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

  \[ S(i_1, q) = 0.90 \quad S(i_2, q) = 0.88 \quad S(i_3, q) = 0.86 \quad S(i_4, q) = 0.76 \]

  “Player number 8 kicked the soccer ball with his foot.”

  \( q \) \( i_1 \) \( i_2 \) \( i_3 \) \( i_4 \)

● 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

[s input vectors] -> set of image regions or word vectors

Transformer Encoder

- **Add & Norm**
- **Multi-Head Attention**
  
  \[ \text{Q, K, V} \]
  
  \[ \text{softmax (per row)} \]
  
  \[ \text{Q} \times \text{K} \times \text{V} \]

- **Feed-Forward Network**

- **Output vectors**
  
  set of **contextualized** image regions or word vectors

A tennis player serving a ball on the court
Transformer Encoder Reasoning Network

A tennis player serving a ball on the court

N. Messina et al. Transformer Reasoning Network for Image-Text Matching and Retrieval
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

\[ DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i + 1)} \]

\[ rel_i = \text{ROUGE-L}(q, C_i) \]

\[ rel_i = \text{SPICE}(q, C_i) \]

- 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

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-L</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSE-0</td>
<td>0.702</td>
<td>0.616</td>
</tr>
<tr>
<td>VSE++</td>
<td>0.712</td>
<td>0.617</td>
</tr>
<tr>
<td>VSRN</td>
<td>0.723</td>
<td>0.620</td>
</tr>
<tr>
<td><strong>TERN (our)</strong></td>
<td><strong>0.725</strong></td>
<td><strong>0.653</strong></td>
</tr>
</tbody>
</table>

**NDCG, 1K test set**

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-L</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSE-0</td>
<td>0.633</td>
<td>0.549</td>
</tr>
<tr>
<td>VSE++</td>
<td>0.656</td>
<td>0.577</td>
</tr>
<tr>
<td>VSRN</td>
<td><strong>0.676</strong></td>
<td>0.596</td>
</tr>
<tr>
<td><strong>TERN (our)</strong></td>
<td><strong>0.665</strong></td>
<td><strong>0.600</strong></td>
</tr>
</tbody>
</table>

**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."

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

"A tennis player serving a ball on the court"
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