





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



Matteo Stefanini







Rita Cucchiara

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

Problem Formulation

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





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



- Both tasks require a method to summarize their content in order to <u>compare</u> or <u>fuse</u> different information.
- Compact representations should be guided by the <u>context</u> of both modalities.





I

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:

$$oldsymbol{Y}(oldsymbol{X},oldsymbol{Z}) = \sum_{i=0}^{n_q} oldsymbol{S}_i(oldsymbol{X},oldsymbol{Z})\cdotoldsymbol{X}_i$$

 $S(X, Z) = \operatorname{softmax} (\operatorname{ScoreAttn}(Q, \mathcal{K}, \mathcal{V}))$

 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:

$$\mathsf{ScoreAttn}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \mathrm{fc}\left(\left[\operatorname{softmax}\left(\frac{\boldsymbol{Q}_{h}\boldsymbol{K}_{h}^{T}}{\sqrt{d}}\right)\boldsymbol{V}_{h}\right]_{h}\right)$$





- 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



	Validation				Test-Dev			
Aggregation Function	All	Yes/No	Number	Others	All	Yes/No	Number	Others
Mean Pooling	54.87	71.50	37.93	46.69	56.05	71.00	38.88	47.19
Max Pooling	56.73	75.68	37.64	47.37	57.95	75.14	38.48	47.69
LogSumExp Pooling	54.61	70.94	38.27	46.53	55.68	70.36	38.72	47.00
1D Convolution	56.87	72.35	39.18	49.79	57.79	71.71	39.97	49.96
CLS Token	58.31	74.29	39.89	51.03	59.40	74.26	40.31	51.07
Ours $(k = 1)$	60.73	77.68	41.86	52.84	62.05	77.84	42.47	53.03
Ours $(k=2)$	60.76	78.06	42.32	52.48	62.06	78.26	42.62	52.66
Ours $(k=3)$	60.50	77.82	41.56	52.33	61.80	78.22	41.69	52.35
Ours $(k=5)$	60.99	78.62	42.53	52.46	62.17	78.52	42.27	52.74
Ours $(k=7)$	60.95	78.40	42.65	52.53	62.43	78.75	43.33	52.83
Ours $(k = 10)$	59.94	77.30	40.82	51.80	61.16	77.39	40.69	51.97











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

the train?

Ours: graffiti





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



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



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



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

Retrieval Results



	Тех	at Retri	eval	Image Retrieval			
Aggregation Function	R@ 1	R@5	R@ 10	R@ 1	R@5	R@10	
Mean Pooling	69.66	93.12	97.64	50.42	82.27	90.83	
Max Pooling	69.04	92.68	96.98	51.20	83.27	91.52	
LogSumExp Pooling	64.20	91.52	96.84	47.22	82.26	91.23	
1D Convolution	65.66	91.86	96.58	49.25	81.43	90.42	
CLS Token	70.30	93.38	97.24	51.05	83.28	91.80	
Ours $(k = 1)$	70.80	93.16	97.24	50.77	82.76	91.31	
Ours $(k = 2)$	70.36	93.46	97.20	51.31	83.38	91.69	
Ours $(k = 3)$	70.42	93.34	97.22	50.98	83.17	91.65	
Ours $(k = 4)$	70.14	93.42	97.76	50.82	82.66	91.14	



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









Rita Cucchiara

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