Do Neural Network Cross-modal Mappings Really Bridge Modalities?

Guillem Collell, Marie-Francine Moens

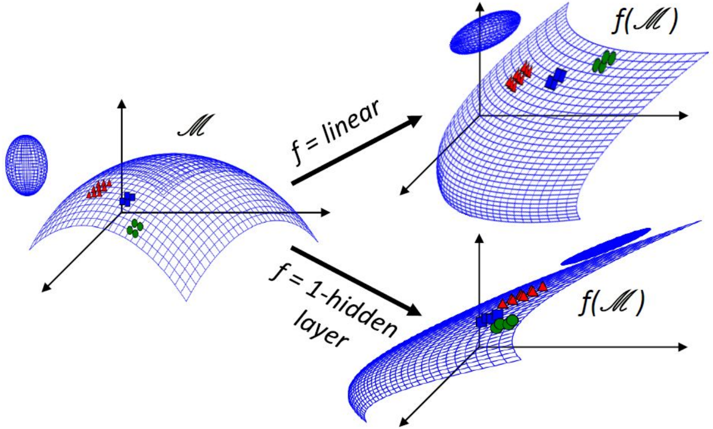
ACL 2018

Cross-modal mappings

Objective: Learn a mapping function from representations of one modality to another.

Applications:

- Cross-modal retrieval (e.g. image search)
- Zero-shot learning
- Word translation
- Building multimodal representations



Desirable(?) property: Mapped representations should have a similar *neighborhood structure* to the true target representations.

Approach

Instead of using similarity metrics like

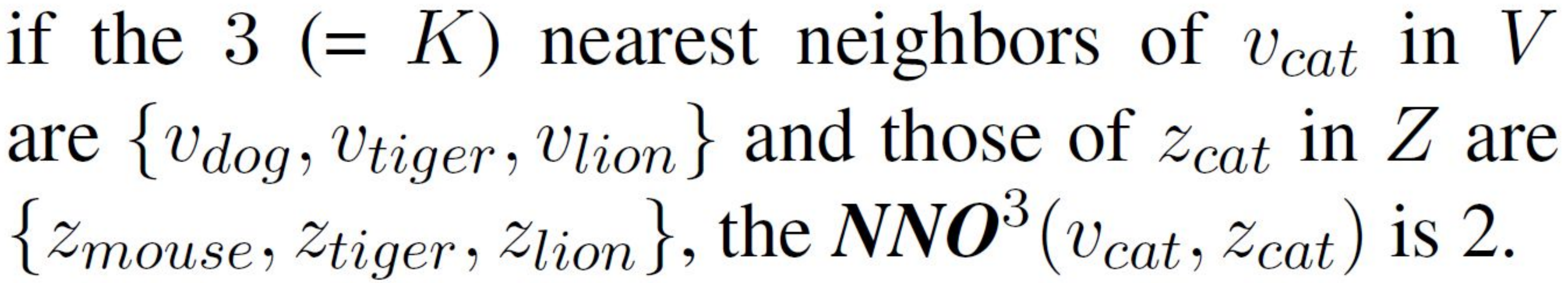
- MSE loss
- Cosine similarity
- Max-margin loss

Measure the similarities of the **neighborhood structures** of sets of vectors.

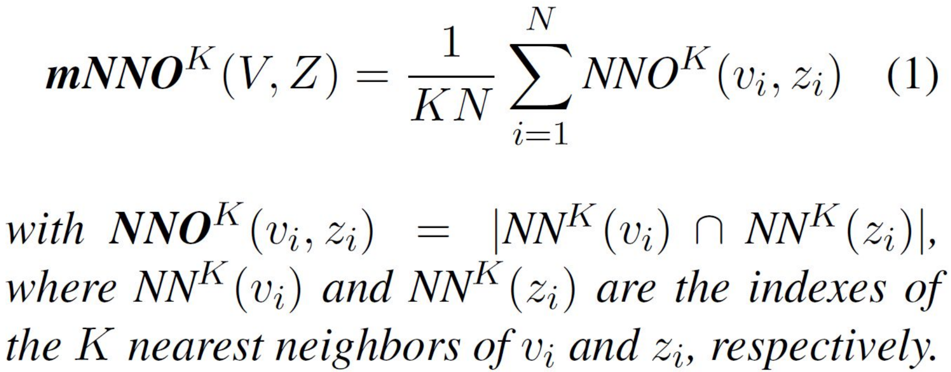
New metric: **mean nearest neighbor overlap**

- Count the (average) proportion of neighbors that appear in the neighborhoods of two vectors.a representation and its mapping.

Example



if the 3 ($= K$) nearest neighbors of v_{cat} in V are $\{v_{dog}, v_{tiger}, v_{lion}\}$ and those of z_{cat} in Z are $\{z_{mouse}, z_{tiger}, z_{lion}\}$, the $\boldsymbol{NNO}^3(v_{cat}, z_{cat})$ is 2.



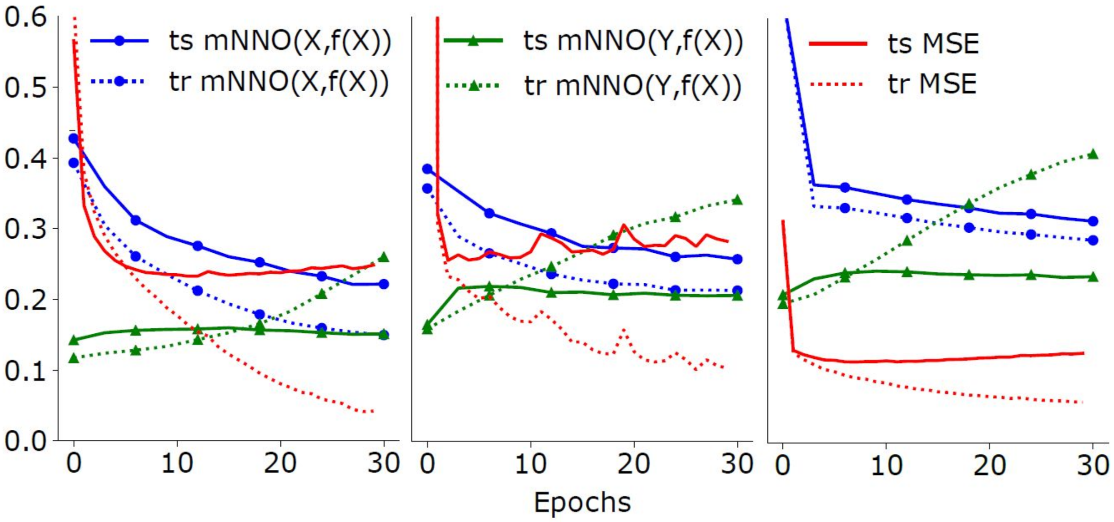
$$\boldsymbol{mNNO}^K(V, Z) = \frac{1}{KN} \sum_{i=1}^{N} \mathit{NNO}^K(v_i, z_i) \quad (1)$$

with $\boldsymbol{NNO}^K(v_i, z_i) = |\mathit{NN}^K(v_i) \cap \mathit{NN}^K(z_i)|$, *where* $\mathit{NN}^K(v_i)$ *and* $\mathit{NN}^K(z_i)$ *are the indexes of the* K *nearest neighbors of* v_i *and* z_i*, respectively.*

Results

Evaluated mappings between images and text and vice versa.

- Mapped representations have more similar neighborhood structures with the original representations, instead of the target representations.



Discussion points

- What kind of similarity constraints should be used for learning coordinated multimodal representations?
- Could this be applied to multimodal learning with missing modalities?
- Is the “desirable property” always desirable?