Transformer Reasoning Network for Image-Text Matching and Retrieval

http://aimir.isti.cnr.it

Nicola Messina, Fabrizio Falchi, Andrea Esuli, Giuseppe Amato

nicola.messina@isti.cnr.it

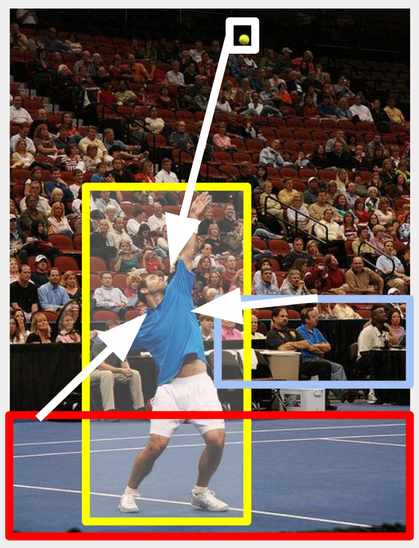
SCALABLE VISUAL TEXTUAL RETRIEVAL

Query: "Player number 8 kicked the soccer ball with his foot."

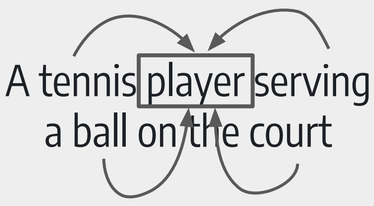
CHALLENGES

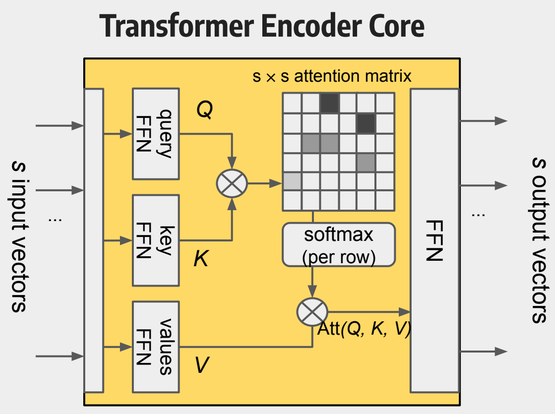
- Produce fixed-sized **visual** and **textual** features
- ➕ Easy to compare
- ➕ **Can be indexed** using already existing text-based or metric space approaches
- They must carry **contextual information**
- ➖ It is **difficult** to represent **relationships between objects** within a single feature

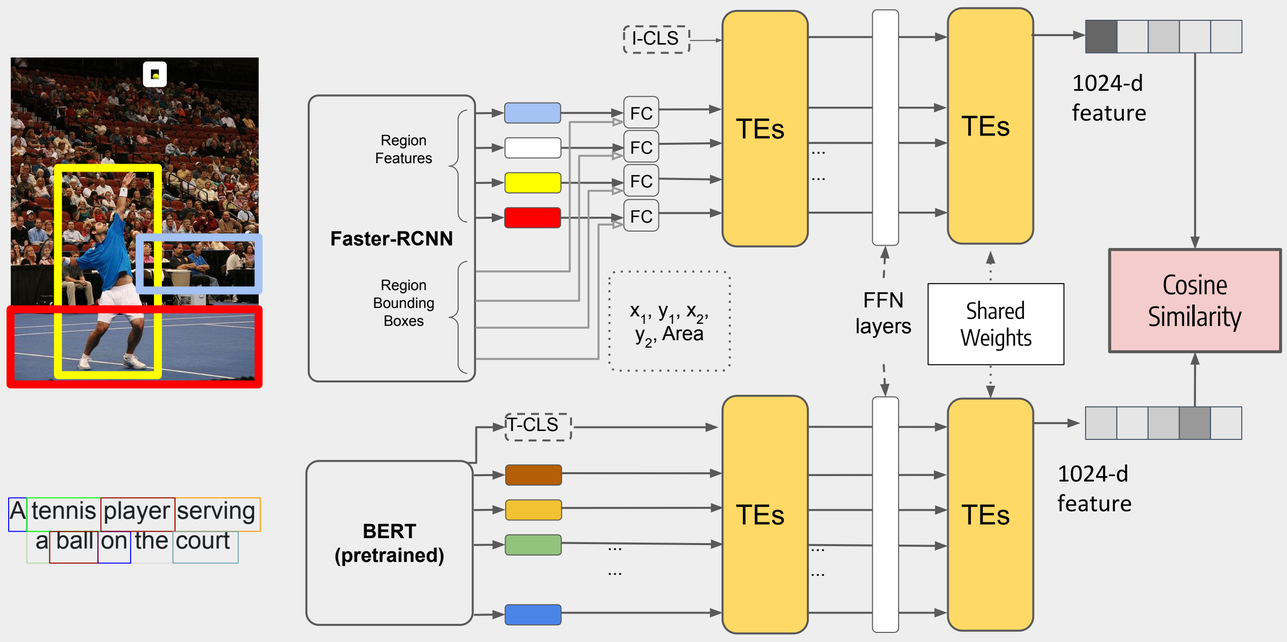
TRANSFORMER ENCODER REASONING NETWORK



- Every region or word should look at its surroundings
- The information is accumulated through the **Transformer Encoder attention mechanism**







Training: Hinge-Based Triplet Ranking Loss

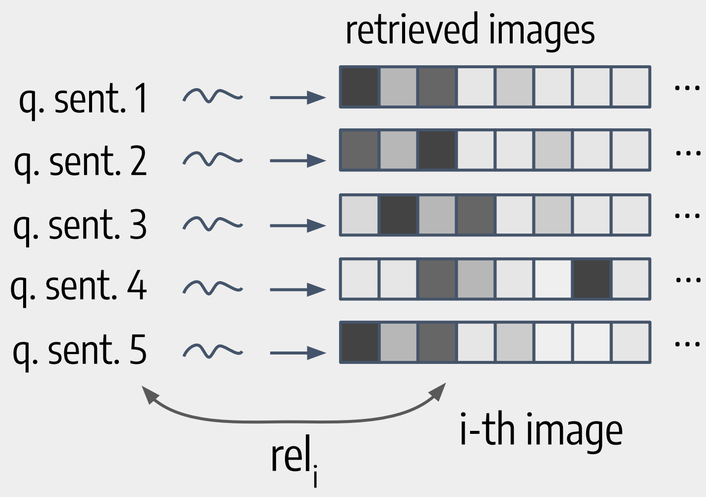
$$L_m(i,c) = \max_{c'}[\alpha + S(i,c') - S(i,c)]_+ + \max_{i'}[\alpha + S(i',c) - S(i,c)]_+$$

Testing:

- Compute features
- Compute similarities
- Rank by decreasing sim.

EVALUATION: NDCG

- Non-exact matches
- High-level semantics



$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{\mathrm{rel}_i}{\log_2(i+1)}$$

where

$$\mathrm{rel}_i = \mathrm{ROUGE\text{-}L}(q, C_i)$$

$$\mathrm{rel}_i = \mathrm{SPICE}(q, C_i)$$

QUANTITATIVE RESULTS

- ➢ MS-COCO dataset (5 human-written sentences per image)
- ➢ 1K test set: 5-fold validation

Model	ROUGE-L	SPICE
VSE-0	0.702	0.616
VSE++	0.712	0.617
VSRN	0.723	0.620
TERN (our)	**0.725**	**0.653**

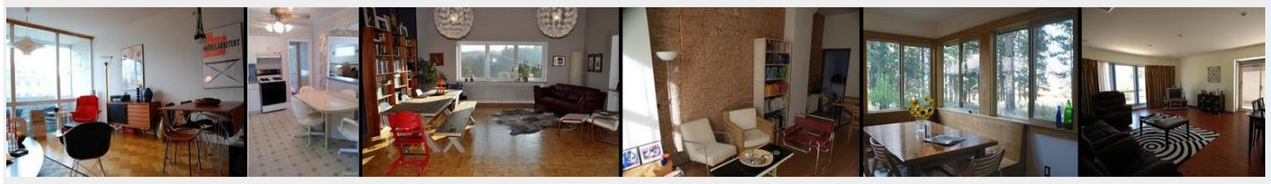
1K test set

Model	ROUGE-L	SPICE
VSE-0	0.633	0.549
VSE++	0.656	0.577
VSRN	**0.676**	0.596
TERN (our)	0.665	**0.600**

5K test set

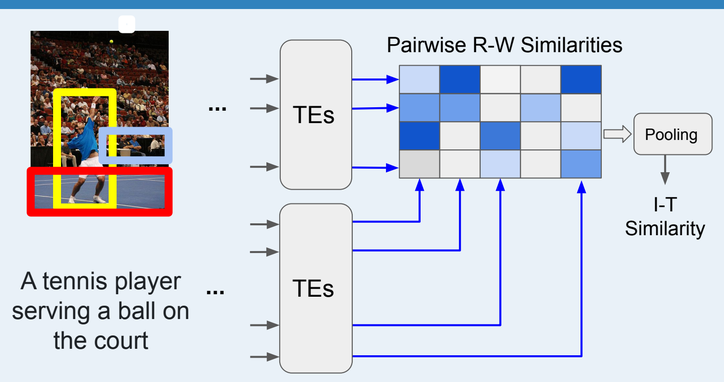
QUALITATIVE RESULTS

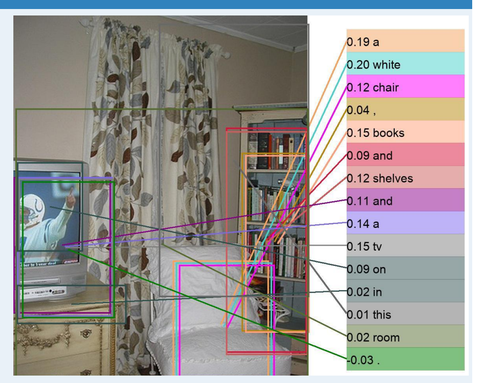
Query: A large jetliner sitting on top of an airport runway.



Query: An eating area with a table and a few chairs.

TOWARDS FINE-GRAINED ALIGNMENT: TERAN





Model	ROUGE-L	SPICE
TERN (our)	0.725	0.653
TERAN (our)	**0.741**	**0.668**

1K test set

Model	ROUGE-L	SPICE
TERN (our)	0.665	0.600
TERAN (our)	**0.682**	**0.610**

5K test set

"*Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders.*"
preprint arXiv:2008.05231 (2020) - **submitted to TOMM journal**