AI VIETNAM
Seminar

AI CITY CHALLENGE 2024
Track 2

Xuan Khai Trinh

Year 2024

AI VIETNAM
Seminar

Outline

- Track 2 Problem
- Data Descriptions and Annotations
- EDA Data: BDD-PC-5K (External)
- EDA Data: WTS (Internal)
- Metrics
- Solution

AI VIETNAM
Seminar

Outline

- Track 2 Problem
- Data Descriptions and Annotations
- EDA Data: BDD-PC-5K (External)
- EDA Data: WTS (Internal)
- Metrics
- Solution

AI VIETNAM
Seminar

Track 2 Problem

Challenge Track 2: Traffic Safety Description and Analysis

Track Problem: **video captioning** về các tình huống an toàn giao thông liên quan đến tai nạn cho người đi bộ.

Recognition: The pedestrian, unaware of the vehicle approaching, was standing diagonally to the right in front of the vehicle. His body was perpendicular to the vehicle and was facing in the direction of travel. Slowly looking around...

Action: He appeared to notice the vehicle and was aware of its presence. In front of him, he planned to continue going straight ahead, despite traveling in a car lane. His speed was slow, matching his cautious actions.

AI VIETNAM
Seminar

Outline

- Track 2 Problem
- **Data Descriptions and Annotations**
- EDA Data: BDD-PC-5K (External)
- EDA Data: WTS (Internal)
- Metrics
- Solution

AI VIETNAM
Seminar

Data Descriptions and Annotations

WTS cung cấp số lượng video lớn nhất với các mô tả video dài chi tiết cùng với thông tin không gian 3D.

Datasets	Videos (total)	Type	Domain	Captions num.	Avg. caption len.	3D Gaze	Year
MSVD [4]	1,970	scene	open	80,380	7.14	-	2011
TACoS [16]	7,206	scene	cooking	18,227	8.27	-	2013
MPII-MD [17]	68,327	scene	movie	68,375	11.05	-	2015
M-VAD [15]	46,589	scene	movie	46,589	12.44	-	2015
MSR-VTT [24]	507,502	scene	open	200,000	9.27	-	2016
Charades [18]	9,848	scene	daily indoor	25,032	23.91	-	2015
Charades-Ego [19]	7,860	scene	daily indoor	14,039	26.30	-	2016
TGIF [11]	125,782	scene	open	125,781	11.28	-	2016
ActivityNet Caps. [6]	19,994	instance	human activity	72,976	14.72	-	2017
VATEX [22]	34,991	scene	open	349,910	15.25	-	2019
HowTo100M [12]	139,668,840	scene	instruction	139,668,840	4.16	-	2019
TRECVID-VTT'20 [1]	9,185	scene	open	28,183	18.90	-	2020
Gaze360 [9]	9	-	indoor	-	-	Y	2019
Dynamic3D [13]	17	-	indoor + one outdoor	-	-	Y	2022
GFIE [8]	n/a	-	indoor	-	-	Y	2023
BDD-X [10]	6,984	scene	traffic + outdoor	26,228	14.5	-	2018
WTS	**6,061 (1,200+4,861)**	**instance**	**traffic + outdoor**	**49,860**	**58.7**	**Y**	**2023**

AI VIETNAM
Seminar

Data Descriptions and Annotations

Video data format:

```
videos
├── train
│   ├── 20230707_12_SN17_T1  ##scenario index
│   │   ├── overhead_view    ##different overhead view about the scenario
│   │   │   ├── 20230707_12_SN17_T1_Camera1_0.mp4
│   │   │   ├── 20230707_12_SN17_T1_Camera2_3.mp4
│   │   │   ├── 20230707_12_SN17_T1_Camera3_1.mp4
│   │   │   └── 20230707_12_SN17_T1_Camera4_2.mp4
│   │   └── vehicle_view  ##vehicle ego-view about the scenario
│   │       └── 20230707_12_SN17_T1_vehicle_view.mp4
│   ├── 20230707_15_SY4_T1
│   │   ├── overhead_view
│   │   │   ├── 20230707_15_SY4_T1_Camera1_0.mp4
│   │   │   ├── 20230707_15_SY4_T1_Camera2_1.mp4
│   │   │   └── 20230707_15_SY4_T1_Camera3_2.mp4
│   │   └── vehicle_view
│   │       └── 20230707_15_SY4_T1_vehicle_view.mp4
...
```

1200 video với overhead view và vehicle view

```
external/
└── BDD_PC_5K
    ├── videos
    │   ├── train
    │   │   ├── video1004.mp4
    │   │   ├── video1006.mp4
    │   │   ├── video1009.mp4
    │   │   ├── video100.mp4
    │   │   ├── video1015.mp4
...
```

4861 video với vehicle view

AI VIETNAM
Seminar

Data Descriptions and Annotations

Annotation format:

```
annotations
├─caption/
  ├── train
  │   ├── 20230707_12_SN17_T1 ##scenario index
  │   │   ├── overhead_view
  │   │   │   └── 20230707_12_SN17_T1_caption.json
  │   │   └── vehicle_view
  │   │       └── 20230707_12_SN17_T1_caption.json
  │   ├── 20230707_15_SY4_T1
  │   │   ├── overhead_view
  │   │   │   └── 20230707_15_SY4_T1_caption.json
  │   │   └── vehicle_view
  │   │       └── 20230707_15_SY4_T1_caption.json
...
```

```
{
    "id": 722,
    "overhead_videos": [ ## caption related videos
        "20230707_8_SN46_T1_Camera1_0.mp4",
        "20230707_8_SN46_T1_Camera2_1.mp4",
        "20230707_8_SN46_T1_Camera2_2.mp4",
        "20230707_8_SN46_T1_Camera3_3.mp4"
    ],
    "event_phase": [
        {
            "labels": [
                "4"  ##segment number
            ],
            "caption_pedestrian": "The pedestrian stands still on the left si
            "caption_vehicle": "The vehicle was positioned diagonally to ..."
            "start_time": "39.395",  ##start time of the segment in seconds,
            "end_time": "44.663"     ##end time of the segment in seconds
        },
...
```

AI VIETNAM
Seminar

Data Descriptions and Annotations

BBox format:

```
{
    "annotations": [
        {
            "image_id": 904, ## frame ID
            "bbox": [
                1004.4933333333333, ## x_min
                163.28666666666666, ## y_min
                12.946666666666667, ## width
                11.713333333333333  ## height
            ],
            "auto-generated": false,  ##human annotated frame
            "phase_number": "0"
        },
        {
            "image_id": 905,
            "bbox": [
                1007.1933333333333,
                162.20666666666668,
                12.946666666666667,
                11.713333333333333
            ],
            "auto-generated": true,  ##generated bbox annotation for the frame
            "phase_number": "0"
        },
...
```

AI VIETNAM
Seminar

Data Descriptions and Annotations

Submission format: Mỗi video submit 5 phase và mỗi phase bao gồm 2 loại caption: pedestrian và vehicle.

```
"video3334": [  ##scneario index for multiple view situations OR video name for single view "BD
    {
        "labels": [  ##segment number, this is known information will be given
            "4"
        ],
        "caption_pedestrian": "",  ##caption regarding pedestrian
        "caption_vehicle": ""      ##caption regarding vehicle
    },
    {
        "labels": [
            "3"
        ],
        "caption_pedestrian": "",
        "caption_vehicle": ""
    },
```

Lưu ý:

- Submit file JSON bao gồm tất cả video trong tập test.
- Submit đủ cả ”caption_pedestrian” and “caption_vehicle”

AI VIETNAM
Seminar

Outline

- Track 2 Problem
- Data Descriptions and Annotations
- **EDA Data: BDD-PC-5K (External)**
- EDA Data: WTS (Internal)
- Metrics
- Solution

AI VIETNAM
Seminar

EDA Data: BDD-PC-5K (External)

Phase: 0 [location] [attention] [behaviour] [context attribute]

Pedestrian caption: The pedestrian, a male in his 30s, was standing still on a main road in an urban area. He was wearing a gray T-shirt and black short pants. The weather was cloudy with bright visibility, and the road surface was dry and level. The pedestrian's body was oriented in the same direction as the vehicle that approached him. He was positioned diagonally to the right, in front of the vehicle. Despite being unaware of the vehicle's presence, he closely watched the parked vehicle in his line of sight. The relative distance between the pedestrian and the vehicle was far. The road had a one-way traffic flow with three lanes and sidewalks on both sides. It was a weekday, and the pedestrian seemed to have no awareness of the usual traffic volume on this road.

Vehicle caption: The vehicle was moving at a constant speed of 10km/h. It was positioned behind a pedestrian and was quite far away from them. The vehicle had a clear view of the pedestrian. It was going straight ahead without any change in direction. The environment conditions indicated that the pedestrian was a male in his 30s with a height of 160 cm. He was wearing a gray T-shirt and black short pants. The event took place in an urban area on a weekday. The weather was cloudy but the brightness was bright. The road surface was dry and level, made of asphalt. The traffic volume was usual on the main road that had one-way traffic with three lanes. Sidewalks were present on both sides of the road.

EDA Data: BDD-PC-5K (External)

AI VIETNAM
Seminar

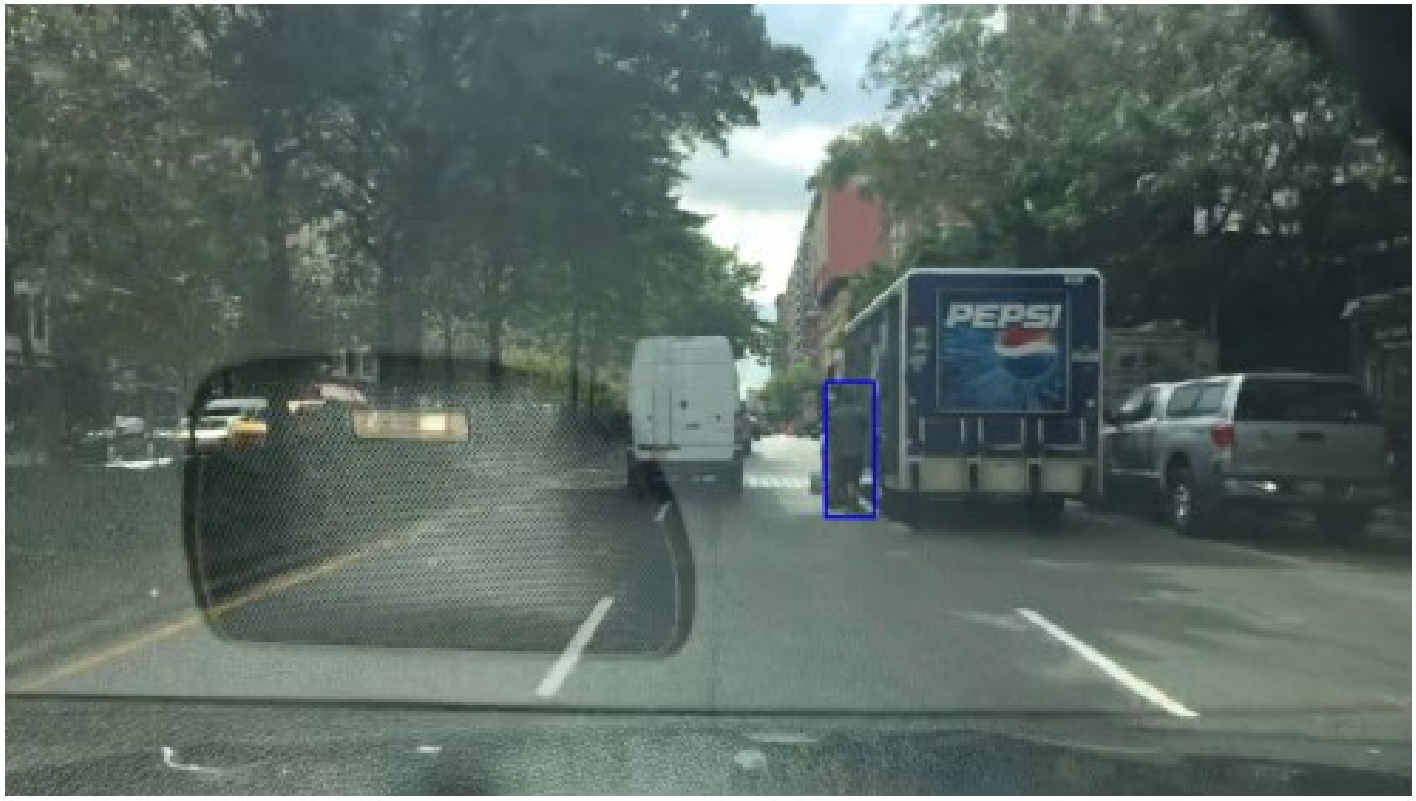
EDA Data: BDD-PC-5K (External)

Phase: 1 [location] [attention] [behaviour] [context attribute]

Pedes caption: A man in his 30s, with a height of 160 cm, dressed in a gray T-shirt and black shorts, is standing still diagonally to the right in front of a vehicle on a bright, cloudy weekday. The vehicle is traveling on a dry, level asphalt road, part of a main road with one-way traffic and three lanes. The pedestrian's body is oriented in the same direction as the vehicle, with his line of sight focused ahead in the direction of travel. Despite being far from the vehicle, he is closely watching and is almost aware of its presence. The surroundings are urban, with both sides of the road having sidewalks. The pedestrian seems to be waiting or observing something, possibly getting ready to cross the road. The traffic volume is usual, and the road surface conditions are favorable for safe movement. Overall, it appears to be a calm and routine situation in which the pedestrian and the vehicle are momentarily sharing the road space.

Vehicle caption: The vehicle is moving at a constant speed of 10km/h behind a pedestrian. The pedestrian is far from the vehicle and is visible within the vehicle's field of view. The vehicle is going straight ahead on a main road with one way and three lanes. The vehicle is in an urban environment with usual traffic volume on a dry and level asphalt road surface. The pedestrian, a male in his 30s, is wearing a gray T-shirt and black short pants. It is a weekday, and the weather is cloudy with bright brightness. Both sides of the road have sidewalks.

EDA Data: BDD-PC-5K (External)



AI VIETNAM
Seminar

Outline

- Track 2 Problem
- Data Descriptions and Annotations
- EDA Data: BDD-PC-5K (External)
- **EDA Data: WTS (Internal)**
- Metrics
- Solution

EDA Data: WTS (Internal)

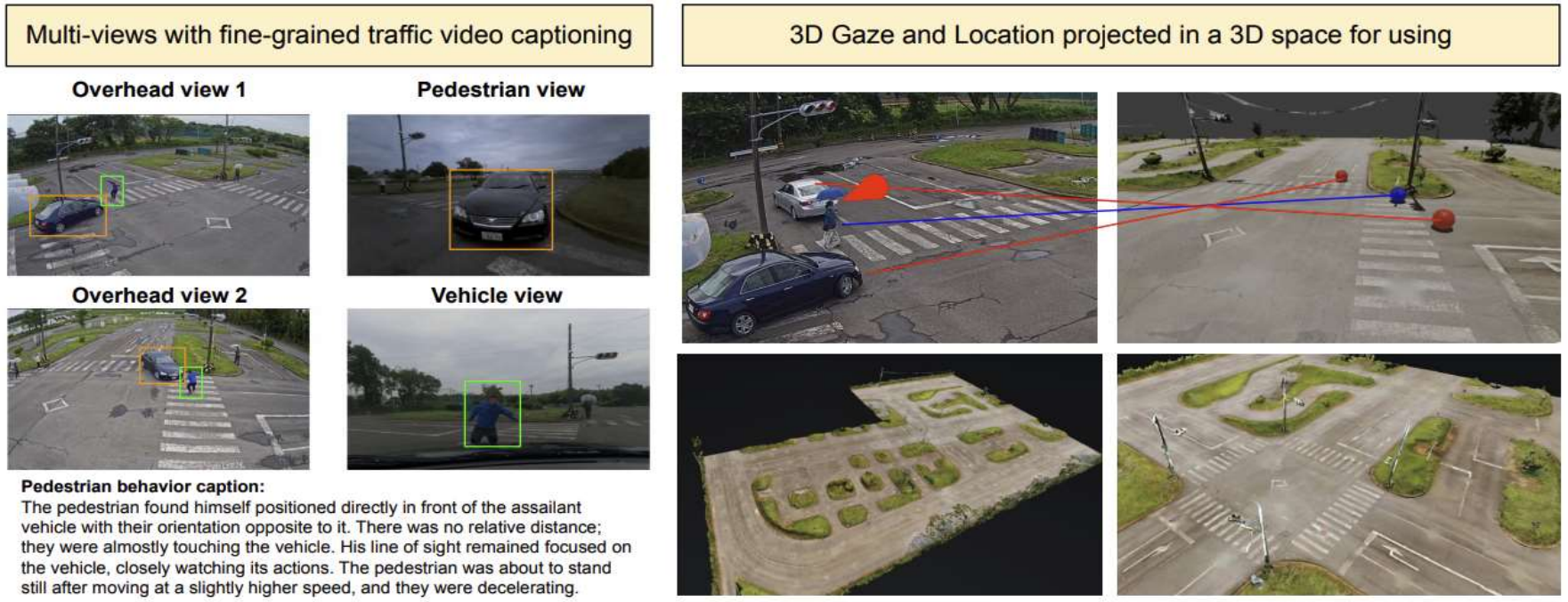
Bộ internal của WTS dataset sẽ cung cấp thêm overhead view - số lượng view tùy thuộc vào video dao động từ 1 tới 4.

EDA Data: WTS (Internal)

Bộ internal của WTS dataset sẽ cung cấp thêm overhead view - số lượng view tùy thuộc vào video dao động từ 1 tới 4.

EDA Data: WTS (Internal)

Bên cạnh đó để thuận tiện cho quá trình phân tích vị trí giữa xe và người đi đường, bộ dataset cũng cung cấp 3D Gaze (chưa công bố).



AI VIETNAM
Seminar

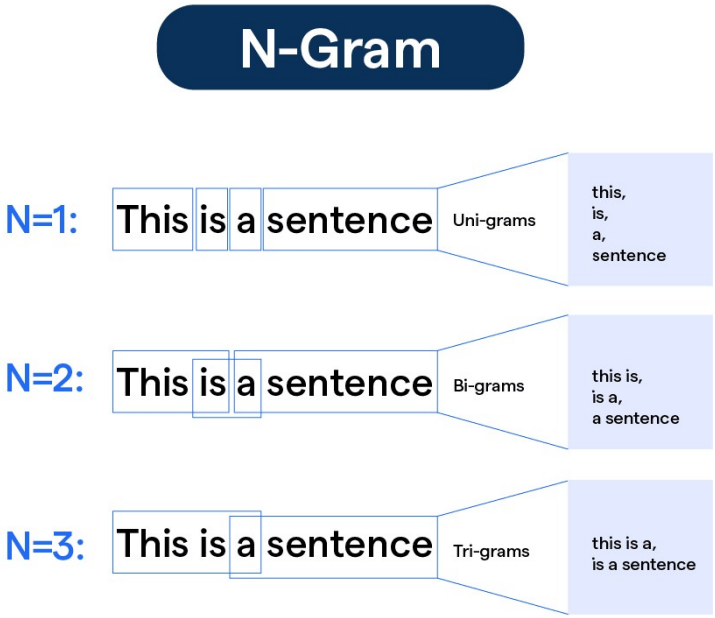
Outline

- Track 2 Problem
- Data Descriptions and Annotations
- EDA Data: BDD-PC-5K (External)
- EDA Data: WTS (Internal)
- **Metrics**
- Solution

Metrics

Metric 1: BLEU-4

- N-gram: các chuỗi gồm n kí tự (hoặc từ) liên tiếp trong bộ văn bản mẫu.



Metrics

Metric 1: BLEU-4

- Các bước để tính BLEU-4:
  + Bước 1: Tokenize đầu vào thành từ/cụm từ.
  + Bước 2: Tính Clipped Precision: $CP_n = \frac{Clip(\text{số lượng } n-gram \text{ dự đoán đúng})}{\text{tổng số lượng } n-gram \text{ trong câu dự đoán}}$

Generated cation

Target cation

So sánh giữa Clipped Precision và Precision:

$$CP_1 = \frac{Clip(\text{số lượng } 1-gram \text{ dự đoán đúng})}{\text{tổng số lượng } 1-gram \text{ trong câu dự đoán}} = \frac{1}{6}$$

$$P_1 = \frac{\text{số lượng } 1-gram \text{ dự đoán đúng}}{\text{tổng số lượng } 1-gram \text{ trong câu dự đoán}} = \frac{6}{6} = 1$$

Clipped Precision giúp giải quyết vấn đề lặp từ trong câu mô tả.

Metrics

Metric 1: BLEU-4

+ Bước 3: Weighted Geometric Mean Precision:

$$\text{Weighted Geometric Mean Precision}_N = \prod_{n=1}^{N} CP_n^{w_n}$$

+ Bước 4: Brevity Penalty: penalty cho câu mô tả có độ dài ngắn hơn câu mục tiêu:

$$\text{Brevity Penalty} = \begin{cases} 1, & \text{nếu } c > r \\ e^{(1-\frac{r}{c})}, & \text{nếu } c \leq r \end{cases}$$

+ Bước 5: BLEU score:

$$\text{BLEU}_N = \text{Brevity Penalty} \times \text{Weighted Geometric Mean Precision}_N$$

Metrics

Metric 2: METEOR

- Các bước để tính METEOR:

+ Bước 1: Xác định tập hợp các ánh xạ giữa các unigram của câu mô tả và câu mục tiêu.

+ Bước 2: Precision: $P = \frac{số\ lượng\ uni-gram\ dự\ đoán\ đúng}{tổng\ số\ lượng\ uni-gram\ trong\ câu\ dự\ đoán}$

+ Bước 3: Recall: $R = \frac{số\ lượng\ uni-gram\ dự\ đoán\ đúng}{tổng\ số\ lượng\ uni-gram\ trong\ câu\ mục\ tiêu}$

+ Bước 4: F-score: $F_{mean} = \frac{10PR}{R+9P}$

Metrics

Metric 2: METEOR

+ Bước 5: Chunk Penalty:

$$p = 0.5(\frac{c}{u_m})^3$$

trong đó: c là số lượng chunks (tập hợp các unigram liền kề) và u_m là số lượng unigram đã được map trong câu mục tiêu.

Lưu ý: để tính chunk penalty, ta cần tìm cách map unigram để số lượng chunks là ít nhất.

2 cách map chunks khác nhau

+ Bước 6: METEOR score:

$$M = F_{mean}(1 - p)$$

Metrics

Metric 3: ROUGE-L

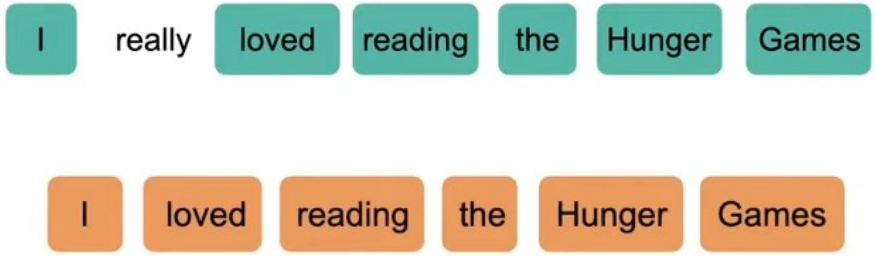
- Các bước để tính ROUGE-L:

+ Bước 1: Rouge-L Precision: $P = \frac{LCS(gen, ref)}{tổng\ số\ lượng\ uni-gram\ trong\ câu\ dự\ đoán}$

+ Bước 2: Rouge-L Recall: $R = \frac{LCS(gen, ref)}{tổng\ số\ lượng\ uni-gram\ trong\ câu\ mục\ tiêu}$

+ Bước 3: Rouge-L F1-Score: $F_1 = 2(\frac{P \times R}{P+R})$

Trong đó LCS là dãy con chung dài nhất (không yêu cầu các từ phải liên tiếp nhưng vẫn theo thứ tự) giữa câu mô tả và câu mục tiêu.



Metrics

Metric 4: CIDEr

- Metric CIDEr cải tiến so với các metric đã được đề cập ở việc nó có thể tận dụng nhiều câu mô tả mục tiêu khác nhau của cùng 1 bức ảnh.
- Các bước để tính CIDEr:
  + Bước 1: Áp dụng kĩ thuật stemming (tối giản một từ thành từ gốc của nó).
    Ví dụ: “fishes”, “fishing”, “fished” → ”fish”

  + Bước 2: Gọi n-gram ω_k là tập hợp của một hoặc nhiều từ. Ω là bộ vocabulary bao gồm tất cả n-grams. Giả sử ta cần đánh giá câu mô tả dự đoán c_i với tập hợp những câu mô tả mục tiêu $S_i = \{s_{i1}, \ldots, s_{im}\}$ của ảnh I_i. Ta cần tính trọng số TF-IDF:

Số lần n-gram ω_k xuất hiện trong câu s_{ij}

Nếu n-gram ω_k xuất hiện trong câu mô tả mục tiêu trên các ảnh trong bộ dataset thì cho bằng 1

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

Tổng số n-grams trong câu s_{ij}

Term Frequency: đặt trọng số cao nếu n-gram xuất hiện thường xuyên trong câu s_{ij}

Inverse Document Frequency: giảm trọng nếu n-gram xuất hiện thường xuyên trên toàn bộ dataset

Metrics

Metric 4: CIDEr

+ Bước 3: Tính CIDEr Score cho từng n-gram có độ dài n:

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

trong đó: $g^n(c_i)$ là vector có trọng số $g_k(c_i)$ của các n-gram ω_k có độ dài n.

+ Bước 4: Tính trung bình CIDEr Score cho tất cả n-gram:

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

trong đó: ta có thể chọn $w_n = \frac{1}{N}$ và $N = 4$

AI VIETNAM
Seminar

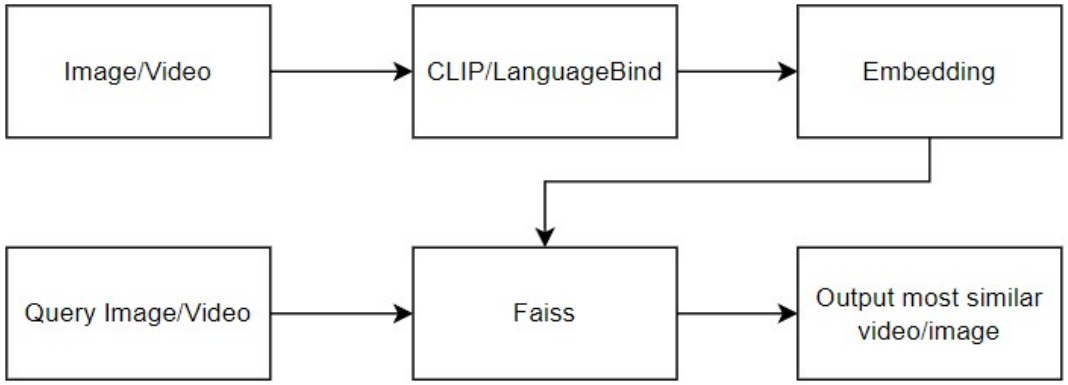
Outline

- Track 2 Problem
- Data Descriptions and Annotations
- EDA Data: BDD-PC-5K (External)
- EDA Data: WTS (Internal)
- Metrics
- Solution

Solution

Solution 1: Image hoặc Video Retrieval

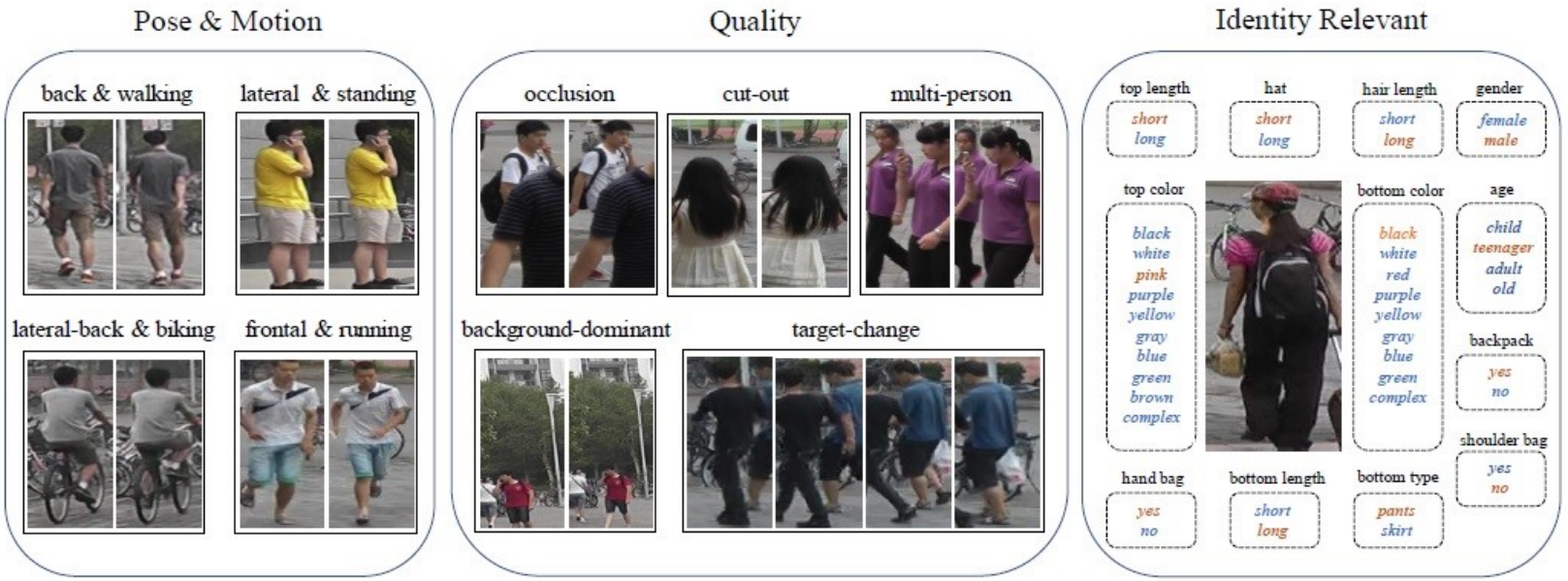
- Sử dụng CLIP hoặc LanguageBind để lấy embedding của video hoặc ảnh
- Tạo Faiss index để lưu trữ vector thuận tiện cho quá trình retrieval
- Thực hiện video-video hoặc image-image retrieval để lấy ra caption gần giống nhất video cần caption.



Solution

Solution 2: Kết hợp với VTF-PAR

- Một trong những điểm yếu của phương pháp số 1 là không sử dụng thông tin vị trí của người cần caption.
- Do đó ta có thể sử dụng VTF-PAR để trích xuất một số thông tin về người để kết hợp hỗ trợ quá trình retrieval



Solution

Solution 3: Zero shot Multimodal LLM với few shot LLM để rewrite lại kết quả

- Ở phương pháp này chúng ta sẽ áp dụng kĩ thuật zero shot cho 1 số multimodal LLM (LLaVA) để tạo ra prompt cho từng segment.
- Sau đó kết hợp với retrieval để thực hiện few shot LLM hỗ trợ rewrite để sinh ra kết quả cuối cùng.