NAPReg: Nouns As Proxies Regularization For Semantically Aware Cross-Modal Embeddings

Bhavin Jawade*, Deen Dayal Mohan*, Naji Mohamed Ali, Srirangaraj Setlur, Venu Govindaraju
{bhavinja, dmohan, najmoh, setlur, govind}@buffalo.edu

Motivation

- Text-to-image matching is the most common form of cross-modal retrieval.
- Existing methods use dual encoders with an attention mechanism and a ranking loss to learn embeddings for retrieval.
- These methods do not have explicit supervision to enforce semantic alignment between visual regions and textual words.
- We propose NAPReg, a regularization formulation that projects high-level semantic entities into the embedding space as shared learnable proxies.

This allows the attention mechanism to learn better word-region alignment and build a more generalized latent representation for semantic concepts.

Our method outperforms existing methods in cross-modal metric learning for text-image and image-text retrieval tasks.

Problem Statement

- Consider, visual features of an image \( V = \{v_1, v_2, v_3\} \) and textual features \( T = \{t_1, t_2, t_3\} \)
- Fine-grained similarity between image and text can be given as:

\[
S(V, T) = f(\Phi(V; \theta_1), \Psi(T; \theta_2))
\]

Stacked Cross Attention (Lee et al. [2])

For each visual location, an attended combination of word representation a \( \text{word} \) (i.e., the attended sentence vector \( a_{j} \), with respect to the \( i^{th} \) image region \( a_{i} \)) is constructed as defined below:

\[
s_{ij} = \frac{a_{i}^T t_{j}}{||a_{i}|| \cdot ||t_{j}||}, \quad i \in [1, n], j \in [1, m]
\]

\[
\omega_{ij} = \exp(\sigma_{ij}), \quad \sigma_{ij} = \sum_{k=1}^{m} \exp(\tau_{ik}), \quad \tau_{ik} = a_{i}^T t_{k}
\]

\[
a_{j} = \sum_{i=1}^{n} \omega_{ij} \cdot t_{i}
\]

The overall cosine similarity between the image-text pair is given by:

\[
S(V, T) = \frac{1}{n} \sum_{i=1}^{n} a_{i}^T t_{i}
\]

Method (Nouns As Proxies)

\[\text{Method (Nouns As Proxies)}
\]

Qualitative Results

Proxes

- Shows the top 2 regions attended by each proxy word in the image on the left and heatmap between the similarity of selected visually relevant regions and the word proxies on the right
- The similarity score of the visual region containing the cat and the dog is highest for the corresponding word in the text
- The magnitude of the scores has also increased in comparison to the model without the proposed regularization.

Ablation Study

Table: Quantitative Results

Recall@K(%) performance on MSOCO dataset

Recall@K(%) performance on Flickr30K dataset

References