**XÂY DỰNG BỘ DỮ LIỆU TIẾNG VIỆT**

**CHO BÀI TOÁN TRẢ LỜI CÂU HỎI TRỰC QUAN**

**(VISUAL QUESTION ANSWERING)**

*VĨNH ANH NGHIÊM QUÂN\*, NGUYỄN THỊ LAN ANH  
NGUYỄN LÊ TRUNG THÀNH, ĐINH THỊ DIỆU MINH  
Khoa Tin học, Trường Đại học Sư phạm, Đại học Huế*

*\*Email: vinhanhnghiemquan@dhsphue.edu.vn*

**Tóm tắt:** Bài toán trả lời câu hỏi trực quan (VQA) là bài toán kết hợp thị giác máy tính và xử lý ngôn ngữ tự nhiên. Trong bài báo này, chúng tôi trình bày các chiến lược xây dựng bộ dữ liệu VQA mới bằng tiếng Việt, Flickr8k-ViQA, bằng cách sinh tự động các cặp câu hỏi – câu trả lời từ những chú thích của bộ dữ liệu Flickr8k. Chúng tôi mô tả các công cụ đã sử dụng, phân tích những trở ngại gặp phải trong quá trình xây dựng và cách khắc phục, đồng thời đề xuất cách đánh giá các mô hình trên bộ dữ liệu này.

**Từ khóa:** Trả lời câu hỏi trực quan, sinh câu hỏi tự động, sinh câu hỏi trực quan.

1. MỞ ĐẦU

Được đề xuất từ năm 2015 [1], bài toán trả lời câu hỏi trực quan (Visual Question Answering) kết hợp hai lĩnh vực quan trọng của học máy (Machine Learning) là thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (Natural Language Processing). Dựa vào một hình ảnh và một câu hỏi ngôn ngữ tự nhiên về hình ảnh đó, mô hình phải đưa ra một câu trả lời tương ứng bằng ngôn ngữ tự nhiên.

Do câu hỏi có thể tập trung vào các vùng khác nhau của hình ảnh (tiền cảnh – foreground, hậu cảnh – background, ngữ cảnh – context hoặc các chi tiết khác) nên đòi hỏi mô hình vừa phải nhận biết được các bộ phận của ảnh, vừa phải kết hợp các bộ phận đó với câu hỏi và suy luận ra câu trả lời [1].

Các nghiên cứu bằng tiếng Việt về bài toán này hiện nay gặp một trở ngại chung, đó là thiếu các bộ dữ liệu huấn luyện và kiểm thử bằng tiếng Việt. Do khác biệt về ngữ pháp và văn phạm nên việc huấn luyên trên bộ dữ liệu tiếng Anh hay ngôn ngữ khác rồi áp dụng vào tiếng Việt là không khả thi.

Từ ý tưởng của Ren và cộng sự [2], trong bài báo này, chúng tôi đề xuất một thuật toán cho phép sinh ra các cặp câu hỏi – câu trả lời từ những câu mô tả hình ảnh, vốn tồn tại với số lượng lớn hơn. Từ đây, chúng tôi tạo ra một bộ dữ liệu mới cho bài toán VQA tiếng Việt: Flickr8k-ViQA.

Cấu trúc tiếp theo của bài viết như sau: mục 2 trình bày trình bày quá trình xây dựng bộ dữ liệu; mục 3 mô tả cách đánh giá các mô hình trên bộ dữ liệu và mục 4 là phần kết luận.

2. XÂY DỰNG BỘ DỮ LIỆU

**2.1. Cách tiếp cận**

Trong bài báo này, chúng tôi hướng đến xây dựng bộ dữ liệu VQA dựa trên bộ dữ liệu Flickr8k [3], gồm 8,000 hình ảnh và 40,000 câu chú thích (5 câu chú thích cho mỗi hình ảnh).

Khi xem xét các hướng tiếp cận để xây dựng bộ dữ liệu, cũng như Ren và cộng sự trong [2], chúng tôi cân nhắc các phương án sau:

1. Thuê nhân lực đặt câu hỏi – câu trả lời.
2. Sinh câu trả lời tự động từ các nhãn vật thể (object label) phát hiện được từ ảnh.
3. Sinh câu trả lời tự động từ các chú thích ảnh (image caption) được cung cấp sẵn trong bộ dữ liệu.

Phương án (1) đem lại chất lượng về mặt ngôn ngữ cao nhất cho dữ liệu. Tuy nhiên, dễ thấy cách này đòi hỏi quy mô lớn về kinh phí và thời gian dù chỉ áp dụng cho bộ dữ liệu kích cỡ trung bình như Flickr8k (8,000 hình ảnh và 40,000 câu chú thích).

Việc sinh tự động các cặp câu hỏi – câu trả lời dựa vào các chú thích ảnh có sẵn theo phương án (3) đem lại nhiều lợi ích: (i) Có thể tận dụng số lượng có sẵn các chú thích ảnh từ các bộ dữ liệu; (ii) Như minh họa trong Hình 1, các cặp câu hỏi sinh ra bằng cách này vẫn giữ được tính đa dạng về mặt ngôn ngữ của các chú thích, khiến chúng gần với câu hỏi do con người đặt ra hơn so với các câu hỏi sinh ra từ các nhãn (label) của các vật thể phát hiện trong ảnh; (iii) Việc sinh tự động giúp hạn chế tối đa việc cần đến sự can thiệp của con người trong quá trình xây dựng bộ dữ liệu, cho phép sinh ra số lượng lớn mẫu dữ liệu VQA trong thời gian ngắn với chi phí tối thiểu.

**Flickr8k 2218743570\_9d6614c51c**

**Nhãn:** cat, dog

**Câu hỏi đặt theo nhãn:** Which animal is in the picture? How many animals are in the picture?

**Chú thích #0:** A black cat hugging a white and gray dog.

**Câu hỏi đặt theo chú thích:** Which animal is hugging the white and gray dog? The black cat is hugging which animal? What color is the cat? What color is the cat hugging the white and gray dog? What color is the dog? What color is the dog being hugged by the black cat? How many cats are hugging the white and gray dog? How many dogs are being hugged by the black cat?

Hình 1. *Việc sinh tự động câu hỏi từ chú thích (caption) thay vì nhãn (label) giúp giữ được tính đa dạng về mặt ngôn ngữ của câu hỏi.*

Chú thích #0 là chú thích đầu tiên trong 5 chú thích của ảnh có id là 2218743570\_9d6614c51c.

**2.2. Công cụ**

Để có thể chuyển đổi các câu chú thích thành câu hỏi, trước tiên phải xác định được các thành phần ngữ pháp của câu chú thích đó thông qua việc gắn nhãn (part-of-speech tagging) cho các từ trong câu.

Chúng tôi xem xét sử dụng các thư viện: Underthesea Toolkit [7] của Under The Sea, Stanza [8] của Stanford NLP Group và Trankit [9] của Oregon NLP Group. Trong đó, Underthesea là bộ công cụ được thiết kế chuyên biệt cho tiếng Việt nên chỉ hỗ trợ nhãn từ loại chuyên biệt XPOS. Hai bộ công cụ còn lại có thể xử lý đa ngôn ngữ (trong đó có tiếng Việt) và hỗ trợ cả UPOS (nhãn từ loại thống nhất). Danh sách các nhãn này được liệt kê trong Bảng 1.

Qua một số thực nghiệm như minh họa trong Hình 2 và Hình 3, chúng tôi quyết định sử dụng Trankit vì đem lại kết quả gắn nhãn đầy đủ và chính xác hơn.

**Underthesea (XPOS)**

**Stanza (UPOS)**

**Trankit (UPOS)**



Hình 2. *Ví dụ về gắn nhãn của 3 bộ công cụ Underthesea, Stanza và Trankit.*

Dễ thấy Underthesea phát hiện tốt các danh từ chỉ loại “con” (Nc trong XPOS) và phụ từ “đang” (R trong XPOS), nhưng để lọt giới từ “qua” (E trong XPOS hay ADP trong UPOS) – cơ sở quan trọng để đặt câu hỏi về địa điểm (câu hỏi với từ nghi vấn “đâu”).

**Stanza (UPOS)**

**Trankit (UPOS)**



Hình 3. *Ví dụ về gắn nhãn cho cấu trúc câu phức tạp của Stanza và Trankit.*

Trong câu đa giới từ, Trankit vẫn phát hiện tốt và gắn nhãn đầy đủ cả hai giới từ lần lượt là “vào” và “bằng”, trong khi Stanza để lọt giới từ đầu tiên.

Bảng 1. *Tập nhãn từ loại thống nhất (Universal POS tags / UPOS) [4] và nhãn từ loại chuyên biệt cho từng ngôn ngữ (XPOS)*[6] *mà cụ thể ở đây là tiếng Việt*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **Tag** | **Description** |  | **STT** | **Tên** | **Chú thích** |
| 1 | ADJ | adjective |  | 1 | N | Danh từ |
| 2 | ADP | adposition |  | 2 | Np | Danh từ riêng |
| 3 | ADV | adverb |  | 3 | Nc | Danh từ chỉ loại |
| 4 | AUX | auxiliary |  | 4 | Nu | Danh từ đơn vị |
| 5 | CCONJ | coordinating conjunction |  | 5 | V | Động từ |
| 6 | DET | determiner |  | 6 | A | Tính từ |
| 7 | INTJ | interjection |  | 7 | P | Đại từ |
| 8 | NOUN | noun |  | 8 | L | Định từ |
| 9 | NUM | numeral |  | 9 | M | Số từ |
| 10 | PART | particle |  | 10 | R | Phụ từ |
| 11 | PRON | pronoun |  | 11 | E | Giới từ |
| 12 | PROPN | proper noun |  | 12 | C | Liên từ |
| 13 | PUNCT | punctuation |  | 13 | I | Thán từ |
| 14 | SCONJ | subordinating conjunction |  | 14 | T | Trợ từ, tiểu từ, từ tình thái |
| 15 | SYM | symbol |  | 15 | U | Từ đơn lẻ |
| 16 | VERB | verb |  | 16 | Y | Từ viết tắt |
| 17 | X | other |  | 17 | X | Các từ không phân loại được |

**2.3. Tiền xử lý**

Chúng tôi thực hiện các bước tiền xử lý như trong [2], nhưng với nhiều thay đổi cho phù hợp với bài toán tiếng Việt.

***2.3.1. Tách câu ghép thành câu đơn***

Trong [2], tác giả đề xuất phân tách câu ghép thành các câu bằng cách đơn giản là xác định vị trí các từ liên kết (conjunctive word, nhãn CCONJ hoặc SCONJ) như “and” (và) và “while” (trong khi), sau đó tách các mệnh đề tại vị trí đó thành các câu đơn.

Bảng 2. *Ví dụ về phân tách mệnh đề câu ghép thành các câu đơn*

|  |  |  |
| --- | --- | --- |
| **STT** | **Câu gốc** | **Sau khi xử lý** |
| 1 | Một đứa trẻ tập đánh bóng chày vào lưới **trong khi** hai người lớn đứng xem. | Một đứa trẻ tập đánh bóng chày vào lưới |
| Hai người lớn đứng xem. |
| 2 | Một cô gái châu Á đội mũ xanh **và** đeo tạp dề đang phục vụ đồ uống trên khay. | Một cô gái châu Á đội mũ xanh. |
| \* Đeo tạp dề đang phục vụ đồ uống trên khay. |

Ở ví dụ 2, liên từ được sử dụng để kết nối cụm động từ khiến việc tách câu không còn chính xác do đây vốn không phải câu ghép.

Như ví dụ trong Bảng 2, cách này không tính đến trường hợp các liên từ được sử dụng để kết nối các cụm danh từ / cụm động từ thay vì các mệnh đề. Để khắc phục nhược điểm này, chúng tôi chỉ tiến hành tách câu tại vị trí liên từ khi xác định được trước và sau liên từ đó là hai mệnh đề (có danh từ - nhãn NOUN - đứng trước động từ - nhãn VERB).

***2.3.2. Chuyển mạo từ xác định thành không xác định***

Ren và cộng sự [2] thay thế các mạo từ không xác định “a(n)” (nhãn DET) trong câu chú thích thành mạo từ xác định “the” khi chuyển thành câu hỏi như ví dụ 1 trong Bảng 3.

Trong trường hợp này, chúng tôi nhận thấy:

1. Mạo từ không xác định “a(n)” sẽ trở thành số từ (nhãn NUM) chỉ số ít “một” khi dịch qua tiếng Việt.
2. Trong tiếng Việt không phân biệt chặt chẽ danh từ xác định và không xác định bằng mạo từ như tiếng Anh.

Do đó, chúng tôi đề xuất: (i) lược bỏ số từ trong trường hợp số từ chỉ số ít số ít (ví dụ 2 trong Bảng 3) và (ii) thay thế số từ bằng định từ “những” trong trường hợp số từ chỉ số nhiều số nhiều (ví dụ 3 trong Bảng 3).

Bảng 3. *Ví dụ về chuyển đổi mạo từ trong tiếng Anh và số từ trong tiếng Việt khi chuyển từ câu chú thích (thể xác định) thành câu hỏi (thể nghi vấn)*

|  |  |  |
| --- | --- | --- |
| **STT** | **Câu chú thích** | **Câu hỏi** |
| 1 | **A** small yellow dog runs on a beach. | Where does **the** small yellow dog run? |
| 2 | **Một** con chó nhỏ màu vàng chạy trên bãi biển. | Con chó nhỏ màu vàng chạy trên đâu? |
| 3 | **Hai** con chó chơi trên bãi biển. | **Những** con chó chơi trên đâu? |

***2.3.3. Chuyển vị-wh***

“Wh” trong chuyển vị-wh xuất phát từ một điểm thú vị của tiếng Anh là đa phần các từ nghi vấn đều bắt đầu bằng “wh”: who, what, where, which, v.v.

Theo [5], đối với các ngôn ngữ nghịch đảo (cấu trúc câu hỏi ngược với câu xác định) như tiếng Anh và nhiều ngôn ngữ phương Tây khác, việc chuyển từ câu xác định thành câu hỏi phải trải qua một quy tắc biến hình là chuyển vị-wh (wh-movement) như ví dụ 1 trong Bảng 4. Đối với tiếng Anh, quá trình chuyển vị-wh có thể tóm tắt như sau:

1. Yếu tố được hỏi được thay thế bằng từ nghi vấn (wh-word).
2. Từ nghi vấn được di chuyển lên đầu câu (do đó phép chuyển vị-wh còn được biết với tên wh-fronting trong tiếng Anh).
3. Thêm trợ động từ (auxiliary verb) nếu chưa có sẵn và thực hiện phép đảo trật tự nếu từ nghi vấn không phải chủ ngữ.

Phép chuyển vị-wh trở nên khó xử lý khi cấu trúc câu trở nên phức tạp hơn như ví dụ 2 trong Bảng 4.

Tuy nhiên, trong tiếng Việt, các từ nghi vấn không bị chuyển vị mà thường giữ nguyên vị trí của chúng trong câu xác định (wh-in situ, tạm dịch là *wh tại vị*). Việc đảo ngữ cũng không diễn ra trong câu nghi vấn tiếng Việt. Do đó, khi chuyển đổi từ các câu chú thích (thể xác định) thành câu hỏi (thể nghi vấn), chỉ cần giữ nguyên câu trúc ban đầu và thay thế các yếu tố được hỏi bằng các từ nghi vấn như ví dụ 3 trong Bảng 4.

Bảng 4. *Minh họa về chuyển vị từ nghi vấn trong tiếng Anh và tiếng Việt*

|  |  |  |
| --- | --- | --- |
| **STT** | **Câu chú thích** | **Câu hỏi** |
| 1 | A grey dog is leaping over **a fallen tree**. | **What** is the grey dog leaping over? |
| 2 | A person eats takeout while watching **a small television**. | **What** does the person watch while eating takeout? |
| 3 | Một người ăn đồ ăn mang đi trong khi xem **một chiếc tivi nhỏ**. | Một người ăn đồ ăn mang đi trong khi xem **gì**? |

Trong ví dụ 2, phép chuyển vị-wh tương đối phức tạp hơn khi vị ngữ chứa hai cụm động từ. Cũng là câu này nhưng khi dịch thành tiếng Việt như trong ví dụ 3, việc chuyển đổi thành câu hỏi trở nên dễ dàng hơn rất nhiều do chỉ việc đưa từ nghi vấn “gì” vào thay thế yếu tố được hỏi ở cuối câu mà không cần thực hiện phép chuyển vị về đầu câu.

**2.4. Sinh câu hỏi tự động**

Trong [2], tác giả tập trung vào bốn dạng câu hỏi wh sau:

1. Câu hỏi *“what”*: Ý tưởng của Ren và cộng sự trong trường hợp này là duyệt câu chú thích đã tiền xử lý, xác định các danh từ chỉ gì (noun categories) bằng WordNet và thay thế chúng bằng “what” trước khi thực hiện phép chuyển vị. Do chưa có mạng từ tiếng Việt hoàn chỉnh tương đương như WordNet, chúng tôi tạm sử dụng danh sách danh từ tiếng Việt trích xuất từ dự án Mạng từ tiếng Việt và chỉ tập trung thay thế các danh từ chỉ người (mục noun.person) thành “ai” và danh từ chỉ động vật (mục noun.animal) thành “con gì”.
2. Câu hỏi *“how many”*: Như trong [2], chúng tôi tìm các số từ trong câu và thay thế chúng bằng “có bao nhiêu” nếu ở đầu câu và “bao nhiêu” nếu ở giữa câu. Để xác định số từ, chúng tôi chọn các từ có nhãn UPOS là NUM.
3. Câu hỏi *“what color”*: Ren và cộng sự tìm kiếm các tính từ chỉ màu sắc, sau đó tách riêng danh từ đi liền với tính từ đó để đặt câu hỏi “what color”, lược bỏ hoàn toàn nội dung còn lại của câu chú thích. Do trong Mạng từ tiếng Việt không có sẵn danh sách các tính từ chỉ màu sắc, chúng tôi dùng cách tiếp cận như trong mã nguồn của [2] là tự xây dựng danh sách này thông qua trích lọc tất cả các tính từ (nhãn ADJ) tìm thấy khi gắn nhãn cho bộ dữ liệu. Sau đó, chúng tôi thay thế tất cả tính từ nằm trong danh sách này bằng “màu gì”.
4. Câu hỏi *“where”*: Đối với trường hợp này, Ren và cộng sự [2] chỉ xét các cụm giới từ bắt đầu bằng giới từ “in” và loại bỏ các trường hợp nội dung sau đó chỉ trang phục (Vd: a man **in** black shirt), sau đó thay thế nội dung sau đó bằng “where”. Chúng tôi mở rộng danh sách các giới từ (nhãn ADP) bao gồm “trên”, “dưới”, “trong”, “vào”, “lên”, “trước”, “sau” và thay thế nội dung sau đó bằng “đâu”.

Ren và cộng sự [2] chỉ sử dụng chính yếu tố được hỏi làm câu trả lời (câu trả lời một từ duy nhất) như ví dụ 1 trong Bảng 5. Điều này giúp đưa bài toán VQA trở về bài toán phân lớp, với từ được dùng để trả lời chính là nhãn đúng của lớp.

Trong bài báo này, chúng tôi đề xuất thêm câu trả lời đầy đủ bằng cách đưa ngược yếu tố được hỏi vào thay thế từ nghi vấn trong câu hỏi như ví dụ 2 trong Bảng 5.

Bảng 5. *Câu trả lời một từ (one word answer) và câu trả lời đầy đủ*

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Câu chú thích** | **Câu hỏi** | **Câu trả lời** |
| 1 | A little girl in a **pink** dress going into a wooden cabin | **What color** is the dress? | pink |
| 2 | Một cô bé mặc váy **hồng** đi vào một căn nhà gỗ. | Cô bé mặc váy **màu gì** đi vào căn nhà gỗ? | hồng |
| Cô bé mặc váy **hồng** đi vào căn nhà gỗ. |

**2.5. Bộ dữ liệu flickr8k-viqa**

Để có được các chú thích bằng tiếng Việt, chúng tôi sử dụng phương thức dịch máy từ tiếng Anh sang tiếng Việt cho bộ dữ liệu Flickr8k.

Sau khi áp dụng các bước đã được đề cập ở các mục trước, chúng tôi xây dựng được bộ dữ liệu mới: Flickr8k-ViQA (**Flickr8k** – **Vi**etnamese **Q**uestion **A**nswering) với số lượng hơn 200,000 cặp câu hỏi – câu trả lời.

Bảng 6. *Thống kê sơ bộ bộ dữ liệu Flickr8k-ViQA*

|  |  |  |
| --- | --- | --- |
| LOẠI CÂU HỎI | SỐ LƯỢNG | % |
| AI | 29,792 | 14.87% |
| CON GÌ | 11,992 | 5.98% |
| BAO NHIÊU | 110,125 | 54.96% |
| MÀU GÌ | 20,747 | 10.35% |
| Ở ĐÂU | 27,728 | 13.84% |
| TỔNG CỘNG | 200,384 | 100.00% |

**Flickr8k 332045444\_583acaefc3**

**Chú thích #0:** (Tiếng Anh) A little white curly haired dog runs across the pavement with a stick in its mouth. (Tiếng Việt) Một con chó lông xoăn nhỏ màu trắng chạy trên vỉa hè với một cái que trong miệng.

**Câu hỏi:** 1. **Con gì** lông xoăn nhỏ màu trắng chạy trên vỉa hè với một cái que trong miệng? 2. **Có bao nhiêu** con chó lông xoăn nhỏ màu trắng chạy trên vỉa hè với một cái quen trong miệng? 3. Con chó lông xoăn nhỏ **màu gì** chạy trên vỉa hè với một cái quen trong miệng? 4. Con chó lông xoăn nhỏ màu trắng chạy trên **đâu**?

**Câu trả lời một từ:** 1. chó 2. một 3. trắng 4. vỉa\_hè

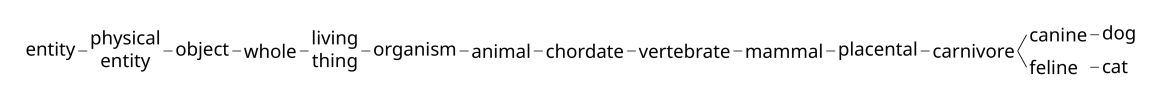
Hình 4. *Ví dụ một mẫu dữ liệu trong Flickr8k-ViQA.  
Từ một chú thích ban đầu có thể sinh ra 4 cặp câu hỏi – câu trả lời khác nhau.*

3. THANG ĐO

Vấn đề của bài toán VQA là làm sao để đánh giá xem câu trả lời của mô hình đã chính xác hay chưa. Trong [2], tác giả chỉ xét các câu trả lời một từ để đưa bài toán về bài toán phân lớp, sau đó sử dụng độ tương đồng Wu-Palmer (WUPS – **Wu**-**P**almer **S**imilarity) [11] để đánh giá câu trả lời:

Trong đó:

* s1 và s2 là các tập synset (tập khái niệm) lần lượt chứa w1 và w2
* depth là độ sâu của tập synset
* lcs (least common subsumer) hay lowest common ancestor của s1 và s2 là tập synset sâu nhất chứa hai tập s1 và s2



Hình 5. *Ví dụ về độ tương đồng Wu-Palmer với WordNet tiếng Anh.*

Trong trường hợp này, “cat” và “dog” có cùng độ sâu là 14, synset chung gần nhất của chúng là “carnivore” có độ sâu 12. Theo công thức trên, WUPS(“dog”,”cat)≈0.8571

Do độ đo này phụ thuộc chặt chẽ vào WordNet tiếng Anh và mạng từ tiếng Việt [10] vẫn chưa hoàn thiện (như đã đề cập ở các mục trước), chúng tôi đề xuất hai phương án đánh giá câu trả lời một từ như sau:

1. Dịch câu trả lời dự đoán của mô hình và câu trả lời đúng trong bộ dữ liệu ra tiếng Anh rồi tính độ tương đồng Wu-Palmer. Vì đây là các câu trả lời một từ nên việc dịch ra tiếng Anh sẽ không quá phức tạp so với việc dịch câu trả lời đầy đủ. Cách này cũng giúp tận dụng được WordNet tiếng Anh để tính WUPS.
2. Biểu diễn câu trả lời dự đoán của mô hình lẫn câu trả lời đúng trong bộ dữ liệu lên không gian vectơ (chẳng hạn Word2Vec), sau đó tính độ tương đồng của các vectơ này bằng công thức:

Trong đó A và B là biểu diễn vectơ (word embedding) của câu trả lời dự đoán của mô hình và câu trả lời đúng của bộ dữ liệu.

Cách này giúp tận dụng tính chất của các word embedding là các từ có chức năng ngữ nghĩa tương đồng nhau sẽ chỉ về các hướng giống nhau trong không gian vectơ. Do đó giá trị cosine của góc giữa hai vectơ này có thể được sử dụng để đại diện cho độ tương đồng về mặt ngữ nghĩa giữa hai từ.

Tuy nhiên, khác với độ tương đồng Wu-Palmer, độ tương đồng tính theo cách này phụ thuộc rất lớn vào mô hình được lựa chọn (CBOW – **C**ontinuos **B**ag **O**f **W**ords hay SkipGram) và các ngữ liệu văn bản (text corpora) được dùng để huấn luyện mô hình. Do đó, để đảm bảo tính thống nhất về kết quả đánh giá cũng như để tiện đối chiếu và so sánh, chúng tôi kiến nghị cần nêu rõ mô hình đã chọn và dữ liệu đã dùng để huấn luyện khi sử dụng độ tương đồng này.

Đối với câu trả lời hoàn chỉnh, chúng tôi đề xuất sử dụng các độ đo vẫn hay được sử dụng trong dịch máy (machine translation) như BLEU (**B**ilingua**l** **E**valuation **U**nderstudy) [12].

4. KẾT LUẬN

Qua bài báo này, chúng tôi đã xây dựng được bộ dữ liệu VQA bằng tiếng Việt mang tên Flickr8k-ViQA từ bộ dữ liệu Flickr8k. Cách sinh tự động các cặp câu hỏi – câu trả lời từ các câu chú thích trong bài báo này có thể được áp dụng cho các bộ dữ liệu tương tự. Chúng tôi cũng trình bày sơ bộ thang đo giúp đánh giá các mô hình sử dụng bộ dữ liệu này.

Do hạn chế về mặt thời gian, chúng tôi chưa cung cấp hiệu năng tối thiểu (baseline) để làm giá trị tham khảo khi xây dựng các mô hình. Việc so sánh, đối chiếu hiệu năng của các mô hình sử dụng bộ dữ liệu này là tiền đề quan trọng để phát triển bài toán VQA bằng tiếng Việt trong tương lai.

**Flickr8k 1093737381\_b313cd49ff**

**Chú thích #0:** (Tiếng Anh) A woman in a blue dress stands with a chalk portrait of two children playing by a stream. (Tiếng Việt) Một người phụ nữ mặc váy xanh đứng với bức chân dung bằng phấn của hai đứa trẻ đang chơi đùa bên dòng suối.

**Câu hỏi:** 1. **Có bao nhiêu** người phụ nữ mặc váy xanh đứng với bức chân dung bằng phấn của hai đứa trẻ đang chơi đùa bên dòng suối? 2. Người phụ nữ mặc váy xanh đứng với bức chân dung bằng phấn của **bao nhiêu** đứa trẻ đang chơi đùa bên dòng suối? 3. Người phụ nữ mặc váy **màu gì** đứng với bức chân dung bằng phấn của hai đứa trẻ đang chơi đùa bên dòng suối?

**Câu trả lời một từ:** 1. một 2. hai 3. xanh

Hình 6. *Một ví dụ khác trong Flickr8k-ViQA.  
Từ một chú thích ban đầu có thể sinh ra 4 cặp câu hỏi – câu trả lời khác nhau*

**TÀI LIỆU THAM KHẢO**

1. Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425-2433. 2015.
2. Ren, Mengye, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *arXiv preprint arXiv:1505.02074* (2015).
3. Hodosh, Micah, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013): 853-899.
4. Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.
5. Dũng Vũ (2003). *Tiếng Việt và Ngôn ngữ học hiện đại – Sơ khảo về cú pháp.* Stuttgart: VIET.
6. Thái, N. P., Lương, V. X., & Huyền, N. T. M (2008). Xây dựng Treebank tiếng Việt. *Kỷ yếu hội thảo ICTrda08-VLSP*.
7. Under The Sea (2021). Vietnamese NLP Research Group, Underthesea - Open-source Vietnamese Natural Language Process Toolkit.. [Online]. Available: https://github.com/undertheseanlp/underthesea [Access Mar 20, 2021].
8. Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
9. Nguyen, Minh, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen (2021). Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. *arXiv preprint arXiv:2101.03289*.
10. Asianwordnet (2018). Mạng từ tiếng Việt. [Online]. Available: http://viet.wordnet.vn/wnms/ (Access Apr 15 2018).
11. Wu, Zhibiao, and Martha Palmer (1994). Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.
12. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311-318.

**Title:** VISUAL QUESTION ANSWERING FOR VIETNAMESE: DATASET AND EVALUATION METRICS

**Abstract:** Visual Question Answering (VQA) combines both computer vision and natural language processing. In this paper, we present strategies for building a new Vietnamese VQA dataset, Flickr8k-ViQA, by automatically generating question-answer pairs from Flickr8k dataset annotations. We describe the tools used for the project, discuss the challenges we learned and solved during the process, and propose the metrics to evaluate models on this data set.

**Keywords:** Visual question answering, automatic question synthesis, visual question generation.