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

Sandro Pezzelle, Ravi Shekhar, Raffaella Bernardi
{firstname.lastname}@unitn.it

Motivation
Conceptual combination is the cognitive process by which two existing concepts are combined to form new complex concepts [1].

In language, this mechanism can be observed in the formation and lexicalization of compound words like boathouse, swordfish, etc.

Composition of concepts/words is something more than a simple addition, but additive models are effective in language (DSMs) [2].

Research question
Can the visual representation of a complex concept (clipboard) be obtained by summing up its parts (clip, board) as in language?

We expect this procedure to work in some cases (parts still visible), but fails where more abstract operations are needed.

Dataset
List of noun-noun compounds annotated for imageability from [3].

1. Filtering based on imageability > 5 (visually well-defined items)
2. Only genuine noun-noun combinations were retained
3. Selection driven by average quality of top-25 Google images for both the compound (bagpipe) and its constituents (bag, pipe)

→ resulting list including 115 items.

Dataset construction
- compositional group: 38 manually-selected items involving either superimposition (air+plane) or concatenation (bag+pipe)
- control group: 12 randomly-selected from the 115-item list
- full group: compositional + control group (50 items)

→ for each nn-compound and noun in full group: 1 good image

In total: 50 nn-compound images + 79 noun images (129).

Model
We test a simple additive model in both Vision and Language:

bagpipe = bag + pipe

Visual features
For each image: 4096-dimension vector extracted using ConvNets (VGG-19 pretrained on ImageNet, fc6 layer) [4].

Linguistic features
For each word: word2vec 400-dimension vector trained on 3-billion tokens corpus (ukWac + Wikipedia + BNC).

Evaluation
To evaluate the compositional model, we use three measures:

1. Cosine similarity between observed (clipboard) and composed vector (clip+board)
2. CompInfo: difference between composed-observed similarity and observed-closest noun similarity (e.g., clipboard and clip)
   Thus, composition works with CompInfo > 0.
3. Rec@k: retrieving the observed vector in the semantic space using the composed one (with k=1 and k=5).

Results
Vision
1. Composition works in 76.31% cases in the compositional group vs 16.66% cases in the control group (CompInfo > 0).
2. Both similarity and Rec@k are higher in compositional group.

Language
1. Composition works in 76.31% and 58.33% cases, respectively.
2. Both similarity and Rec@k are higher in compositional group.

References
[2] D. Paperno and M. Baroni. To appear. When the whole is less than the sum of the parts... Accepted for publication in Computational Linguistics.