All You Can Embed: Natural Language based Vehicle Retrieval with Spatio-Temporal Transformers

Carmelo Scribano, Davide Sapienza, Giorgia Franchini, Micaela Verucchi, Marko Bertogna
Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia, Modena (Italy)

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA — High Performance Real Time Lab

CVPR VIRTUAL JUNE 19-25

The AI City Challenge Track 5: Contribution

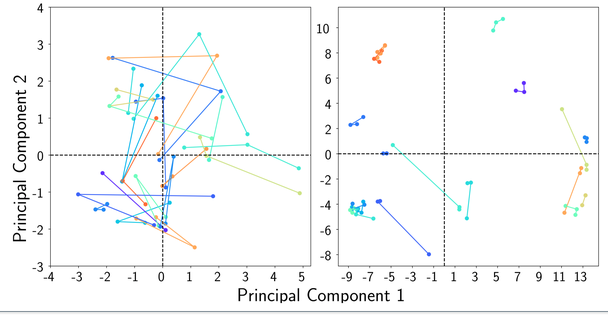
All You Can Embed (**AYCE**), a modular solution to correlate single-vehicle tracking sequences with natural language. The main building blocks of the proposed architecture are (i) BERT to provide an embedding of the textual descriptions, (ii) a convolutional backbone along with a Transformer model to embed the visual information, (iii) a new loss function designed for the track.

Natutal Language Branch (BERT Finetuning)

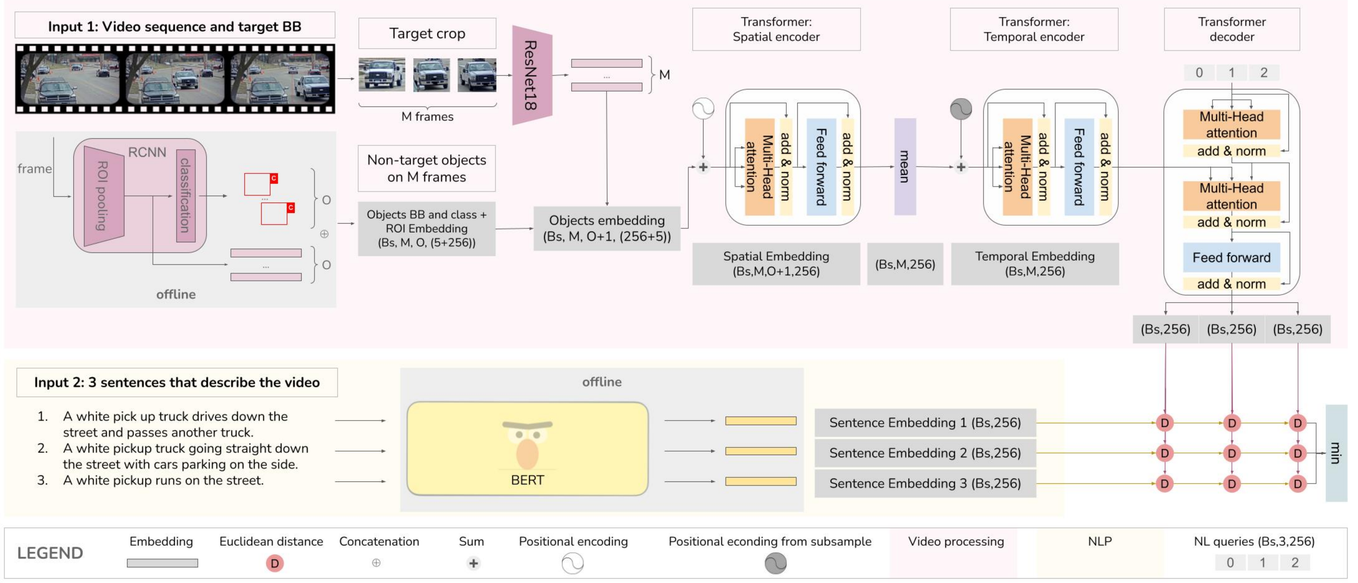
POSITIVE
"A grey sedan turning right."

ANCHOR
"A silver car is driving in the right lane around a curve"

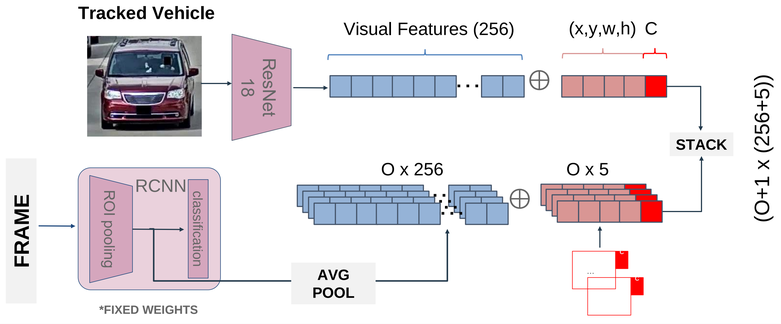
NEGATIVE
"A red SUV runs down the street alongside parked cars."



Architecture Overview



Object Embeddings



The objective function

$$\mathcal{L}_{AYCE}(\boldsymbol{A},\boldsymbol{P},\boldsymbol{N}) = \boldsymbol{TL}(\boldsymbol{A},\boldsymbol{P},\boldsymbol{N}) + \frac{\mathbf{1}}{\boldsymbol{Bs}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{Bs}} \boldsymbol{\beta}\cdot\boldsymbol{\Phi}\left(\boldsymbol{A}^{\boldsymbol{i}},\boldsymbol{P}^{\boldsymbol{i}}\right)$$

Where:

$$\Phi\left(A^i,P^i\right) = \min\left(d\left(A^i_m,P^i_n\right)\right) \quad m,n \in \{1,2,3\}$$

TL is defined as the standard Triplet Loss, the distance measure used in TL is the average of all distance pairs between Language (3) and Visual (3) outputs.

Results and Repository

Rank	Team	MRR
1	Alibaba	0.1869
2	TimeLab	0.1613
3	SBUK	0.1594
11	UNIMORE	0.1078

https://github.com/cscribano/AYCE_2021