Optimizing Visual Representations in Semantic Multi-Modal Models with Dimensionality Reduction, Denoising and Contextual Information

Maximilian Köper and Kim-Anh Nguyen and Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{maximilian.koeper,kim-anh.nguyen,schulte}@ims.uni-stuttgart.de

Abstract

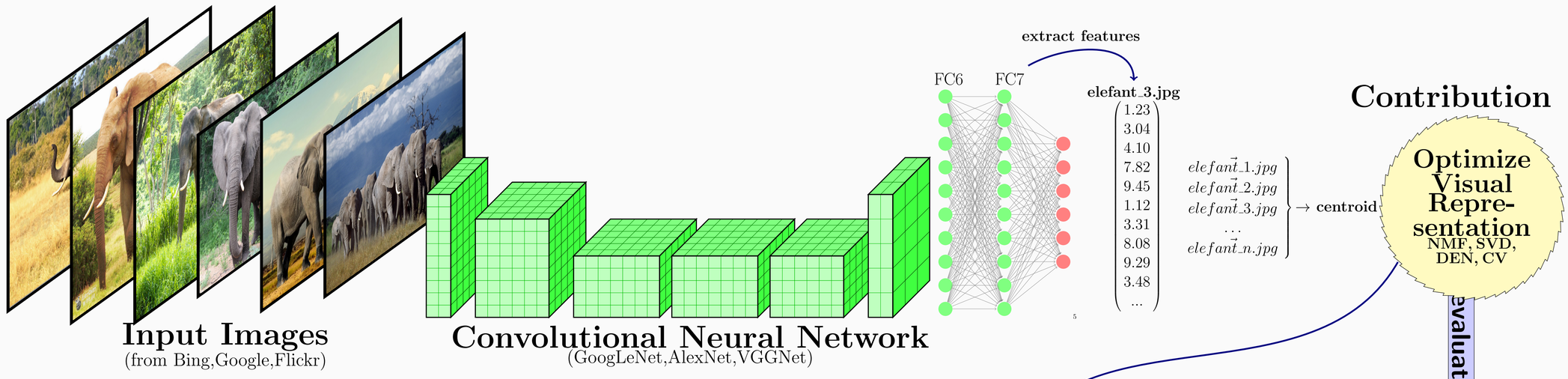
- We improve **visual representations** for multi-modal semantic models by
  → Applying standard **dimensionality reduction** and **denoising** techniques
  → Proposing a novel technique **ContextVision** that takes corpus-based textual information into account when enhancing visual embeddings
- We explore our contribution in a **visual** and a **multi-modal** setup and evaluate on benchmark word **similarity** and **relatedness** tasks

Motivation & Method

- Computational models across tasks potentially profit from combining corpus-based textual information with perceptional information
  → Word meanings are grounded in the external environment. Sensorimotor experience cannot be learned only based on linguistic symbols, cf. the **grounding problem** Harnad (1990)
- Recent advances in computer vision (deep learning) led to the development of better visual representations. Features are extracted from convolutional neural networks (CNNs)
- Dimension reduction techniques & denoising improve performance when applied to word representations Bullinaria and Levy (2012), Nguyen et al. (2016)
  → What about visual representations ?

- Singular Value Decomposition (**SVD**): a matrix algebra operation that can be used to reduce matrix dimensionality yielding a new high-dimensional space
- Non-negative matrix factorization (**NMF**) is a a matrix factorisation approach where the reduced matrix contains only non-negative real numbers
- denoising methods (**DEN**) use a non-linear, parameterized, feedforward neural network as a filter on word embeddings to reduce noise
- Our novel idea *ContextVision* (**CV**) strengthens visual vector representations by performing negative sampling using visual representation and corpus contexts.

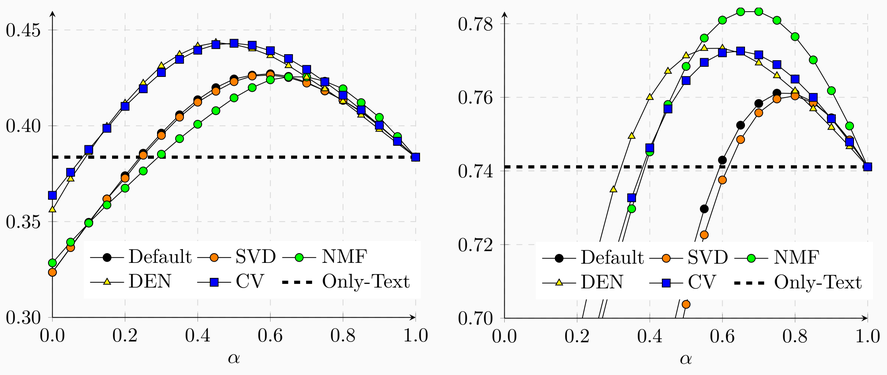
Overview



Multi-Modal Setup

- Varying a weight threshold (α). Similarity is computed as follows:

$$sim(x, y) = \alpha \cdot ling(x, y) + (1 - \alpha) \cdot vis(x, y)$$



- SimLex: bing+AlexNet
- MEN: flickr+AlexNet

Average gain/loss:

	SimLex	MEN	Both
SVD	0.11	-0.20	-0.05
NMF	1.71	10.49	6.10
DEN	1.63	7.34	4.48
CV	3.23	8.29	5.76

Conclusion

We successfully applied dimensionality reduction as well as denoising techniques. Except for SVD, all investigated methods showed **significant improvements** in single- and multi-modal setups on the task of predicting similarity and relatedness.

Results (only visual)

		AlexNet		GoogLeNet		VGGNet	
		SimLex	MEN	SimLex	MEN	SimLex	MEN
bing	Default	.324	.560	.314	.513	.312	.545
	SVD	.324	.557	.316	.513	.314	.544
	NMF	.329	**.610***	.341*	**.612***	.330	**.631***
	DEN	.356*	.582*	.342*	.564*	.343*	.599*
	CV	**.364***	.583*	**.358***	.582*	**.357***	.603*
flickr	Default	.271	.434	.244	.366	.262	.422
	SVD	.270	.424	.245	.364	.264	.418
	NMF	.284	.560*	.280*	.556*	.288	**.581***
	DEN	.276	.566*	.273*	.526*	.280	.570*
	CV	**.310***	**.573***	**.287***	**.589***	**.312***	.540*
google	Default	.354	.526	.358	.517	.346	.535
	SVD	**.355**	.527	.359	.518	.348	.536
	NMF	.353	**.596***	**.367**	**.608***	.366	**.609***
	DEN	.343	.559*	.361	.555*	.356	.560*
	CV	.352	.561*	.362	.573*	**.374**	.556*