Báo cáo đồ án môn
CS406.P11 - Xử lý ảnh và ứng dụng

Đề tài: Image Captioning

Lớp CS406.P11
Sinh viên thực hiện: 21521465 - Trần Ngọc Thiện
Giảng viên hướng dẫn: thầy Cáp Phạm Đình Thăng

New Tab

Giới thiệu

Mô tả bài toán

Bài toán image captioning (tạo chú thích cho hình ảnh) là một lĩnh vực trong thị giác máy tính và xử lý ngôn ngữ tự nhiên. Đây là một bài toán phức tạp, đòi hỏi máy tính không chỉ hiểu được nội dung của hình ảnh mà còn có khả năng diễn đạt chúng bằng ngôn ngữ tự nhiên.

Mô tả bài toán

- Input:
  - Tập train: { (x, y)n }
    - x: hình ảnh
    - y: câu mô tả nội dung ảnh tương ứng
    - n: số mẫu dữ liệu trong tập train
  - X: ảnh đầu vào
- Output: Y: câu mô tả nội dung hình ảnh

Thách thức

- Nhiều cách diễn đạt
- Chi tiết phức tạp
- Ngữ cảnh
- Hình ảnh có nhiều đối tượng, các mối quan hệ phức tạp giữa các đối tượng
- Hình ảnh có đối tượng nhỏ hoặc bị che khuất

Ứng dụng

- Tìm kiếm hình ảnh dựa trên văn bản
- Hỗ trợ người khuyết tật
- Truyền thông xã hội
- Phân tích hình ảnh y tế

Phương pháp

Phương pháp CNN-LSTM

Sử dụng encoder (CNN) để xử lý ảnh và decoder (RNN, LSTM) để tạo caption

Phương pháp CNN-LSTM

Image encoder sử dụng InceptionNet v3

(a) Inception module, naïve version

(b) Inception module with dimension reductions

Phương pháp CNN-LSTM

Image encoder sử dụng InceptionNet v3

Phương pháp CNN-LSTM

Cấu trúc mô hình decoder

Phương pháp CNN-LSTM

Kết hợp đặc trưng hình ảnh và văn bản

Phương pháp CNN-LSTM

- Ưu điểm:
  - Kiến trúc dễ hiểu và triển khai.
  - Đã được chứng minh hiệu quả trên nhiều bài toán khác nhau.
- Nhược điểm:
  - Khó xử lý các chuỗi dài do vấn đề gradient vanishing/exploding.
  - Không tận dụng tốt thông tin toàn cục: RNN chỉ xử lý thông tin theo trình tự tuần tự, khó nắm bắt các mối quan hệ giữa các từ cách xa nhau trong câu.

Phương pháp transformer-based

Sử dụng mô hình vision transformer để encode ảnh

Phương pháp transformer-based

Sử dụng mô hình vision transformer để encode ảnh và GPT2-Vietnamese để tạo câu mô tả.

Phương pháp transformer-based

- Ưu điểm:
  - Xử lý song song
  - Cơ chế attention giúp Transformer nắm bắt các mối quan hệ giữa các từ cách xa nhau trong câu.
  - Hiệu suất cao
- Nhược điểm:
  - Phức tạp
  - Tốn tài nguyên tính toán

Đánh giá

Độ đo BLEU Score

- BLEU (Bilingual Evaluation Understudy) là một trong những metric phổ biến nhất để đánh giá chất lượng của các mô hình tạo văn bản.
- Tại sao sử dụng BLEU?
  - Đơn giản
  - Hiệu quả
  - Phổ biến
- Hạn chế của BLEU
  - Chỉ đo lường sự trùng lặp, không đánh giá được tính lưu loát, ý nghĩa của câu.
  - Không đánh giá sự đa dạng
  - Nhạy cảm với các từ hiếm

Độ đo BLEU Score

- Precision 1-gram

Target Sentence: The guard arrived late because ~~it was raining~~

Predicted Sentence: The guard arrived late because of the rain

- Precision 2-gram

Target Sentence: The guard arrived late because it was raining

Predicted Sentence: The guard arrived late because of the rain

Độ đo BLEU Score

$$
\begin{aligned}
Geometric\ Average\ Precision\ (N) &= exp(\sum_{n=1}^{N} w_n\ log\ p_n) \\
&= \prod_{n=1}^{N} {p_n}^{w_n} \\
&= (p_1)^{\frac{1}{4}} \cdot (p_2)^{\frac{1}{4}} \cdot (p_3)^{\frac{1}{4}} \cdot (p_4)^{\frac{1}{4}}
\end{aligned}
$$

Độ đo BLEU Score

$$Brevity\ Penalty = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c <= r \end{cases}$$

- c là predicted length = số lượng từ có trong predicted sentence
- r là target length = số lượng từ có trong target sentence

Độ đo BLEU Score

$$Bleu\ (N) = Brevity\ Penalty \cdot Geometric\ Average\ Precision\ Scores\ (N)$$

Bleu Score có thể tính toán cho nhiều giá trị N khác nhau. Cụ thể trong trường hợp N = 4.

- BLEU-1 sử dụng unigram Precision Score.
- BLEU-2 sử dụng geometric average of unigram and bigram Precision Score.
- BLEU-3 sử dụng geometric average of unigram, bigram, and trigram Precision Score.
- Và cứ tiếp tục như thế cho đến BLEU-N.

New Tab
Dataset

Dataset

- Sử dụng UIT-ViIC
- Bộ dữ liệu có 3850 hình ảnh chủ đề thể thao lấy từ MS-COCO với 19250 câu caption bằng tiếng Việt (5 câu cho mỗi hình ảnh)

- Một cậu bé đang chống đẩy trên quả banh.
- Một cậu bé đang chống đầu lên quả bóng đá hít đất.
- Một cậu bé đang hít đất đầu trên quả banh trên cỏ.
- Một cậu bé đang hít đất cùng quả bóng và một cậu bé đang quan sát.
- Một cậu bé đang chống đẩy trên một quả bóng đá.

Kết quả

Kết quả

LSTM: BLEU 4 = 0.2978

Transformer: BLEU 4 = 0.3875

Kết quả

- LSTM: Một người đàn ông đang vung gậy để đánh quả bóng chày.
  - BLEU 4: 0.6761
- Transformer: Người đàn ông đang vung gậy bóng chày đánh trả bóng .
  - BLEU 4: 0.1968
- Label caption:
  - Những người đàn ông đang luyện tập bóng chày ở trong khung lưới .
  - Cậu bé đang vung gậy để đánh quả bóng chày ở trước tấm lưới .
  - Cậu bé đang vung gậy để đánh quả bóng chày ở trên sân .
  - Những người đàn ông đang chơi bóng chày ở trong khung lưới .
  - Một cầu thủ bóng chày đang dùng gậy để đánh bóng ở trước tâm lưới .

Kết quả

- LSTM: Một cầu thủ ném bóng đang chuẩn bị ném quả bóng ở trong tay.
  - BLEU 4: 0.0543
- Transformer: Người đàn ông đang cầm gậy bóng chày đứng trên sân .
  - BLEU 4: 0.5747
- Label caption:
  - Cầu thủ bóng chày đang cầm gậy bóng chày sẵn sàng đỡ bóng .
  - Cầu thủ bóng chày đang cầm gậy thi đấu trên sân .
  - Người đàn ông áo trắng đang chơi bóng chày trên sân .
  - Cầu thủ bóng chày cầm gậy đứng trên sân chuẩn bị đỡ bóng .
  - Cầu thủ bóng chày đang vung gậy bóng chày đánh trả bóng .

Kết quả

- LSTM: Các cầu thủ bóng đá đang thi đấu trên sân.
  - 5.63 giây
- Transformer: Các cầu thủ bóng đá đang thi đấu ở trên sân.
  - 31.98 giây

Kết quả

- LSTM: Một trận thi đấu bóng đá đang diễn ra ở trên sân.
  - 5.98 giây
- Transformer: Một trận thi đấu bóng bầu dục nam đang diễn ra ở trên sân.
  - 33.50 giây

New Tab

Hướng phát triển

Hướng phát triển

- Cải thiện chất lượng chú thích (đa dạng, chi tiết, theo ngữ cảnh)
- Xử lý trên dữ liệu đa dạng và phức tạp hơn
- Video captioning
- Multimodal learning: Kết hợp thông tin từ nhiều modal khác nhau (ví dụ: văn bản, âm thanh) để tạo ra các mô tả phong phú hơn.

New Tab
Demo

Cảm ơn thầy và các bạn đã lắng nghe