Transformer Reasoning Network for Image-Text Matching and Retrieval

Nicola Messina, Fabrizio Falchi, Andrea Esuli, Giuseppe Amato

ICPR 2020

Efficient Sentence-to-Image Retrieval

- Problem
  - Efficiently retrieve images given a natural language sentence as a query

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

q

$S(i_1, q)=\mathbf{0.90}$ $S(i_2, q)=\mathbf{0.88}$ $S(i_3, q)=\mathbf{0.86}$ $S(i_3, q)=\mathbf{0.76}$


i_1

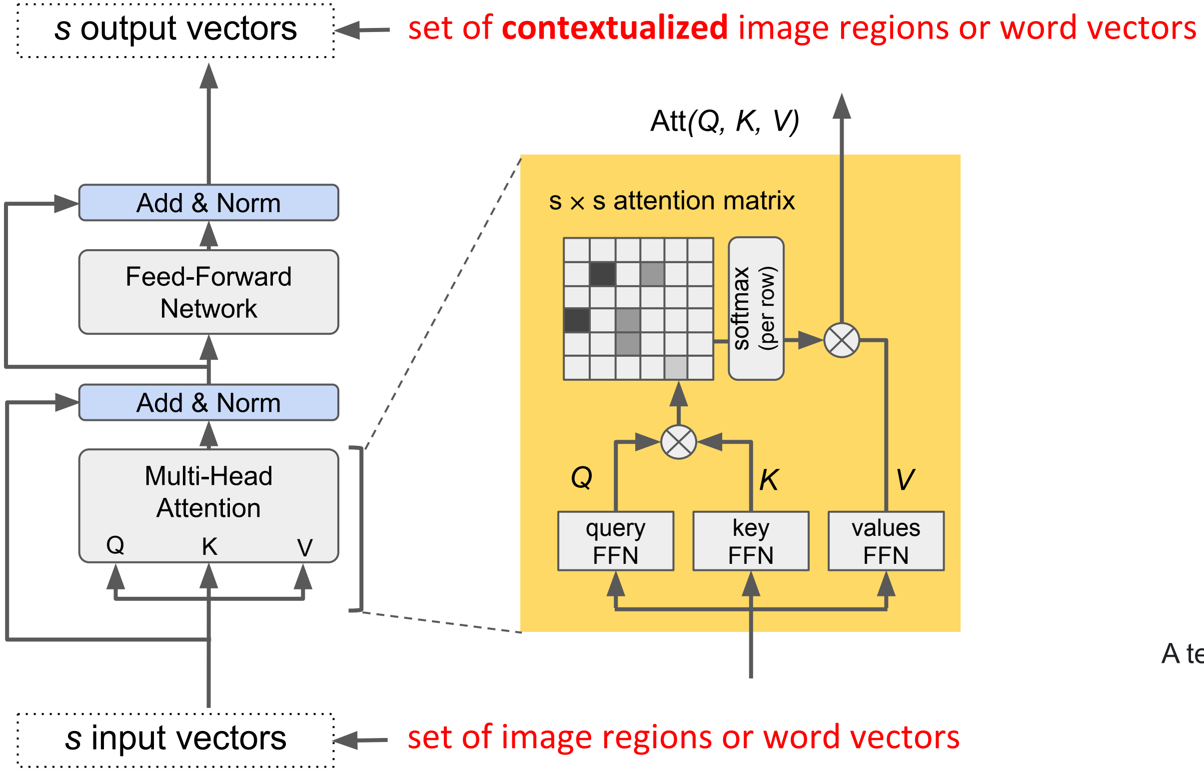

i_2

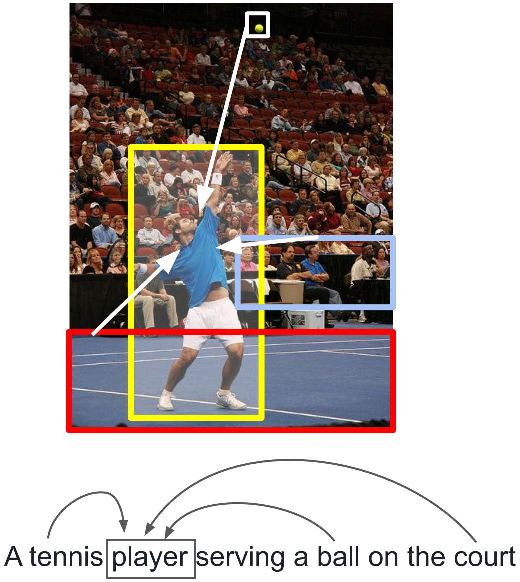

i_3


i_4

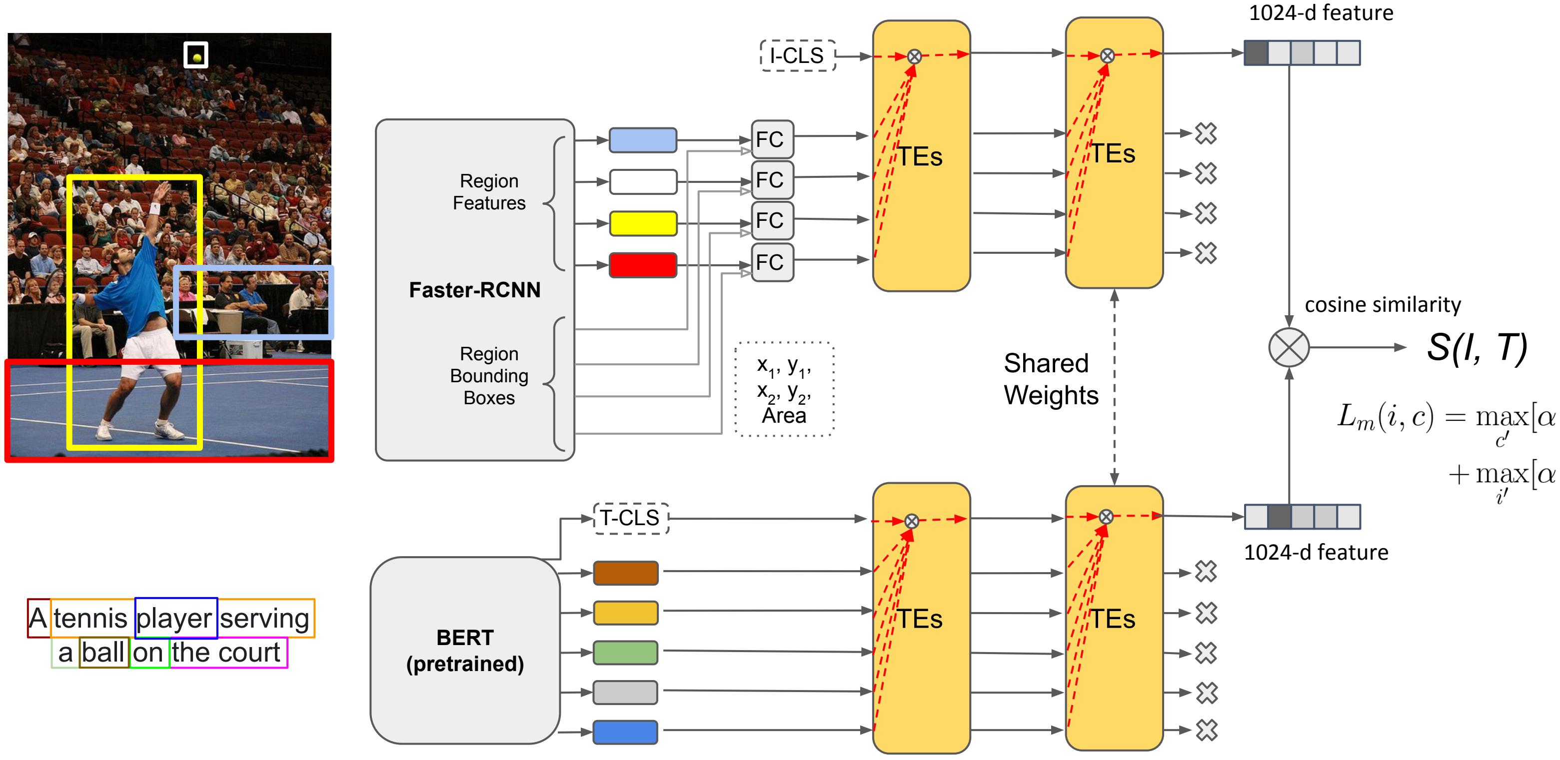
- Challenges
  - Produce compact and very informative **visual** and **textual** features
    - They should be compared using **cosine similarity** to retrieve the I-T similarity score
    - Can be indexed using already existing text-based or metric-space approaches
  - **Effectiveness: context awareness** is important

Transformer Encoder for I-T Processing





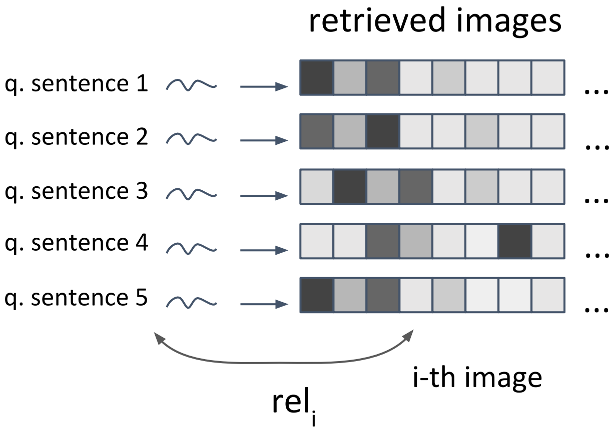
Transformer Encoder Reasoning Network



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

TERN Evaluation

- We used the NDCG metric during evaluation
- It is able to keep into consideration
  - Non-exact matches
  - Highly-semantic aspects of visuals and texts



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

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

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

- Given a pair of sentences, they return a similarity score
- Quite efficient to compute
- SPICE in particular accounts for high-level semantic similarities between sentences

TERN Evaluation

- MS-COCO dataset
  - 5 human-written sentences for each image

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**

NDCG, **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**

NDCG, **5K** test set

TERN Evaluation

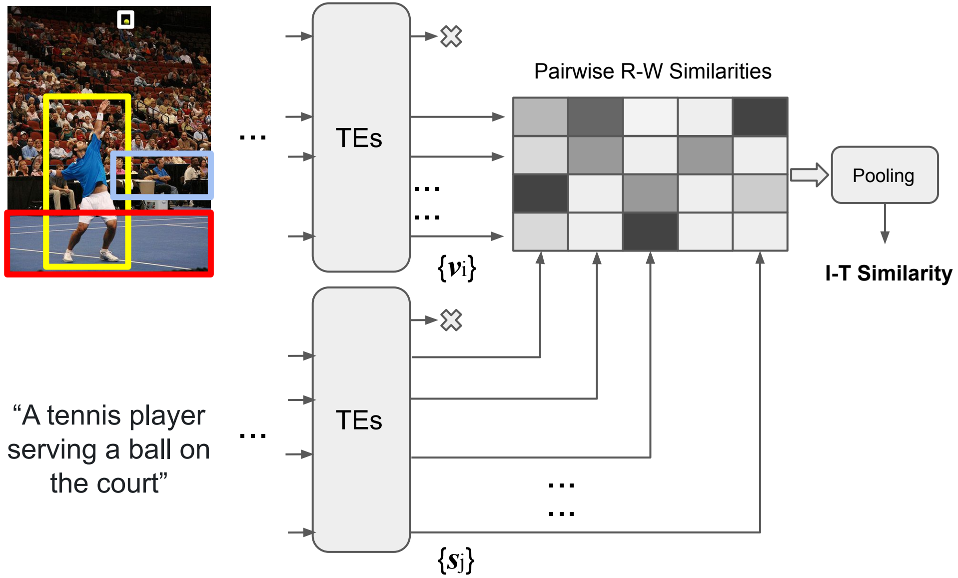
Query: A large jetliner sitting on top of an airport runway.

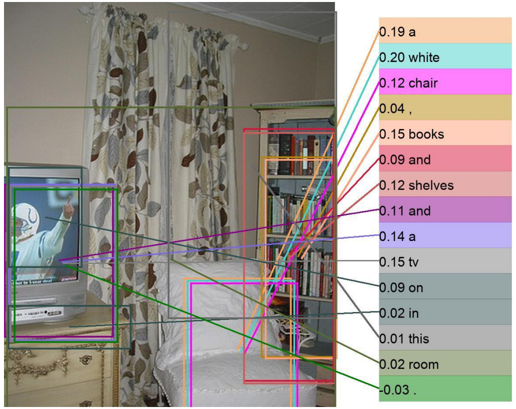
Query: An eating area with a table and a few chairs.

= Exact Match (according to COCO GT)

Transformer Encoder Reasoning and Alignment Network

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





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

Conclusions

- We introduced the TERN architecture
  - TERN produces high-level multi-modal features that can be used in scalable retrieval setups
  - It uses the power of the transformer encoder for obtaining **context-aware** representations.
- We evaluated the retrieval performances using NDCG
  - Relevances computed using SPICE and ROUGE-L textual similarities
- We showed that by enforcing fine-grained R-W alignment we can obtain:
  - interpretable region-word associations
  - better retrieval effectiveness

Thank You!

nicola.messina@isti.cnr.it

GitHub