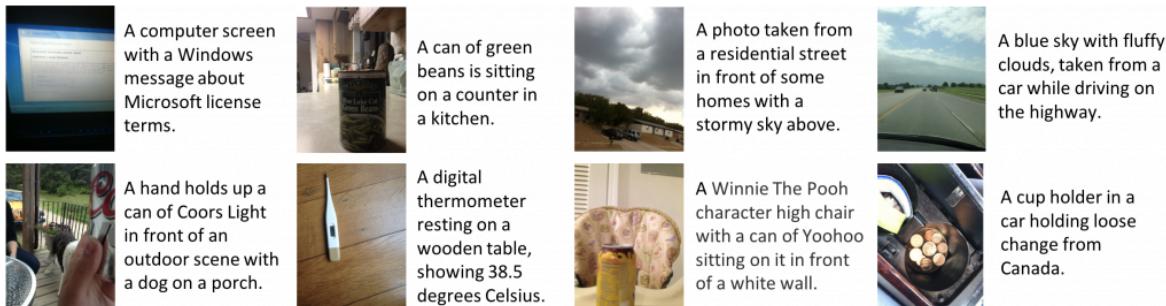
Tuy nhiên, trong môn học này, mô hình chưa hoàn thành việc xây dựng và đánh giá thực nghiệm. Đây sẽ là hướng tiếp theo sau khi kết thúc môn học.

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 [6] 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 P100

RAM 13 Gb

4.2.2 Độ đo

BLEU

BLEU[12] đánh giá sự giống nhau của một câu được sinh c với tập mục tiêu S dựa trên tần suất xuất hiện của các n-grams trong c có xuất hiện trong S . Cụ thể, đối với tập S , đến số lượng xuất hiện của các n-grams trong từng câu. Số lần xuất hiện của 1 n-grams trong tập S được xác định là số lượng lớn nhất n-grams đó xuất hiện trong một câu. Độ đo BLEU được tính bằng số lượng n-grams khớp với tập mục tiêu chia cho tổng số lượng n-grams trong câu c .

Ví dụ:

Câu được sinh: the the the the the the.

Câu mục tiêu 1: The cat is on the mat.

Câu mục tiêu 2: There is a cat on the mat.

Độ đo BLEU1 (unigram) của câu được sinh trong ví dụ trên là 2/7. Do số lần xuất hiện của từ "the" trong tập mục tiêu là 2 (xuất hiện nhiều nhất với 2 lần trong câu 1).

ROUGE_L

ROUGE_L[10] là độ đo đánh giá sự giống nhau giữa hai đoạn, ban đầu được đề xuất để đánh giá hiệu năng cho bài toán tóm tắt văn bản. Ý tưởng chính của ROUGE_L là hai câu văn có đoạn chung lớn nhất (Longest Common Sequence - LCS) càng lớn thì càng giống nhau. Dựa trên đó, ROUGE_L được xây dựng như sau:

$$R_{lcs} = \frac{\sum_{s_i \in S_1} \max_{s_j \in S_2} LCS(s_i, s_j)}{m + n}$$

$$P_{lcs} = \frac{\sum_{s_i \in S_1} \max_{s_j \in S_2} LCS(s_i, s_j)}{n}$$

$$\beta = \frac{P_{lcs}}{R_{lcs}}$$

$$ROUGE_L = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

Trong đó đại lượng R_{lcs} , P_{lcs} lần lượt đại diện cho điểm số Recall và Precision giữa hai tập câu S_1 và S_2 . ROUGE_L đánh giá cao giá trị của độ đo Recall hơn, nghĩa là mong muốn câu được sinh có chuỗi giống với câu mục tiêu là dài nhất.

CIDEr

Độ đo CIDEr đánh giá sự giống nhau của một câu c với tập câu mục tiêu S (reference sentences) dựa trên tần suất xuất hiện của các n-grams. Điểm CIDEr càng cao khi các n-grams trong c xuất hiện càng nhiều trong S , các n-grams không xuất hiện trong S không xuất hiện trong c , các n-grams xuất hiện trong hầu hết các câu của S cũng được đánh trọng số thấp hơn. Để tính toán các tần suất này, TF-IDF được áp dụng để đánh trọng số cho từng n-gram:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}\right)$$

Trong đó w_l là 1 n-gram, Ω là tập từ điển của các n-grams, $h_k(s)$ là số lần n-gram w_k xuất hiện trong câu s , I là tập các ảnh trong bộ dữ liệu và $g_k(s)$ là trọng số cho n-gram w_k trong câu s .

Từ đó, công thức tính điểm CIDEr được xây dựng:

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|}$$

Trong đó, $g^n(c)$ là vecto tạo bởi các $g_k(c)$ có độ dài là n. Nhiều giá trị của n được sử dụng và điểm số CIDEr tổng được tính bằng:

$$CIDEr(c_i, S_i) = \sum_{n=1}^N w_n CIDEr_n(c_i, S_i)$$

Với w_n là trọng số đối với từng giá trị của n.

SPICE

SPICE[2] khác với 2 độ đo trước khi đánh giá cả chất lượng nội dung của câu được sinh c so với một câu dụng tiêu s. Với một câu c, một mô hình biểu diễn ngữ nghĩa (ở đây Scene Graph được sử dụng) được dùng để biểu diễn câu:

$$G(c) = \langle O(c), E(c), K(c) \rangle$$

Trong đó, $G(c)$ là đồ thị biểu diễn câu c , $O(c)$ là tập các đối tượng xuất hiện trong câu c , $E(c)$ là tập các liên kết giữa các đối tượng và $K(c)$ là tập thuộc tính của từng đối tượng.

Các đồ thị này, sau đó được sử dụng để đánh giá mức độ tương đồng giữa các câu. Cụ thể, một hàm T được định nghĩa để lấy ra các tập hợp đối tượng, thuộc tính có ý nghĩa từ đồ thị của câu, mỗi tập hợp có thể bao gồm 1, 2 hoặc 3 thành phần, ví dụ về tập hợp này:

$$\{(girl), (court), (girl, young), (girl, standing), (court, tennis), (girl, on - top - of, court)\}$$

Dộ đo SPICE được xác định như sau:

$$\begin{aligned} P(c, S) &= \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|} \\ R(c, S) &= \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|} \\ SPICE(c, S) &= F_1(c, S) = \frac{2.P(c, S).R(c, S)}{P(c, S) + R(c, S)} \end{aligned}$$

Trong đó, $P(c, S)$ là độ đo precision của câu c đối với tập S , $R(c, S)$ là độ đo recall của câu c đối với tập S . $SPICE(c, S)$ là độ đo SPICE tương ứng cũng là độ đo F1-score của câu c đối với tập S . $|T(G(c)) \otimes T(G(S))|$ thể hiện số lượng các tập hợp giống nhau sau khi sử dụng hàm T của $G(c)$ và $G(S)$.

4.2.3 Show, Attend and Tell

Tiền xử lý dữ liệu: Các câu tiêu đề được đưa về dạng chuẩn (ví dụ: fishes được đưa về fish) sử dụng thư viện nltk, sau đó được tokenize thành các token.

Mô hình: Mạng Encoder sử dụng mô hình ResNet101 đã được pretrained, quá trình thử nghiệm thực hiện đánh giá trong hai trường hợp fine-tune Encoder hoặc không fine-tune.

Hàm tối ưu: Adam với các thông số: $\beta_1 = 0.9$, $\beta_2 = 0.99$, learning rate cho mạng Encoder là $1e^{-5}$, của Decoder là $2e^{-5}$.

Một vài cấu hình khác: Điều chỉnh learning rate khi kết quả không cải thiện qua một vài epoch. Cụ thể, mỗi 8 epoch không có sự cải thiện về chất lượng, learning rate mới được cập nhật bằng 0.8 lần learning cũ. Sử dụng clip gradient để tránh hiện tượng vanishing gradient trong quá trình huấn luyện. Trong quá trình đánh giá, sử dụng kỹ thuật beam search với beam size bằng 5.

4.2.4 Finetuned OSCAR+

Việc finetune ở đây chỉ là tiếp tục huấn luyện mô hình BERT. Việc trích xuất các thông tin từ ảnh (đối tượng và thuộc tính của đối tượng) được thực hiện bởi mô hình Scence Graph Benchmark của Microsoft. Việc tiền xử lý là thực hiện trích xuất thông tin từ ảnh sử dụng Scence Graph Benchmark và tạo các file đầu vào theo định dạng của OSCAR.

Mô hình: Finetune mô hình OSCAR+ base (sử dụng BERT base đã được train với 2000000 step - tương ứng với 2000000 batch).

Cấu hình: Finetune với 60 epoch, Adam với các thông số: $\beta_1 = 0.9$, $\beta_2 = 0.99$, learning rate $8e^{-6}$, sử dụng Linear Scheduler điều chỉnh learning rate. Sử dụng cơ chế label smoothing.

4.3 Kết quả

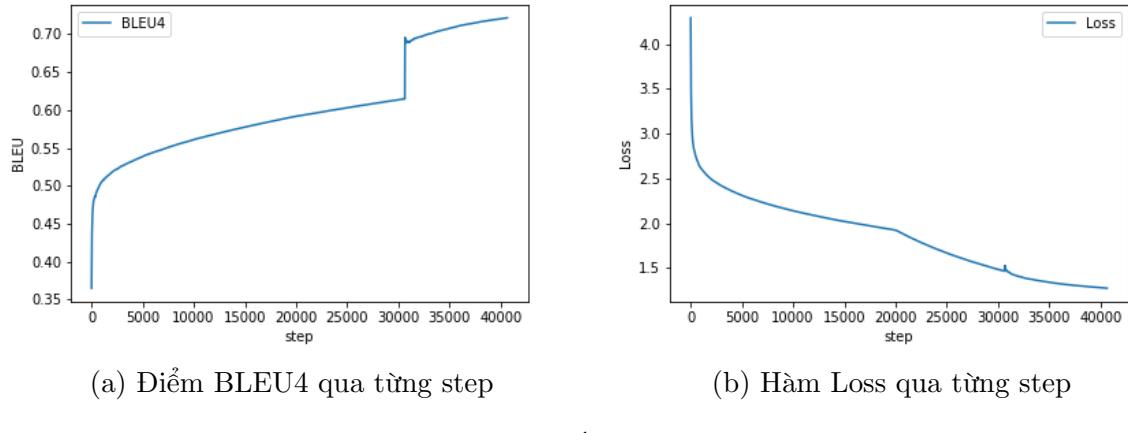
Đối với mô hình Show, attend and tell không thực hiện finetune mạng Encoder các câu tiêu đề xảy ra hiện tượng chỉ có thể sinh ra câu: "Quality issues are too severe to recognize visual content.". Hiện tượng này do trong tập huấn luyện, khá nhiều ảnh có chất lượng thấp và chứa câu trên như một tiêu đề.

Tổng hợp kết quả trên 3 mô hình thực nghiệm

	BLEU4	ROUGE_L	CIDEr	SPICE
Show, attend and tell (no finetune)	35.21	37.11	133.43	13.62
Show, attend and tell (finetune)	34.64	36.86	131.76	13.44
Finetuned OSCAR+	39.83	54.49	146.39	18.70

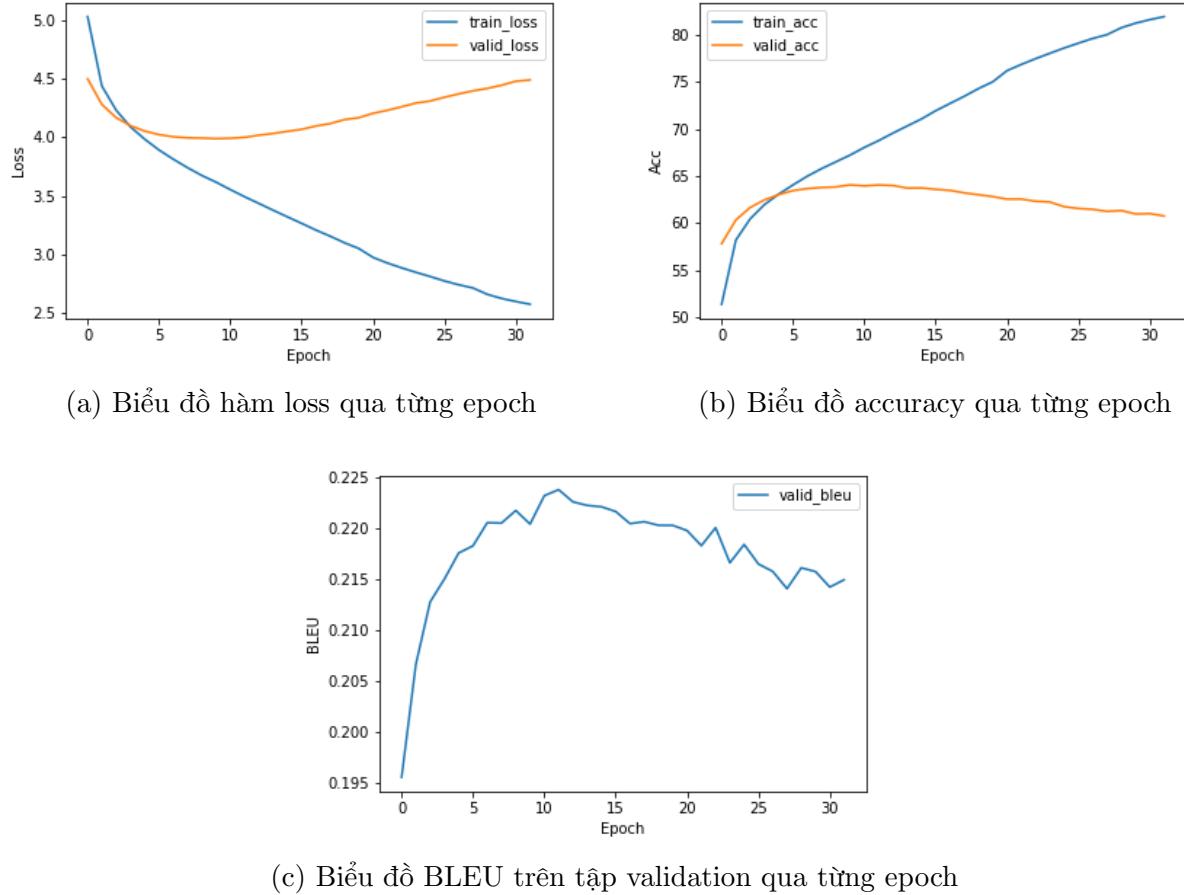
Bảng 4.1: Kết quả thực nghiệm trên 3 mô hình.

Kết quả huấn luyện mô hình finetune OSCAR+:

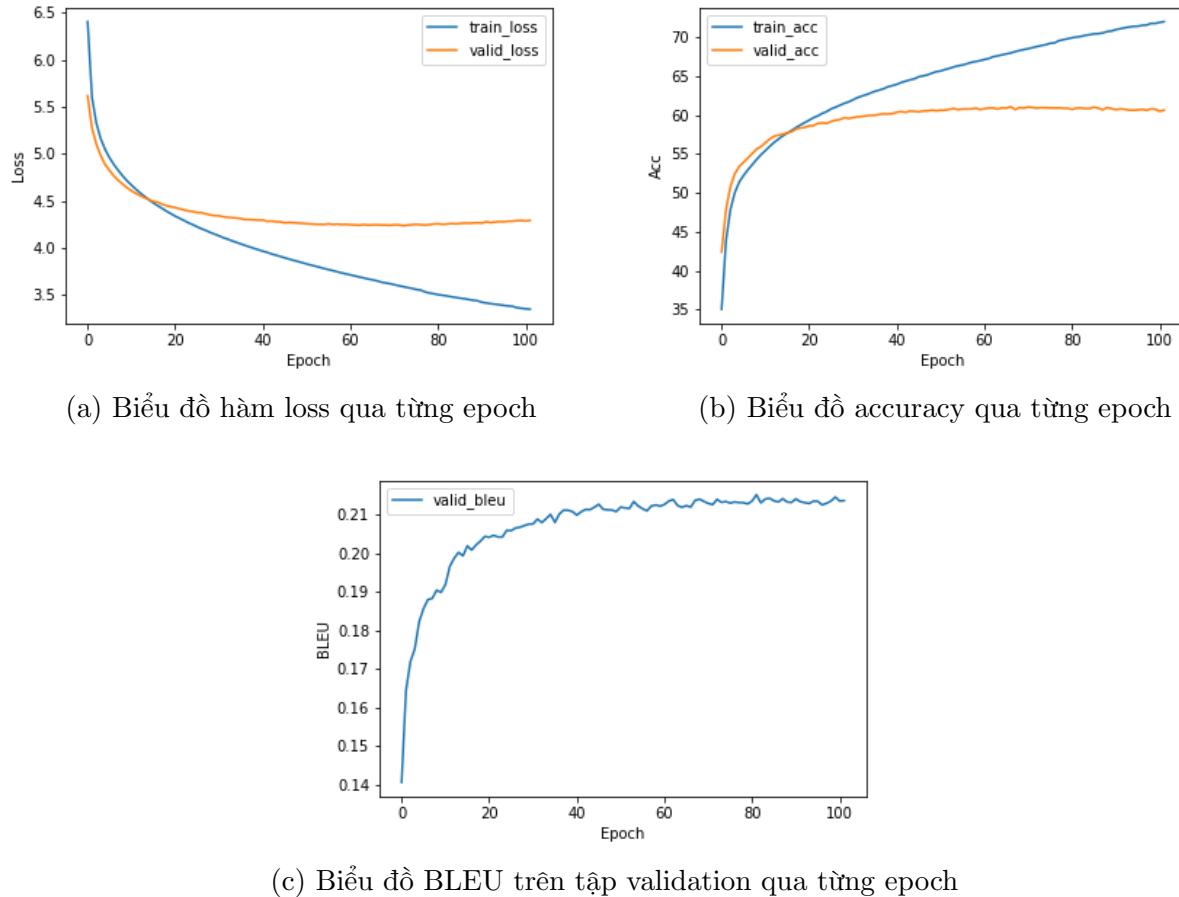


Hình 4.2: Quá trình huấn luyện mạng OSCAR+

Kết quả huấn luyện mô hình Show, Attend and Tell:



Hình 4.3: Quá trình huấn luyện mô hình Show, Attend and Tell không finetune Encoder.



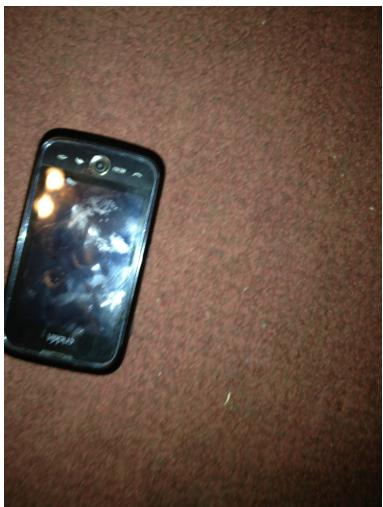
Hình 4.4: Quá trình huấn luyện mô hình Show, Attend and Tell finetune Encoder.

Nhận xét: Các mô hình đều cho kết quả khá tốt trên tập kiểm thử. Mô hình Show, attend and tell không finetune mạng *Encoder* hội tụ nhanh hơn so với khi finetune (quá trình huấn luyện kết thúc sau 32 epoch so với 102 epoch). Mô hình cho kết quả tốt nhất là OSCAR+ với kết quả cao rõ rệt so với 2 mô hình còn lại ở cả 4 độ đo.

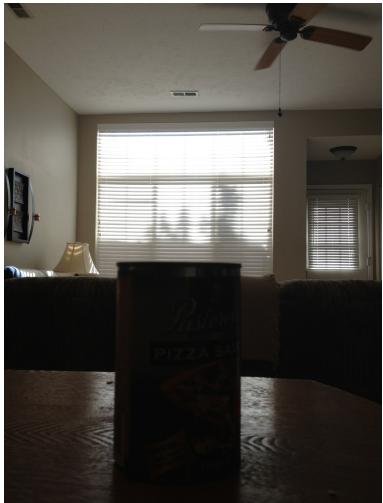
Một số kết quả sinh tiêu đề của 2 mô hình:



- a) a person is holding a black package of food.
- b) quality issues are too severe to recognize visual content.
- c) Here is a white ribbed tank top that covers the entire photo and has some wrinkles in it.



- a) a can of soda is on top of a table.
- b) a cell phone is lying on top of a table.
- c) A small black handheld cell phone on a table.



- a) a picture of the food is on the packaging.
- b) a living room with a fan and a window with blinds.
- c) A can of pizza sauce as you look out over a living room.



- a) a clear plastic bottle of something with a red label on it.
- b) a bottle of hand sanitizer is on a table.
- c) A bottle of famous ballpark mustard from Bertman.

(a) mô hình Show, attend and tell có finetune, (b) mô hình OSCAR+, (c) tiêu đề trong tập dữ liệu.

Chương 5

Tổng kết

Báo cáo này đã trình bày tổng quan về các hướng tiếp cận của bài toán sinh tiêu đề cho ảnh, một số mô hình tiêu biểu của bài toán này. Một bộ dữ liệu với kích thước nhỏ được chuẩn bị và thực hiện đánh giá mô hình thực nghiệm. Mô hình Show, attend and tell với hai cài đặt (finetune và không finetune *Encoder*) và mô hình finetune OSCAR+ được cài đặt để đánh giá kết quả trên bộ dữ liệu đã chuẩn bị. Kết quả cho thấy mô hình OSCAR+ hoàn toàn vượt trội so với mô hình Show, attend and tell, điều này có thể lý giải nhờ vào lượng ảnh và tiêu đề đã pretrain đối với OSCAR+ là lối, kiến trúc mô hình cũng hiện đại và phức tạp hơn. Báo cáo cũng đưa ra ý tưởng để cải thiện chất lượng mô hình cho bài toán. Mô hình GAN đã được áp dụng rất thành công trong bài toán sinh ảnh, tuy nhiên đối với bài toán cross-modal (sinh văn bản từ ảnh) đây là một phương pháp khá mới mẻ. Với những thành công trong bài toán sinh ảnh, hướng đi này cũng hứa hẹn sẽ cho kết quả tốt đối với bài toán sinh tiêu đề - có mô hình tương tự. Tuy nhiên, trong khuôn khổ môn học, báo cáo chưa trình bày được quá trình cài đặt và thực nghiệm trên ý tưởng đã nêu, đó cũng là hướng đi tiếp theo. Những dữ liệu trong bài báo được thu thập từ nhiều nguồn, mong muốn mang đến những thông tin có ích cho người đọc.

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, Basura Fernando, Mark Johnson, and Stephen Gould, *Spice: Semantic propositional image caption evaluation*, European conference on computer vision, Springer, 2016, pp. 382–398.
- [3] 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.
- [4] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju, *Improving image captioning with conditional generative adversarial nets*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8142–8150.
- [5] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth, *Every picture tells a story: Generating sentences from images*, European conference on computer vision, Springer, 2010, pp. 15–29.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] Xiuju 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.

-
- [10] Chin-Yew Lin, *Rouge: A package for automatic evaluation of summaries*, Text summarization branches out, 2004, pp. 74–81.
 - [11] Vicente Ordonez, Girish Kulkarni, and Tamara Berg, *Im2text: Describing images using 1 million captioned photographs*, Advances in neural information processing systems **24** (2011), 1143–1151.
 - [12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, *Bleu: a method for automatic evaluation of machine translation*, Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
 - [13] 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.
 - [14] 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.
 - [15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, *Show and tell: A neural image caption generator*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
 - [16] 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.
 - [17] 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.
 - [18] 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.
 - [19] ———, *Corpus-guided sentence generation of natural images*, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 444–454.
 - [20] 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.
 - [21] 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.
 - [22] 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.