CIMeC UNIVERSITÀ DEGLI STUDI DI TRENTO CLIC Lab

Building a *bagpipe* with a *bag* and a *pipe*: Exploring Conceptual Combination in Vision

S. Pezzelle, R. Shekhar, R. Bernardi

{firstname.lastname@unitn.it}
CIMeC, DISI (University of Trento)

Motivation

Language

Noun-noun compounds (*clipboard, bagpipe,* etc.) are (to some extent) **compositional**: using even simple additive cDSMs we obtain composed vectors that approximate the ones extracted from text corpora

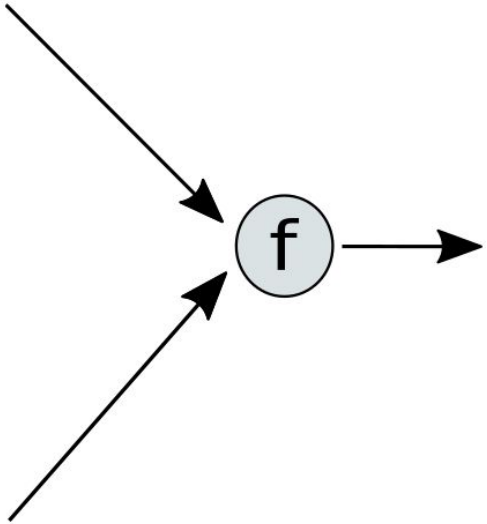
Vision

Can we do the same in vision? That is, does the visual representation of a nn-compound rely on the **combination of its parts**? If so, to what extent (for which cases) does it work?

Visual Composition

clip

board



clipboard

Hypothesis

We should be able to obtain the visual representation of a *clipboard* by combining *clip* and *board* with a **compositional function**, similarly to language

Our hypothesis is that a **simple additive model** should work for some cases, i.e. when the parts are still visible in the representation of the whole:

- superimposed cases (object in a background, e.g. *airplane*)
- concatenation of parts (whole resulting from the parts, e.g. *clipboard*)

Experiment

1. We manually built a dataset of images of visually-compositional nn-compounds (*compositional group*)
2. We randomly chose a *control group*
3. We extracted visual features for NN, N1, N2 with ConvNets (VGG-19 model)
4. We applied additive model to N1,N2 to obtain composed representations N+N
5. We evaluated over:

    a) Cosine similarity between observed and composed NN vectors (**Sim**)

    b) Retrieval (**Rec@k**) with k=1 and 5

    c) **CompInfo**: Sim - similarity observed NN/closest N (working when >0)

Vision vs Language

For each N, NN in the dataset, we built CBOW linguistic vectors with **word2vec**

We experimented with the same additive model and the same evaluation measures

Results

Vision:

Composition works in *compositional group* (76.6%) but not in *control group*

Language:

Same performance in *compositional group* but higher in *control group* (58.3%)

Dataset	$Avg.Similarity$		$\%(CompInfo > 0)$		$Rec@1$		$Rec@5$	
	Vision	Lang	Vision	Lang	Vision	Lang	Vision	Lang
Full	0.6283	0.407	62%	72%	0.34	0.52	0.76	0.88
Compositional	0.6476	0.429	76.31%	76.31%	0.3947	0.57889	0.8158	0.9211
Control	0.5671	0.3377	16.66%	58.33%	0.1667	0.3333	0.5833	0.75

Conclusions

Simple additive model in Vision **works for manually selected cases** (superimposed or concatenated items) but not for combinations where subtler, more abstract interactions are involved (eg. *corkscrew*)

Need for large, annotated datasets

Need for new, more complex compositional models (perhaps combining the 2 modalities) designed for solving the task

Thank you for your attention!