Adaptive Cross-modal Embeddings
for Image-Text Alignment

AUTHORS:

Jônatas Wehrmann
Camila Kolling
Rodrigo C. Barros

CONTACT :

jonatas.wehrmann@edu.pucrs.br
camila.kolling@edu.pucrs.br
rodrigo.barros@pucrs.br

INTRO:

1. Novel approach for training image-text alignment models
2. Retrieve an image given a query text, or captions that successfully label an image
3. It is designed to filter and enhance important information across internal features, allowing for guided vector representations
4. It works similarly to attention modules, though it is far more computationally efficient.
5. Outperforms attention-based SOTA Image Retrieval approaches by relative improvements of 24% on Flickr30k, and by 12% on MSCOCO-1k.

METHOD

1. Projects image-caption pairs into the same space.
2. **Image encoder:** employs a finetuned faster RCNN network on Visual Genome dataset.
3. **Text encoding**: Word-embeddings plus a GRU recurrent network.
4. **Employs ADAPT rather than attention**
5. It improves the embedded representation of instances from the modality *a* based on the global information of the modality *b*.

Formally, ADAPT is given by:

$$A = \psi(\mathcal{A}), B = \phi(\mathcal{B}) \qquad \gamma_a = g(\mathbf{a}, \theta_g), \beta_a = b(\mathbf{a}, \theta_b)$$

$$\overline{B_i} = B_i \odot \gamma_a + \beta_a \qquad M_{ij} = \left(\frac{e^{\overline{B}_{ij}}}{\sum_{i=1}^{n_b} e^{\overline{B}_{ij}}} \lambda \right)$$

$$\overline{\mathbf{b}} = \frac{1}{n_b} \sum_{i=1}^{n_b} \left(\overline{B} \odot M \right)_i$$

LOSS FUNCTION

$$\mathcal{J}_s(\mathbf{a}, \mathbf{b}) = \sum_{b'} [\alpha - s(\mathbf{a}, \mathbf{b}) + s(\mathbf{a}, \mathbf{b}')] + \sum_{\mathbf{a}'} [\alpha - s(\mathbf{b}, \mathbf{a}) + s(\mathbf{b}, \mathbf{a}'))]$$

$$\mathcal{J}_m(\mathbf{a}, \mathbf{b}) = \max_{\mathbf{b}'} [\alpha - s(\mathbf{a}, \mathbf{b}) + s(\mathbf{a}, \mathbf{b}')]_+ + \max_{\mathbf{a}'} [\alpha - s(\mathbf{b}, \mathbf{a}) + s(\mathbf{b}, \mathbf{a}')]_+$$

$$\mathcal{J} = \tau(\epsilon) \cdot \mathcal{J}_m + (1 - \tau(\epsilon)) \cdot \mathcal{J}_s$$
$$\tau = (1 - \eta^{\epsilon})$$

RESULTS

Flickr30k

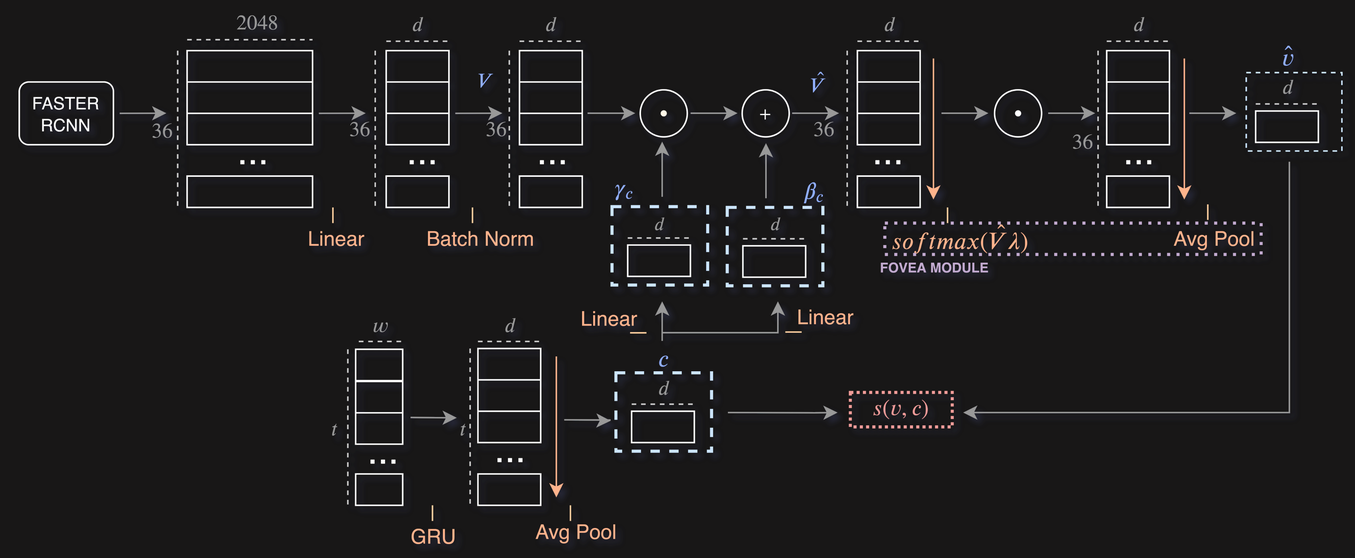
	Annotation	Retrieval
SCAN-T2I	61.8	45.8
SCAN-I2T	67.9	43.9
SCAN-ENS	67.4	48.6
ADAPT-I2T	70.2	55.5
ADAPT-T2I	73.6	57.0
ADAPT-ENS	76.6	60.7

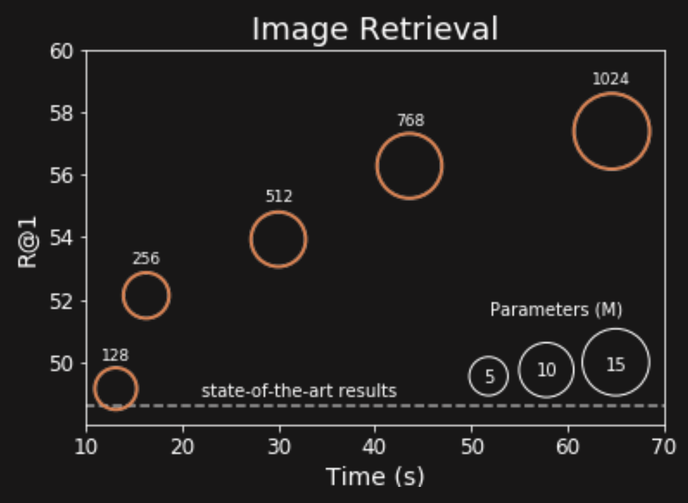
MS COCO

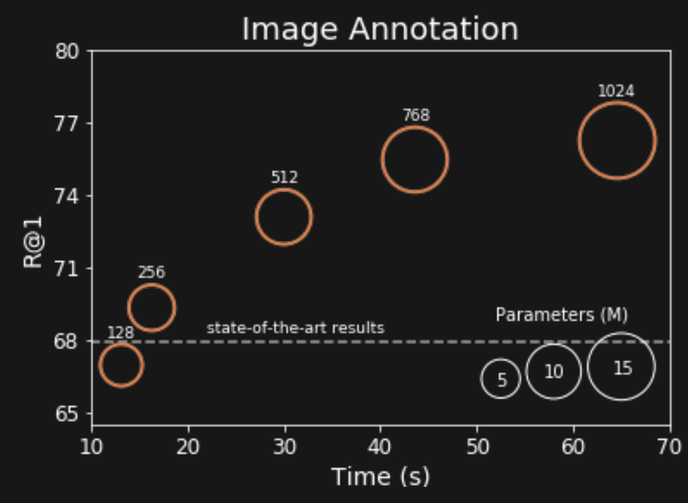
	Annotation	Retrieval
SCAN-T2I	70.9	56.4
SCAN-I2T	69.2	54.4
SCAN-ENS	72.7	58.8
ADAPT-I2T	74.5	62.0
ADAPT-T2I	75.3	63.3
ADAPT-ENS	76.5	62.2

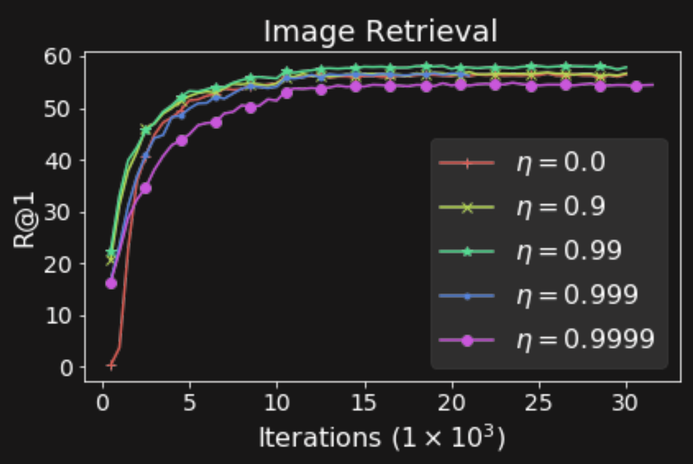
Ablation

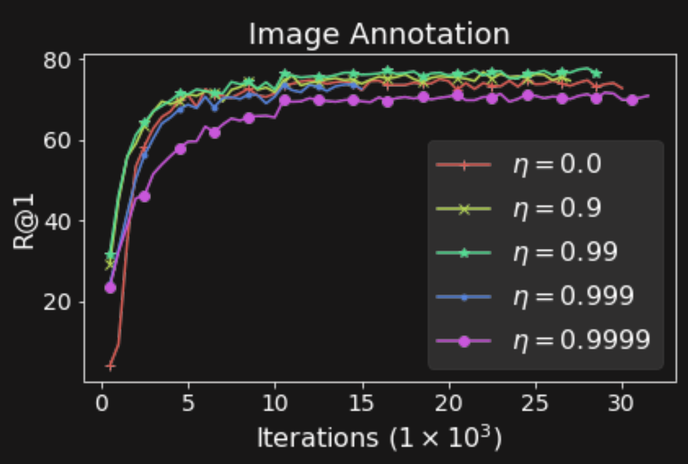
ADAPT-T2I	Annotation	Retrieval
Default	76.2	57.4
No-fovea	72.4	52.9
λ=1	72.3	56.8
λ=5	75.1	57.5
λ=10	76.2	57.4
Δλ	75.7	58.1



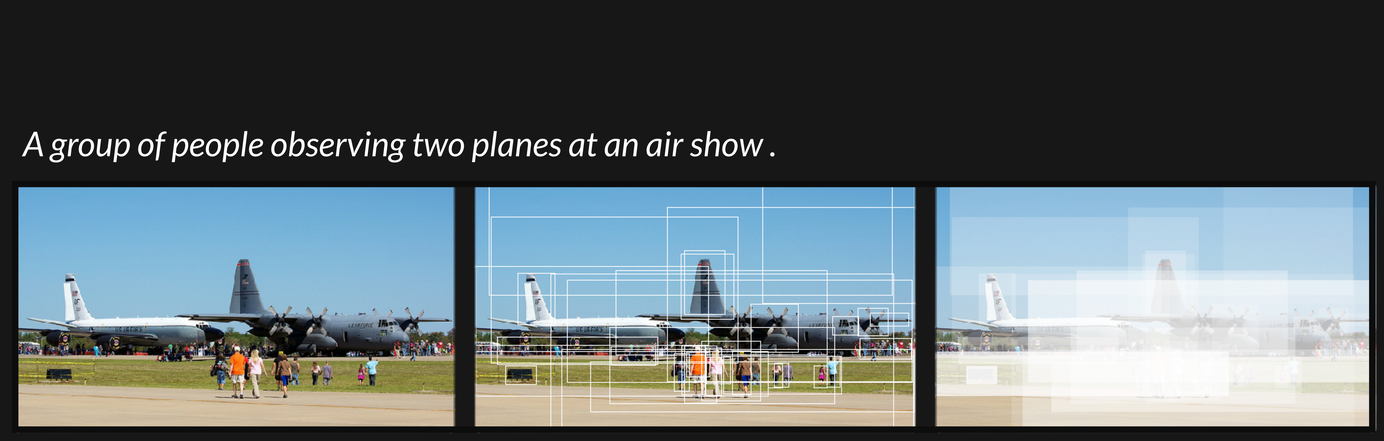




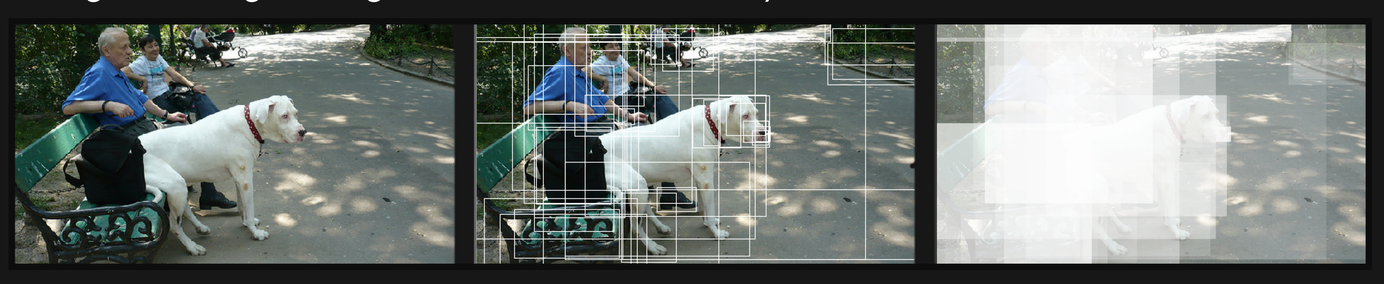




A group of people observing two planes at an air show .



A large white dog is sitting on a bench beside an elderly man .



Project page: https://github.com/jwehrmann/retrieval.pytorch

PUCRS