Do Neural Network Cross-Modal Mappings Really Bridge Modalities?

Guillem Collell & Marie-Francine Moens

Language Intelligence and Information Retrieval group (LIIR)
Department of Computer Science

KU LEUVEN

Story

Collell, G., Zhang, T., Moens, M.F. (2017) *Imagined Visual Representations as Multimodal Embeddings*. **AAAI**

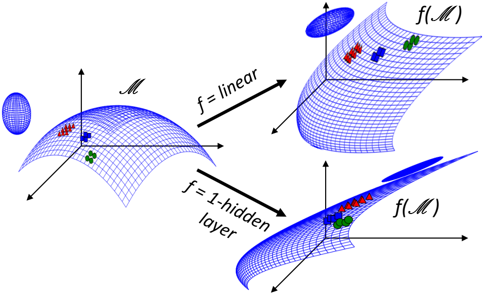
- Learn **mapping** f: *text* $\longrightarrow$ *vision*.
- **Finding 1:** Imagined vectors, ***f(text)***, outperform original visual vectors in 7/7 word similarity tasks.
- So, why are mapped vectors **multimodal**? We conjecture:
  - **Continuity**. Output vector is nothing but the input vector transformed by a continuous map: $f(\vec{x}) = \vec{x}\theta$.
- **Finding 2 (not in AAAI paper):** Vectors imagined with an untrained network do even better.

Motivation

- **Applications** (e.g., *zero-shot image tagging*, *zero-shot translation* or *cross-modal retrieval*):
  - Use **linear** or **NN** maps to bridge modalities / spaces.
  - Then, they tag / translate based on ***neighborhood structure*** of mapped vectors $f(X)$.
- **Research question**: *Is the neighborhood structure of $f(X)$ similar to that of Y? Or rather to X?*
- **How** to measure similarity of 2 sets of vectors from different spaces? **Idea:** mean nearest neighbor overlap (**mNNO**)

General Setting

- **Mappings** $f : \mathcal{X} \to \mathcal{Y}$ to bridge modalities $\mathcal{X}$ and $\mathcal{Y}$:
  - **Linear** (***lin***): $f(x) = W_0 x + b_0$
  - Feed-forward **neural net** (***nn***): $f(x) = W_1 \sigma(W_0 x + b_0) + b_1$



Experiment 1

Definition

Nearest Neighbor Overlap (**NNO**$^K(v_i, z_i)$) = number of K nearest neighbors that two *paired* data points v_i, z_i share in their respective spaces.

The **mean NNO** is:

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

$$\begin{cases} \mathrm{NN}^3(v_{cat}) = \{v_{dog}, v_{tiger}, v_{lion}\} \\ \mathrm{NN}^3(z_{cat}) = \{z_{mouse}, z_{tiger}, z_{lion}\} \end{cases} \Rightarrow \quad \boldsymbol{NNO}^3(v_{cat}, z_{cat}) = 2 \tag{1}$$

Experiment 1

- **Goal**: Learn map $f : X \to Y$ and calculate ***mNNO***$(Y, f(X))$. Compare it with ***mNNO***$(X, f(X))$

Experimental Setup

- **Datasets**: (i) *ImageNet*; (ii) *IAPR TC-12*; (iii) *Wikipedia*
- **Visual features**: VGG-128 and ResNet.
- **Text features**: *ImageNet* (GloVe and word2vec); *IAPR TC-12* & *Wikipedia* (biGRU).
- **Loss**: $MSE = \frac{1}{2}\|f(x) - y\|^2$. We also tried *max-margin* and *cosine*.

Experiment 1: Results

			ResNet		VGG-128	
			$X, f(X)$	$Y, f(X)$	$X, f(X)$	$Y, f(X)$
ImageNet	$I \rightarrow T$	*lin*	**0.681***	0.262	**0.723***	0.236
		nn	**0.622***	0.273	**0.682***	0.246
	$T \rightarrow I$	*lin*	**0.379***	0.241	**0.339***	0.229
		nn	**0.354***	0.27	**0.326***	0.256
IAPR TC-12	$I \rightarrow T$	*lin*	**0.358***	0.214	**0.382***	0.163
		nn	**0.336***	0.219	**0.331***	0.18
	$T \rightarrow I$	*lin*	**0.48***	0.2	**0.419***	0.167
		nn	**0.413***	0.225	**0.372***	0.182
Wikipedia	$I \rightarrow T$	*lin*	**0.235***	0.156	**0.235***	0.143
		nn	**0.269***	0.161	**0.282***	0.148
	$T \rightarrow I$	*lin*	**0.574***	0.156	**0.6***	0.148
		nn	**0.521***	0.156	**0.511***	0.151

Table: $X, f(X)$ and $Y, f(X)$ denote $mNNO^{10}(X, f(X))$ and $mNNO^{10}(Y, f(X))$, respectively.

Experiment 2

- **Goal**: Map X with an ***untrained net*** f and compare performance of X with that of $f(X)$.
- We "ablate" from Experiment 1 the *learning* part and the choices of *loss* and *output vectors*.

Experimental Setup

Evaluate vectors in:

- **(i) Semantic similarity**: *SemSim*, *Simlex-999* and *SimVerb-3500*.
- **(ii) Relatedness**: *MEN* and *WordSim-353*.
- **(iii) Visual similarity**: *VisSim*.

Experiment 2: Results

	WS-353		Men		SemSim	
	Cos	Eucl	Cos	Eucl	Cos	Eucl
f_{nn}(GloVe)	**0.632**	**0.634***	0.795	**0.791***	0.75*	**0.744***
f_{lin}(GloVe)	0.63	0.606	0.798	0.781	0.763	0.712
GloVe	**0.632**	0.601	**0.801**	0.782	**0.768**	0.716
f_{nn}(ResNet)	0.402	0.408*	0.556	**0.554***	0.512	0.513
f_{lin}(ResNet)	**0.425**	0.449	0.566	0.534	0.533	0.514
ResNet	0.423	**0.457**	**0.567**	0.535	**0.534**	**0.516**
	VisSim		**SimLex**		**SimVerb**	
	Cos	Eucl	Cos	Eucl	Cos	Eucl
f_{nn}(GloVe)	0.594*	**0.59***	0.369	**0.363***	0.313	**0.301***
f_{lin}(GloVe)	0.602*	0.576	0.369	0.341	**0.326**	0.23
GloVe	**0.606**	0.58	**0.371**	0.34	0.32	0.235
f_{nn}(ResNet)	0.527*	**0.526***	0.405	**0.406**	0.178	0.169
f_{lin}(ResNet)	0.541	0.498	**0.409**	0.404	0.198	0.182
ResNet	**0.543**	0.501	**0.409**	0.403	**0.211**	**0.199**

Table: Spearman correlations between human ratings and similarities (cosine or Euclidean) predicted from embeddings.

Conclusions and Future Work

Conclusions:

- Neighborhood structure of $f(X)$ more similar to X than Y.
- Neighborhood structure of embeddings not significantly disrupted by mapping them with an *untrained net*.

Future Work: *How to mitigate the problem?*

- Discriminator (adversarial) trying to guess whether the sample is from Y or $f(X)$.
- Incorporate pairwise similarities into loss function.

Thank you!

Questions?