CVPR VIRTUAL JUNE 19-25

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

AI City Challenge 2021

Carmelo Scribano[1,2], Davide Sapienza[1,2], Giorgia Franchini[1,3]
Micaela Verucchi[1] and Marko Bertogna[1]

[1]University of Modena and Reggio Emilia, [2]University of Parma, [3]University of Ferrara

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA High Performance Real Time Lab UNIVERSITÀ DI PARMA Università degli Studi di Ferrara

Problem statement

“Natural language (NL) description offers another useful way to specify vehicle track queries. In this new Challenge Track, participating teams will perform vehicle retrieval given single-camera tracks and corresponding NL descriptions of the targets.”

- AI City Challenge 2021 – Track5

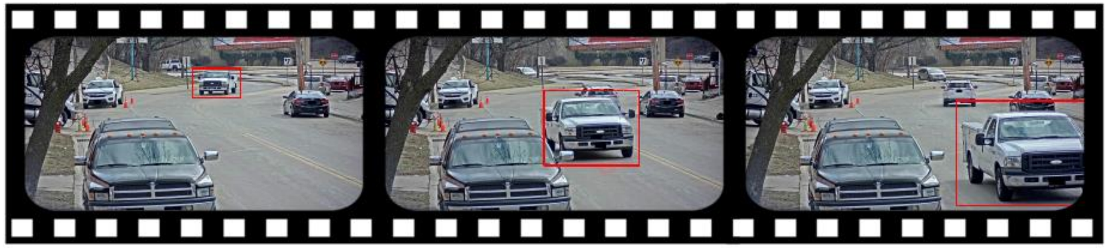
GOAL:

- Matching Single-Vehicle Tracking Sequences with the corresponding Natural Language Descriptions.

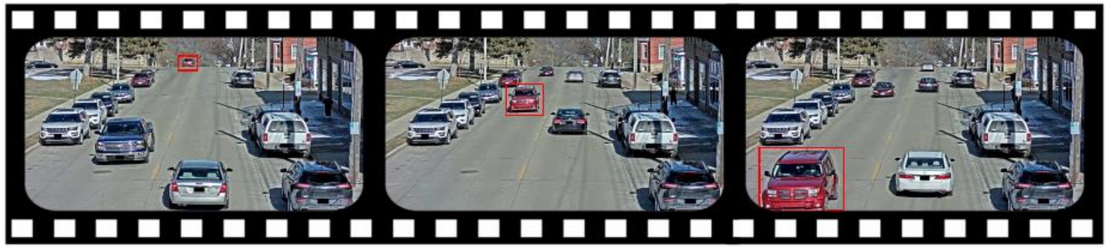
Related tasks:

- Image and Video retrieval
- Multi-Modal video Understanding

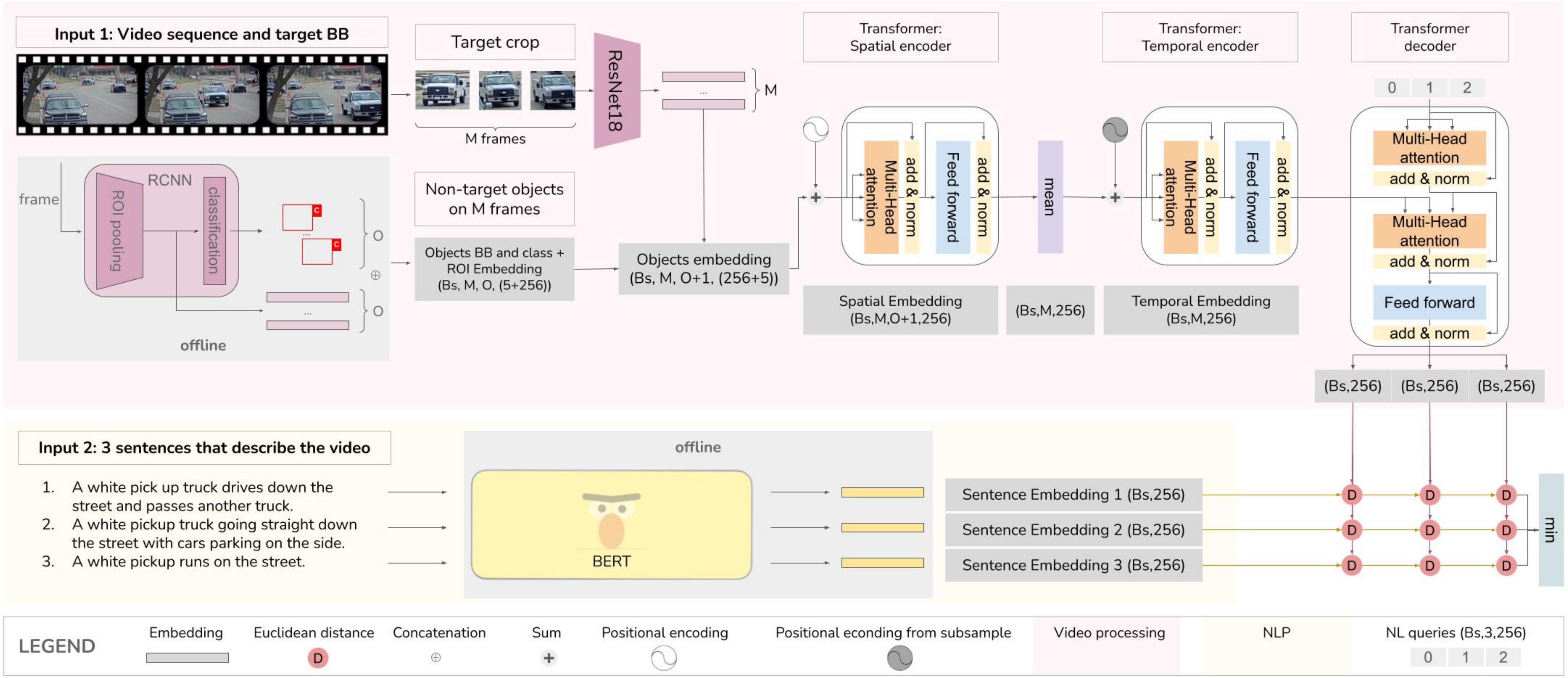
1. A white pick up truck drives down the street and passes another truck.
2. A white pickup truck going straight down the street with cars parking on the side.
3. A white pickup runs on the street.



1. A red SUV runs down the street alongside parked cars.
2. Red SUV keeps straight followed by a maroon car.
3. A red SUV runs down the road followed by a black vehicle.



Model Overview



Input 1: Video sequence and target BB

Target crop

M frames

ResNet18

M

frame

RCNN

ROI pooling

classification

O

O

offline

Non-target objects on M frames

Objects BB and class + ROI Embedding (Bs, M, O, (5+256))

Objects embedding (Bs, M, O+1, (256+5))

Transformer: Spatial encoder

Multi-Head attention

add & norm

Feed forward

add & norm

Spatial Embedding (Bs,M,O+1,256)

mean

(Bs,M,256)

Transformer: Temporal encoder

Multi-Head attention

add & norm

Feed forward

add & norm

Temporal Embedding (Bs,M,256)

Transformer decoder

0 1 2

Multi-Head attention

add & norm

Multi-Head attention

add & norm

Feed forward

add & norm

(Bs,256) (Bs,256) (Bs,256)

Input 2: 3 sentences that describe the video

1. A white pick up truck drives down the street and passes another truck.
2. A white pickup truck going straight down the street with cars parking on the side.
3. A white pickup runs on the street.

offline

BERT

Sentence Embedding 1 (Bs,256)

Sentence Embedding 2 (Bs,256)

Sentence Embedding 3 (Bs,256)

D D D

D D D

D D D

min

LEGEND

Embedding

Euclidean distance D

Concatenation ⊕

Sum +

Positional encoding

Positional econding from subsample

Video processing

NLP

NL queries (Bs,3,256) 0 1 2

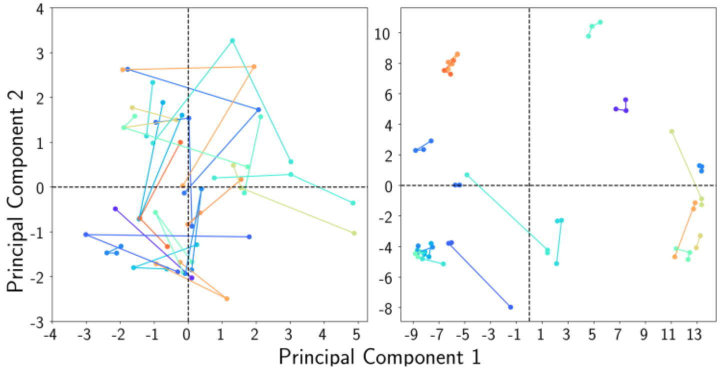
Natural Language Branch

OBJECTIVE:

$$\begin{cases} d\left(T_1^i, T_2^i\right) \ll d\left(T_1^i, T_2^j\right) \\ d\left(T_1^i, T_2^i\right) \approx 0 \end{cases} \quad \forall i : i \neq j$$

Where:

$$T_a^i = BERT\left(t_a^i\right) \qquad d(u,v) = 1 + \frac{u^T\, v}{||u||||v||} \in [0,2]$$



BEFORE:

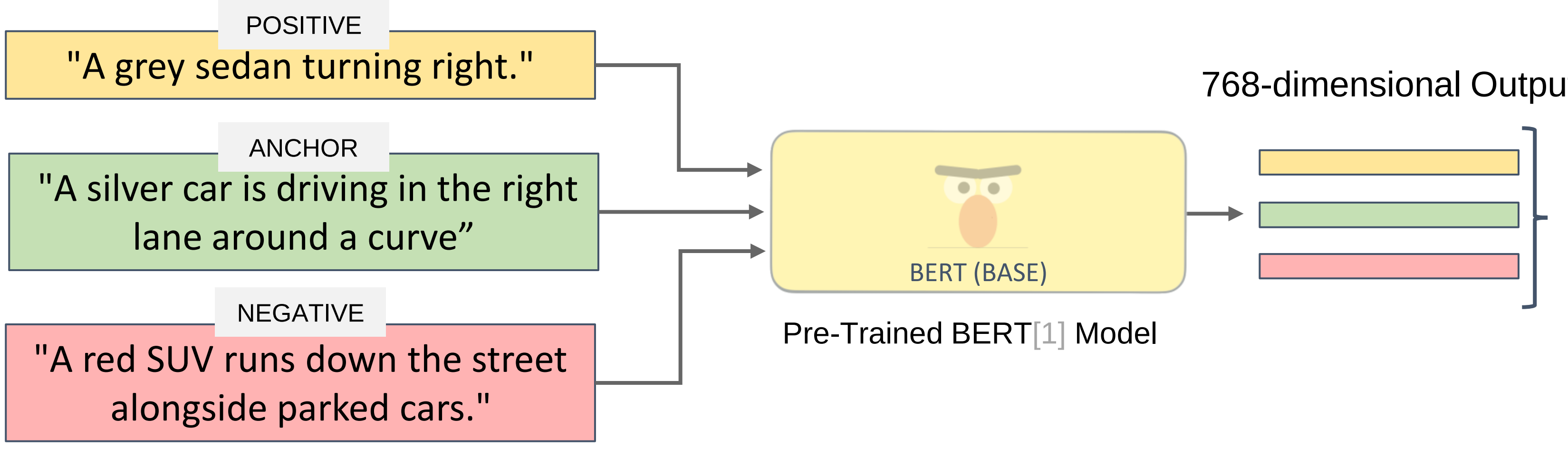
$$mean(d_{INTER}) = 0.1703$$

$$mean(d_{INTRA}) = 0.1899$$

AFTER:

$$mean(d_{INTER}) = \mathbf{0.2089}$$

$$mean(d_{INTRA}) = \mathbf{0.6140}$$


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."
BERT (BASE)
768-dimensional Output

Pre-Trained BERT[1] Model

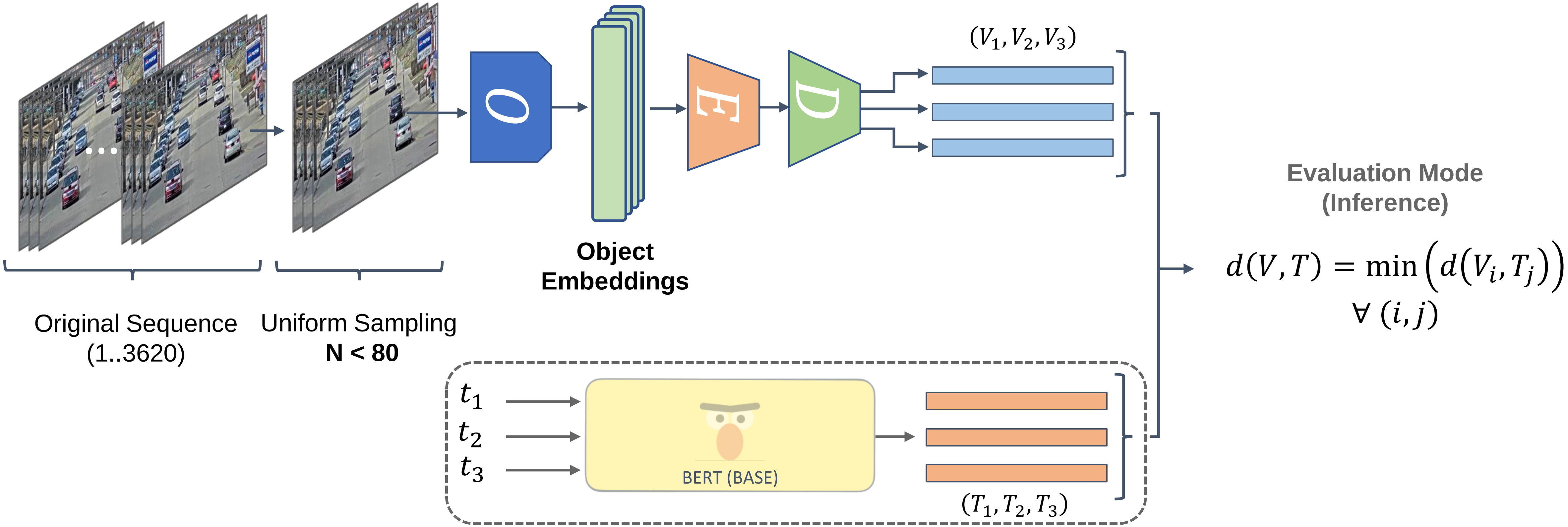
$$TL = \frac{1}{Bs}\sum_{i=1}^{Bs} \max\left(0, d\left(A^i, P^i\right) - d\left(A^i, N^i\right) + m\right)$$

Triplet Loss

Randomly chosen (A,P,N) Triple

[1]: Bert:Pre-training of deep bidirectional transformers for language understanding

Visual Branch



Original Sequence
(1..3620)

Uniform Sampling
N < 80

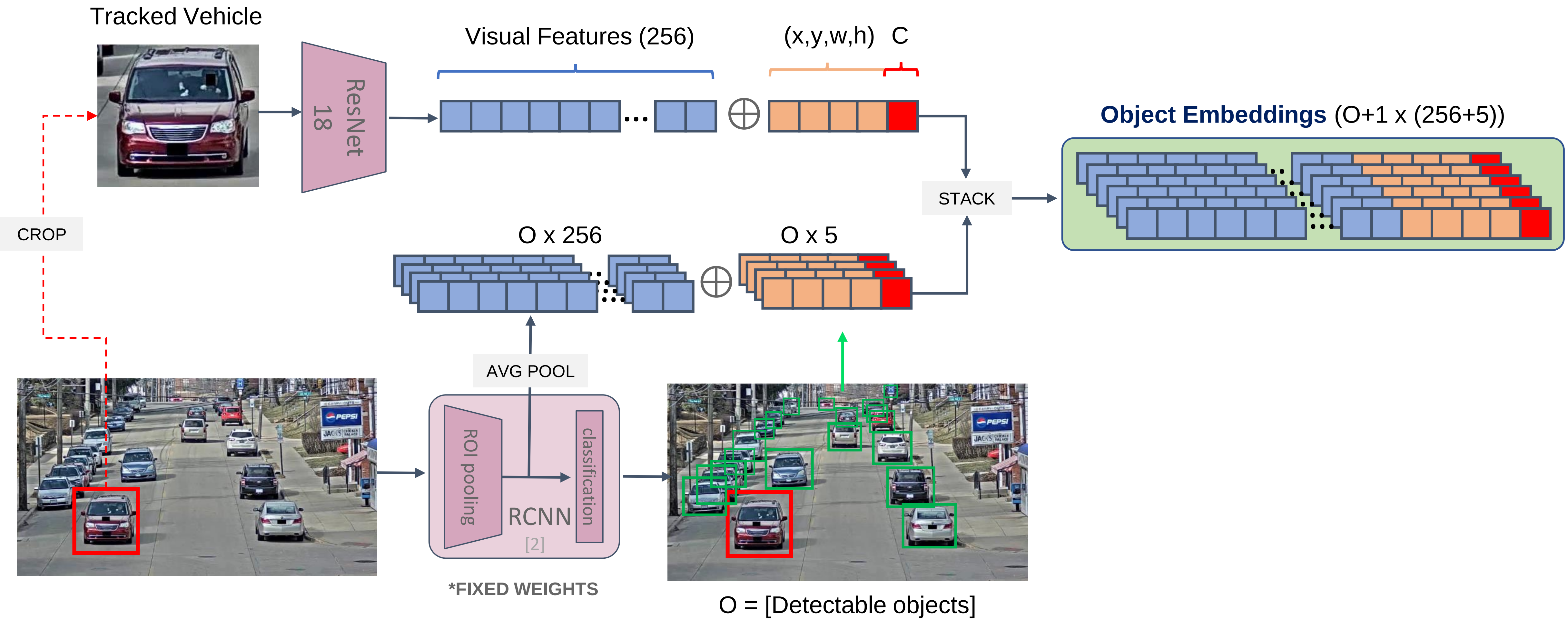
Object Embeddings

Evaluation Mode (Inference)

$$d(V,T) = \min\left(d\left(V_i, T_j\right)\right)$$
$$\forall\ (i,j)$$

Language Branch
*After Finetuning

Object Embeddings



Tracked Vehicle

ResNet 18

Visual Features (256)

(x,y,w,h) C

Object Embeddings (O+1 x (256+5))

STACK

CROP

O x 256

O x 5

AVG POOL

ROI pooling

RCNN [2]

classification

*FIXED WEIGHTS

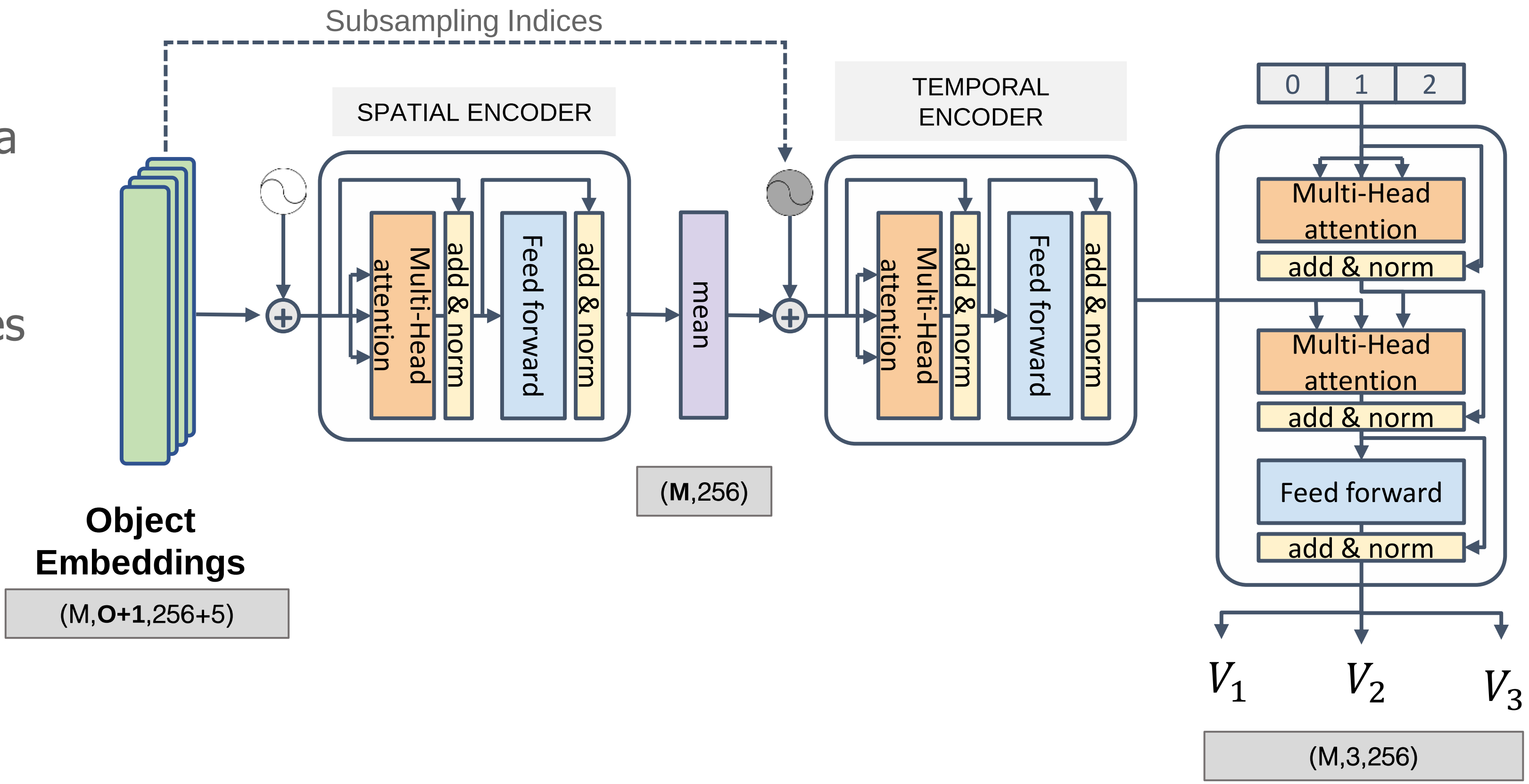
O = [Detectable objects]

[2]: Faster r-cnn: Towards real-time object detection with region proposal networks

Spatio-Temporal Transformers

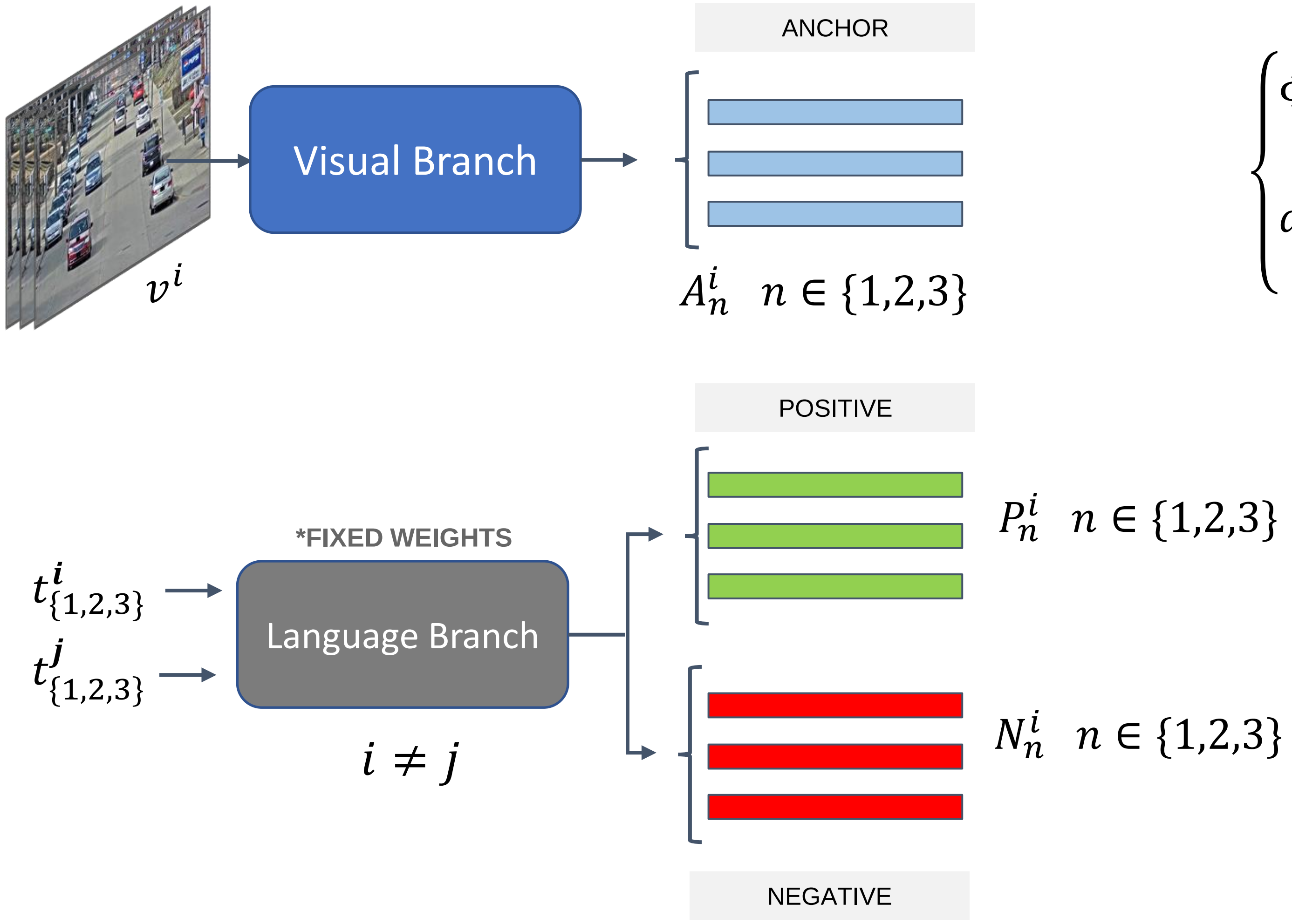
RATIONALE:

- The Spatial Encoder provides a comprehensive frame-level representation
- The Temporal Encoder encodes the sequential meaning of the depicted scene.



[3]: Attention is All you Need

Optimization



ANCHOR

Visual Branch

v^i

$A_n^i \quad n \in \{1,2,3\}$

POSITIVE

*FIXED WEIGHTS

$t^{\boldsymbol{i}}_{\{1,2,3\}}$

$t^{\boldsymbol{j}}_{\{1,2,3\}}$

Language Branch

$i \neq j$

$P_n^i \quad n \in \{1,2,3\}$

$N_n^i \quad n \in \{1,2,3\}$

NEGATIVE

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

$$\begin{cases} \Phi(A,P) = \min\left(d\left(A_{m,}P_n\right)\right) \quad m,n \in \{1,2,3\} \\ d_{TL}(A,[P|N]) = \dfrac{1}{9}\displaystyle\sum_{m=1}^{3}\sum_{n=1}^{3} \|A_m - [P|N]_n\|^2 \end{cases}$$

RESULTS

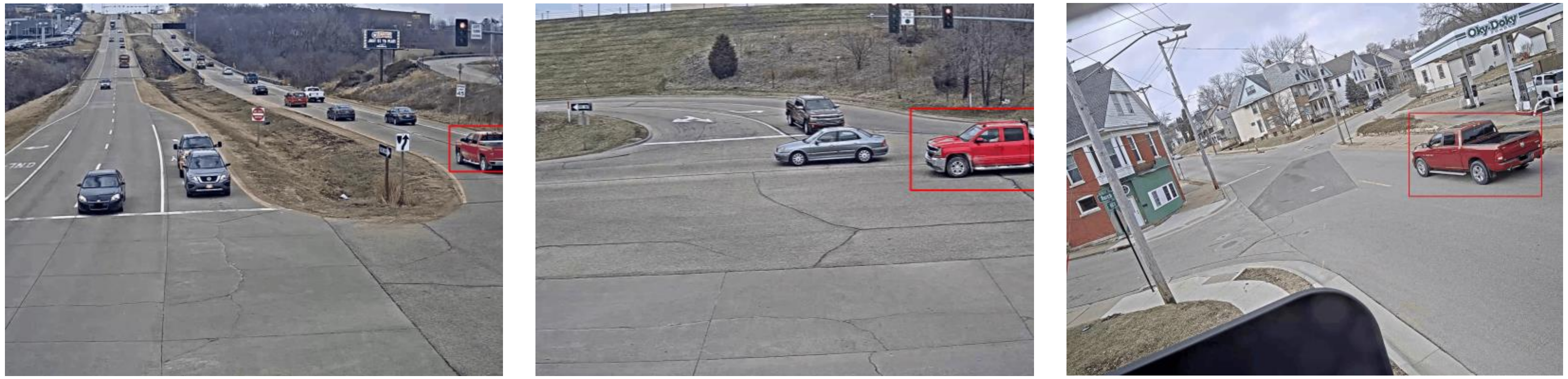
MODEL	MRR	RE@5	RE@10
BASELINE	0.0269	0.0264	0.0491
OURS (BEST – 11°)	**0.1078**	0.1321	0.2491
Alibaba-UTS-ZJU (*1*°)	0.1869	-	-
Sun Asterisk (4°)	0.1571	-	-
Modulab (10°)	0. 1195	-	-

Qualitative Results

"A red pickup drives straight down a *highway.“*

"A red pickup truck runs down the street.“

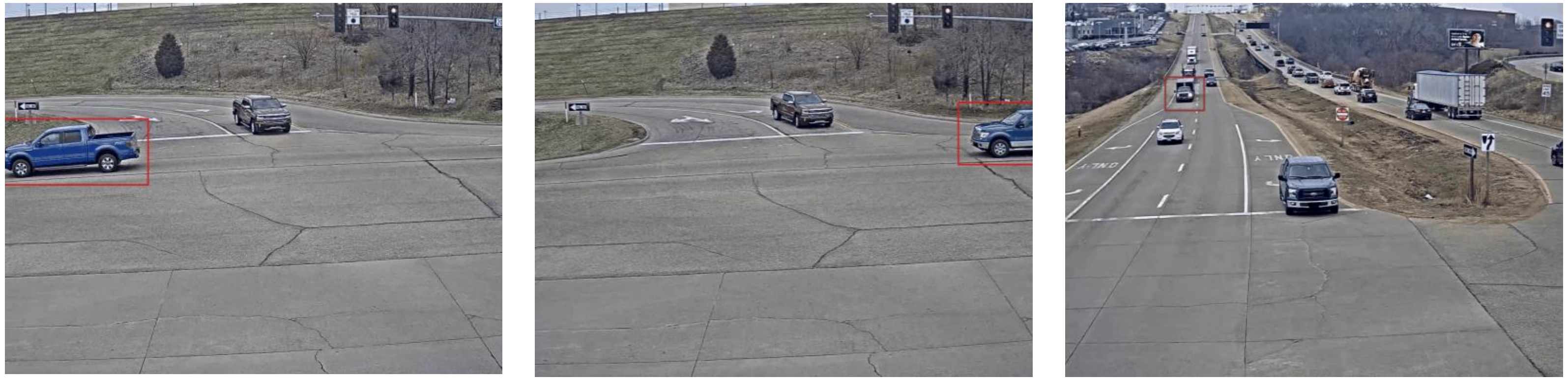
"A red pickup following straight other three car”



"A blue pickup runs down the street."

"A blue pickup truck going straight down the street passing an intersection."

"A blue truck runs down the street.“



"A black sedan crossing an intersection."

"A midsize black sedan goes straight through the intersection."

"A black Sedan runs down the street."