Connecting Language and Vision for Natural Language-Based Vehicle Retrieval

Shuai Bai[1] Zhedong Zheng[2] Xiaohan Wang[3] Junyang Lin[1]
Zhu Zhang[1] Chang Zhou[1] Hongxia Yang[1] Yi Yang[2]
[1]DAMO Academy, Alibaba Group, [2]ReLER Lab, University of Technology Sydney, [3] Zhejiang University

Motivation

1. There is one gap between two modalities, i.e., natural language (NL) descriptions, and vehicle track images. How to construct a **shared semantic representation space** for both images and texts ?

2. Compared with image-based vehicle ReID, natural language-based vehicle retrieval provides more details on the behavioral and environmental information. How to effectively leverage this **spatio-temporal information**?

3. The number of training data is limited, and there exists noise in the training set. How to use data augmentation to improve the **generalization ability** of the learned model?

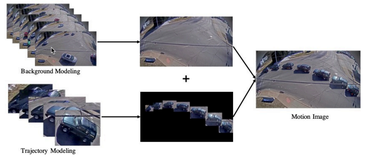
Contribution

- We explore different optimization objectives, including **Symmetric InfoNCE Loss** and **Instance Loss**, to construct a consistent representation space.
- We propose a simple and effective method to **model the motion and background**. A **dual stream** structure is adopted to combine the local and global information.
- We propose a **robust natural language-based vehicle search system** for smart city applications, which arrives **the 1th place** .

Method

Data Augmentation

The data augmentation strategies are applied to strengthen the key information, which include the motion modeling and description augmentation.

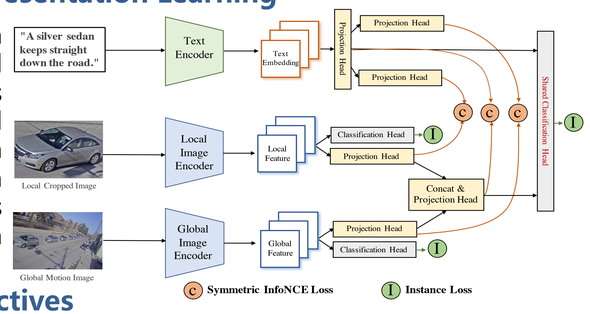


"A mid-sided blue sedan **goes straight** through an **intersection** behind a **blue vehicle**."
"A black sedan **keeping straight** down the street followed by another **black vehicle**."
"A black sedan **goes down** the straight after a **blue sedan**."

" A mid-sided blue sedan. A mid-sided blue sedan **goes straight** through ……. "
" A black sedan. A black sedan **keeping straight** down the street followed ……"
" A black sedan. A black sedan **goes down** the straight after a **blue sedan.** "
" A mid-sided blue sedan. A black sedan. A black sedan "

Cross-Modal Representation Learning

We adopt a dual-stream structure with the local cropped vehicle images and the global background , which learns more position information as well as the environment from motion images.



Optimization Objectives

We explore different optimization objectives, including Symmetric InfoNCE Loss and Instance Loss.

- **Symmetric InfoNCE Loss**

$$\mathcal{L}_{t2i} = \frac{1}{M}\sum_{i=1}^{M} -\log \frac{exp(cos(z_{img,i}, z_{text,i})/\tau)}{\sum_{j=1}^{M} exp(cos(z_{img,j}, z_{text,i})/\tau)} \mathcal{L}_{i2t} = \frac{1}{M}\sum_{i=1}^{M} -\log \frac{exp(cos(z_{img,i}, z_{text,i})/\tau)}{\sum_{j=1}^{M} exp(cos(z_{img,i}, z_{text,j})/\tau)}$$

- **Symmetric Instance Loss**

$$\mathcal{L}_t = -log(W_{shared}\, z_t) \qquad \mathcal{L}_i = -log(W_{shared}\, z_i)$$

- **Loss Function**

$$\mathcal{L} = \lambda_1 \mathcal{L}_{t2i} + \lambda_2 \mathcal{L}_{i2t} + \lambda_3 (\mathcal{L}_i + \mathcal{L}_t)$$

Experiment

Ablation Studies

Method	Performance					
Baseline	✓	✓	✓	✓	✓	✓
Instance Loss		✓	✓	✓	✓	✓
Motion Image			✓	✓	✓	✓
NLP Augmentation				✓	✓	✓
Large Size& Model					✓	✓
Ensemble						✓
MRR(%)	8.25	9.65	13.21	14.56	19.27	20.77

Competition Results

Rank	Team Name	MRR
1	**Alibaba-UTS (Ours)**	**0.1869**
2	TimeLab	0.1613
3	SBUK	0.1594
4	SNLP	0.1571
5	HUST	0.1564

Quantitative Results

Conclusion

To connect the vision and language modalities, we jointly train the state-of-the-art vision model and transformer-based language model with the **symmetric InfoNCE loss and Instance loss**. Further, we design a **two-stream** architecture to incorporate both **local details and global information of vehicles**, and apply the **text augmentation** technique, i.e., backtranslation, to enhance the model robustness. The proposed system has achieved **18.69% MRR** accuracy and arrived **the first place** in the natural language-based vehicle retrieval track of the 5th AICity Challenge.

Repositories : https://github.com/ShuaiBai623/AIC2021-T5-CLV https://github.com/layumi/NLP-AICity2021 https://github.com/layumi/Vehicle_reID-Collection