A Novel Attention-based Aggregation Function to Combine Vision and Language

Matteo Stefanini  
Marcella Cornia  
Lorenzo Baraldi  
Rita Cucchiara

E-mail: {name.surname}@unimore.it
Problem Formulation

How can we effectively aggregate and represent sets/sequences with Transformers?

Best global representation?

CLSL Token

Transformer Encoder

Sequence of words or set of image regions

Vision and Language Fusion: Retrieval and VQA

Text Retrieval

A white boxy birthday cake with red flowers and blue accents on a decorated table with gray cover and candles.

Image Retrieval

An orange is placed on a plate with a cracker.

Visual Question Answering

What is the mustache made of?

AI System

bananas

Both tasks require a method to summarize their content in order to compare or fuse different information.

Compact representations should be guided by the context of both modalities.
Method: Attention Aggregation of feature vectors

- Our method learns a compact representation of sets or sequences of feature vectors.

- Given two modalities $X$ and $Z$, we compute a compressed vector for $X$ as the weighted sum of its vectors:

$$Y(X, Z) = \sum_{i=0}^{n} S_i(X, Z) \cdot X_i$$

$$S(X, Z) = \text{softmax} \left( \text{ScoreAttn}(Q, K, V) \right)$$

- where each weight is a score computed with a function based on cross-attention mechanism, with $Q$ projection of $X$ and $K$-$V$ projections of the other modality $Z$:

$$\text{ScoreAttn}(Q, K, V) = \text{fc} \left( \left[ \text{softmax} \left( \frac{Q_h K_h^T}{\sqrt{d}} \right) V_h \right]_h \right)$$

M. Stefanini, M. Cornia, L. Baraldi, R. Cucchiara. “A Novel Attention-based Aggregation Function to Combine Vision and Language” ICPR 2020
We employ the same simple Visual-Semantic model above to compare different aggregation methods.

Notably, applying our method multiple times we can learn $k$ different vectors representing different global aspects of the inputs.
### Visual Question Answering Results

**M. Stefanini, M. Cornia, L. Baraldi, R. Cucchiara. “A Novel Attention-based Aggregation Function to Combine Vision and Language” ICPR 2020**

<table>
<thead>
<tr>
<th>Aggregation Function</th>
<th>Validation</th>
<th>Test-Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Mean Pooling</td>
<td>54.87</td>
<td>71.50</td>
</tr>
<tr>
<td>Max Pooling</td>
<td>56.73</td>
<td>75.68</td>
</tr>
<tr>
<td>LogSumExp Pooling</td>
<td>54.61</td>
<td>70.94</td>
</tr>
<tr>
<td>1D Convolution</td>
<td>56.87</td>
<td>72.35</td>
</tr>
<tr>
<td>CLS Token</td>
<td>58.31</td>
<td>74.29</td>
</tr>
<tr>
<td>Ours ((k = 1))</td>
<td>60.73</td>
<td>77.68</td>
</tr>
<tr>
<td>Ours ((k = 2))</td>
<td>60.76</td>
<td>78.06</td>
</tr>
<tr>
<td>Ours ((k = 3))</td>
<td>60.50</td>
<td>77.82</td>
</tr>
<tr>
<td>Ours ((k = 5))</td>
<td><strong>60.99</strong></td>
<td><strong>78.62</strong></td>
</tr>
<tr>
<td>Ours ((k = 7))</td>
<td>60.95</td>
<td>78.40</td>
</tr>
<tr>
<td>Ours ((k = 10))</td>
<td>59.94</td>
<td>77.30</td>
</tr>
</tbody>
</table>

**Question:** What color is the car on the right?  
**Ground-truth:** red  
**Mean:** white  
**Ours:** red

**Question:** How many people can you see?  
**Ground-truth:** eight  
**Mean:** five  
**Ours:** seven

**Question:** Is the bear real?  
**Ground-truth:** no  
**Mean:** yes  
**Ours:** no

**Question:** What color is the floor?  
**Ground-truth:** brown  
**Mean:** yellow  
**Ours:** brown

**Question:** Is the girl sitting on the horse?  
**Ground-truth:** yes  
**Mean:** no  
**Ours:** yes

**Question:** what is on the train?  
**Ground-truth:** graffiti  
**Mean:** people  
**Ours:** graffiti

**Question:** What is the yellow food?  
**Ground-truth:** corn  
**Mean:** eggs  
**Ours:** corn

**Question:** how many giraffes are there?  
**Ground-truth:** two  
**Mean:** three  
**Ours:** two
## Retrieval Results

<table>
<thead>
<tr>
<th>Aggregation Function</th>
<th>Text Retrieval</th>
<th>Image Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>Mean Pooling</td>
<td>69.66</td>
<td>93.12</td>
</tr>
<tr>
<td>Max Pooling</td>
<td>69.04</td>
<td>92.68</td>
</tr>
<tr>
<td>LogSumExp Pooling</td>
<td>64.20</td>
<td>91.52</td>
</tr>
<tr>
<td>1D Convolution</td>
<td>65.66</td>
<td>91.86</td>
</tr>
<tr>
<td>CLS Token</td>
<td>70.30</td>
<td>93.38</td>
</tr>
<tr>
<td>Ours (k = 1)</td>
<td>70.80</td>
<td>93.16</td>
</tr>
<tr>
<td>Ours (k = 2)</td>
<td>70.36</td>
<td><strong>93.46</strong></td>
</tr>
<tr>
<td>Ours (k = 3)</td>
<td>70.42</td>
<td>93.34</td>
</tr>
<tr>
<td>Ours (k = 4)</td>
<td>70.14</td>
<td>93.42</td>
</tr>
</tbody>
</table>

**Query Caption:** An orange is placed on a plate with a cracker.

**Top-1 (Ours)**

**Query Caption:** A man in blue jacket standing by a passing train.

**Top-1 (Ours)**

**Query Caption:** Many umbrellas on a beach near a body of water.

**Top-1 (Ours)**

**Query Caption:** Two animals walking through high grass in the woods.

**Top-1 (Ours)**

**Query Caption:** A white boxy birthday cake with red flowers on a decorated table with candles.

**Mean:** A sheet cake sitting on top of a table with lit candles.

**Top-1 (Ours)**

**Query Caption:** A large white blue and red clock shaped like a cup.

**Mean:** A triangle sign with an English and foreign warning.

**Top-1 (Ours)**

**Query Caption:** The boy is getting ready to hit the ball with his bat.

**Mean:** A man is posed in mid swing about to serve a ball in a tennis court.

**Top-1 (Ours)**

**Query Caption:** A dog is running alongside a horse in a coral.

**Mean:** Two brown dogs are playing on the dirt.

**Top-1 (Ours)**

---

M. Stefanini, M. Cornia, L. Baraldi, R. Cucchiara. “A Novel Attention-based Aggregation Function to Combine Vision and Language” ICPR 2020
Thank you

Matteo Stefanini  Marcella Cornia  Lorenzo Baraldi  Rita Cucchiara

E-mail: {name.surname}@unimore.it