

TRANSPARENT ANALYSIS OF MULTI-MODAL EMBEDDINGS

Anita L. Veró and Ann Copestake
Computer Laboratory, University of Cambridge



1. Resources for Multi-Modal Semantics

Distributional models suffer from the *grounding problem*:

Grounding problem: the fact that the meaning of a word is represented as a distribution over other words does not account for the fact that human semantic knowledge is *grounded in physical reality and sensorimotor experience*. (Harnad, 1990)

Multi-modal semantics addresses this by *enhancing* linguistic representations with extra-linguistic perceptual input, usually using **images**.

Open questions about representation learning techniques and data sources:

Does visual data bolster performance on **non-visual tasks**?

If it does, is this only because we add more data or does it convey complementary **quality** information compared to a higher **quantity** of text?

Can we achieve comparable performance using **small-data** if it comes from the right data distribution? Is the modality, the size or the distributional properties of the data that matters?

2. Mid-fusion and Evaluations

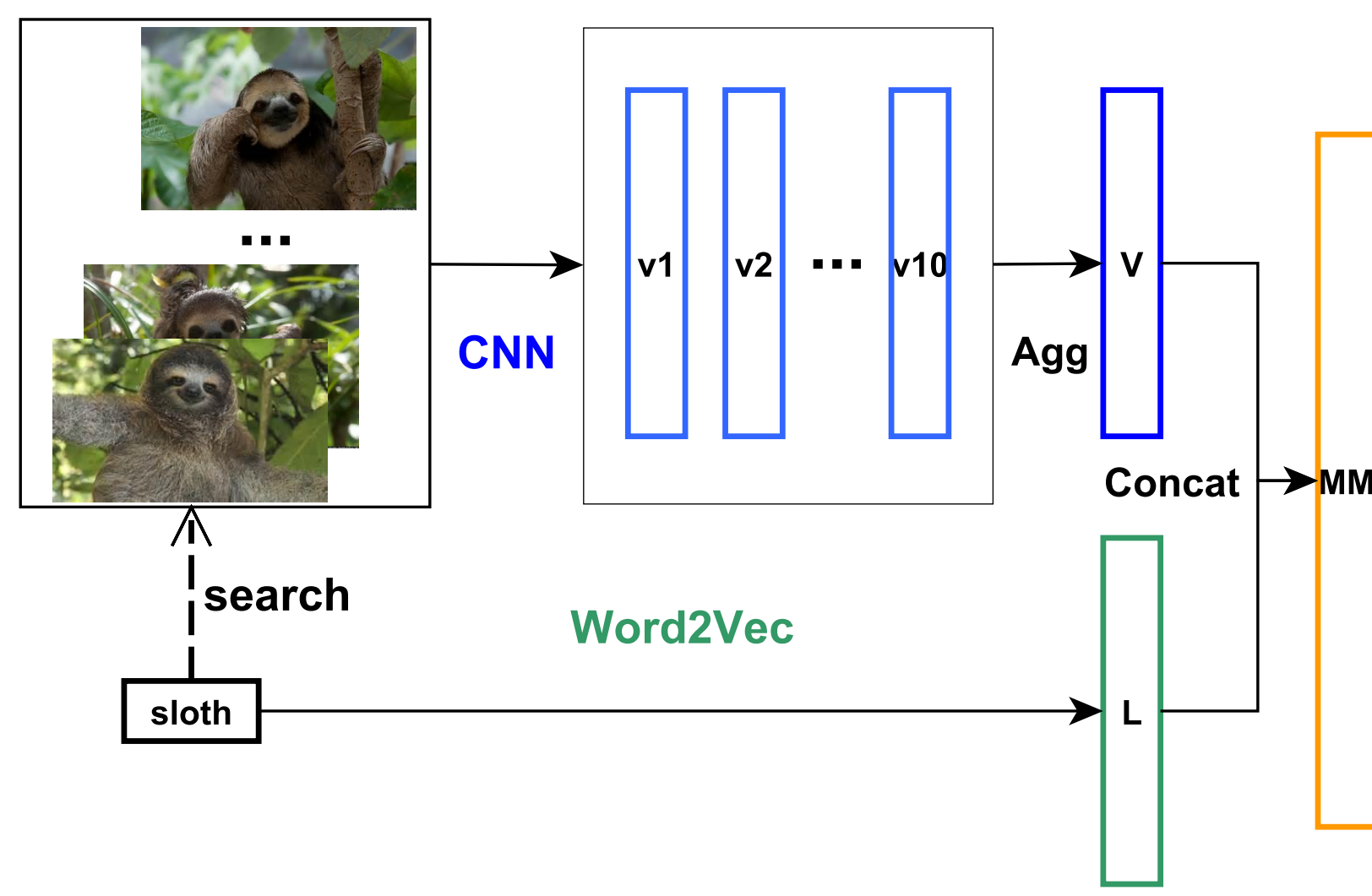
Visual representations

- Transfer last fully-connected layer **convolutional network features**.

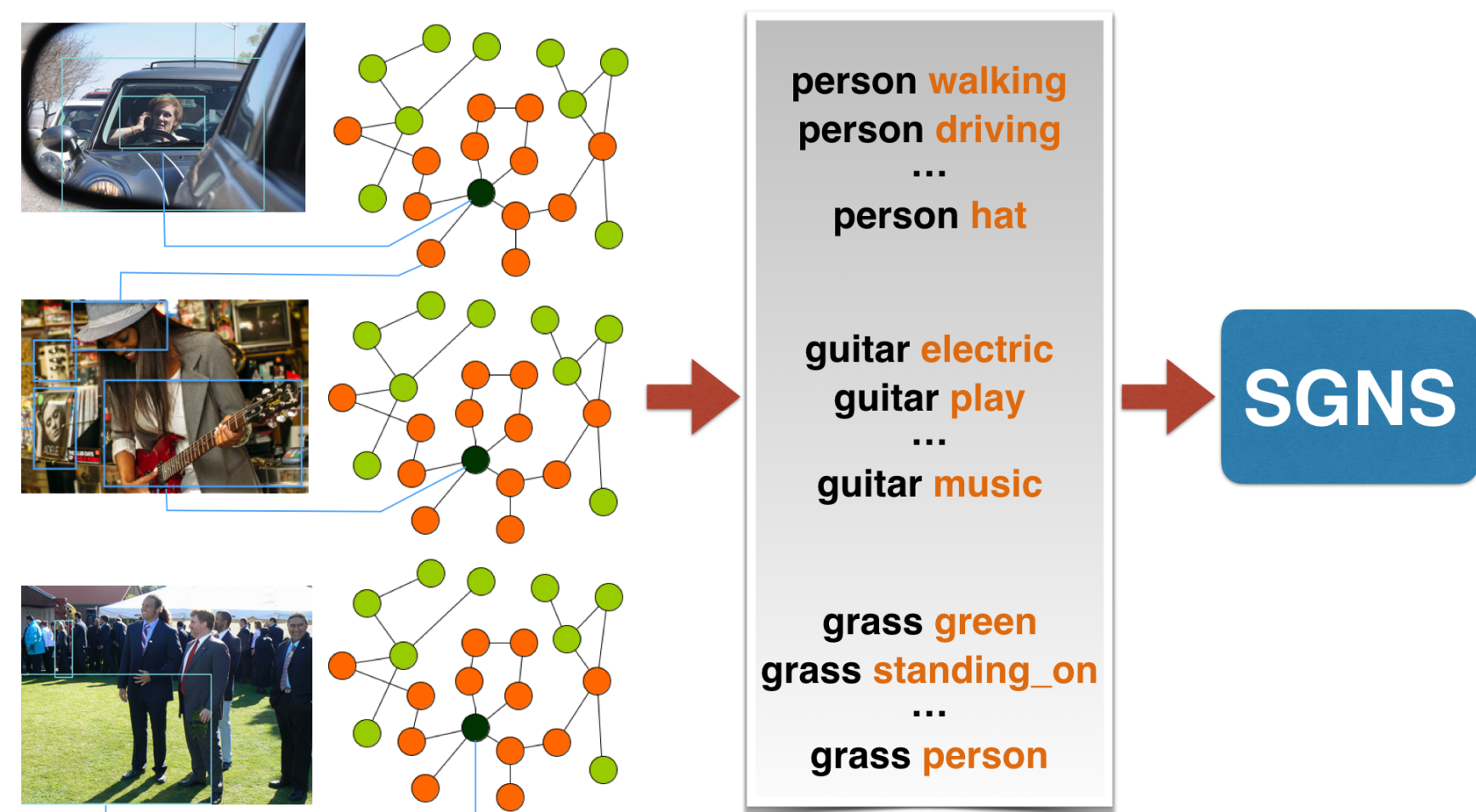
Aggregating image vectors for one word

Multi-modal representation: concatenating visual and textual vectors.

Evaluations: Word similarity and brain imaging.

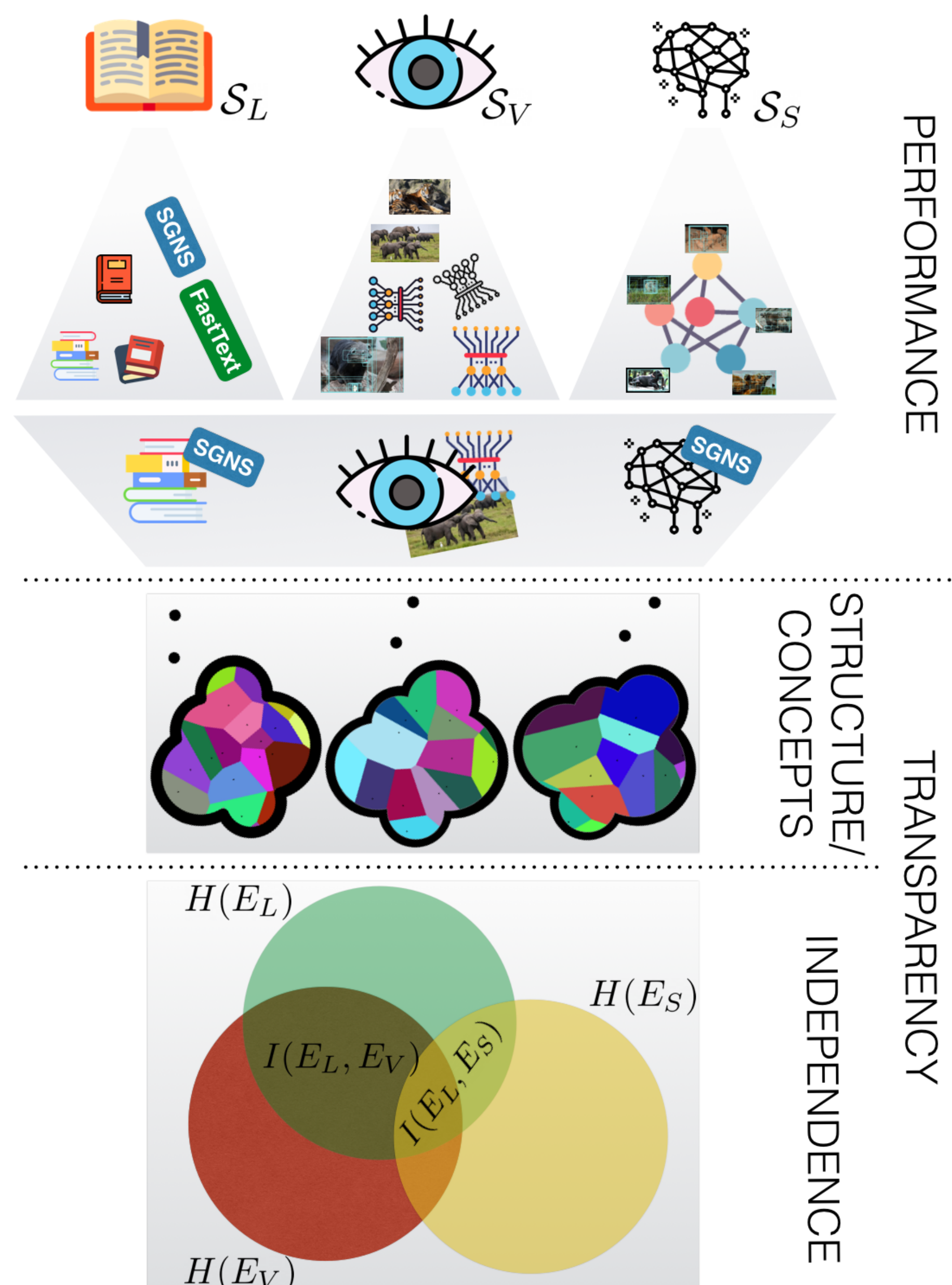


3. Visually Structured Graph Modality



4. Three-pillar Analysis

[EmbEval Toolkit]



5. Models and Data Sources

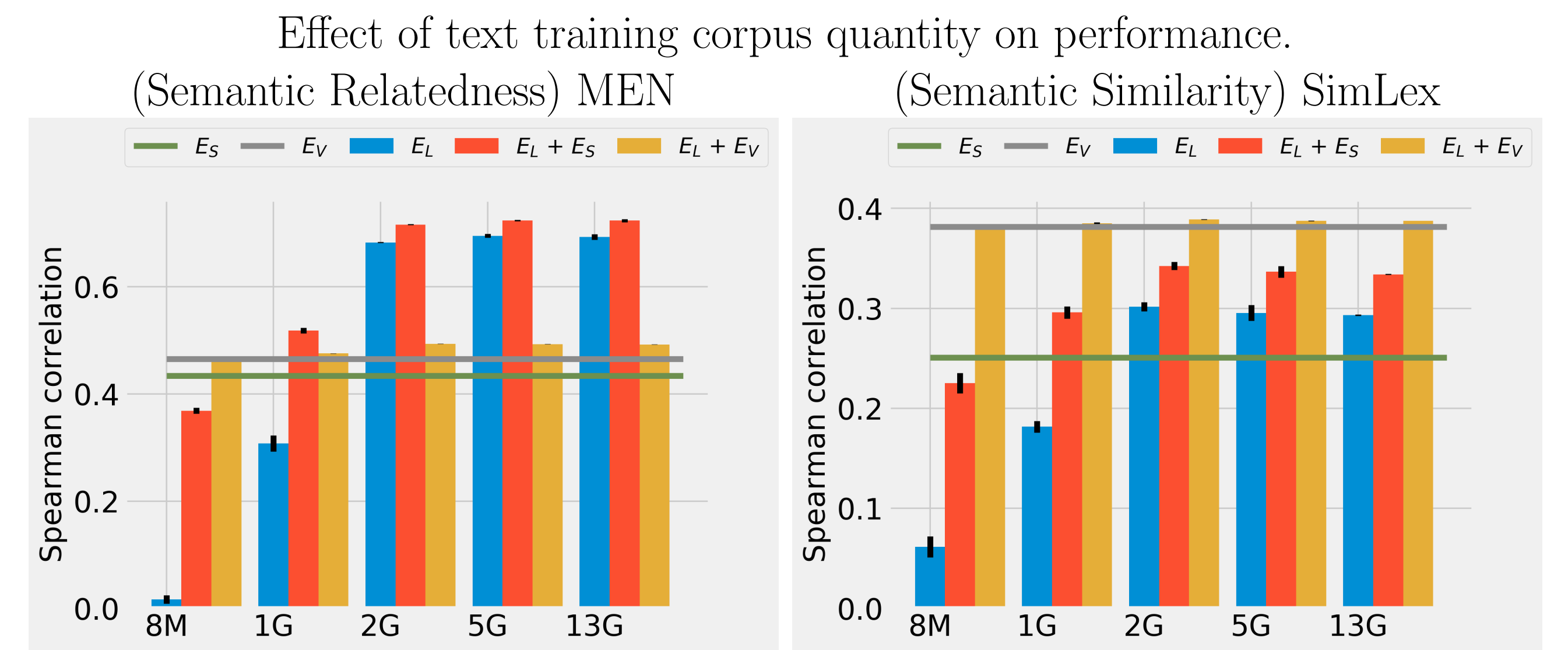
CNNs: **AlexNet**, **GoogLeNet**, **VGGNet**, **ResNet**

Image Datasets: **Google**, **Flickr**, **Bing**, **VisualGenome**

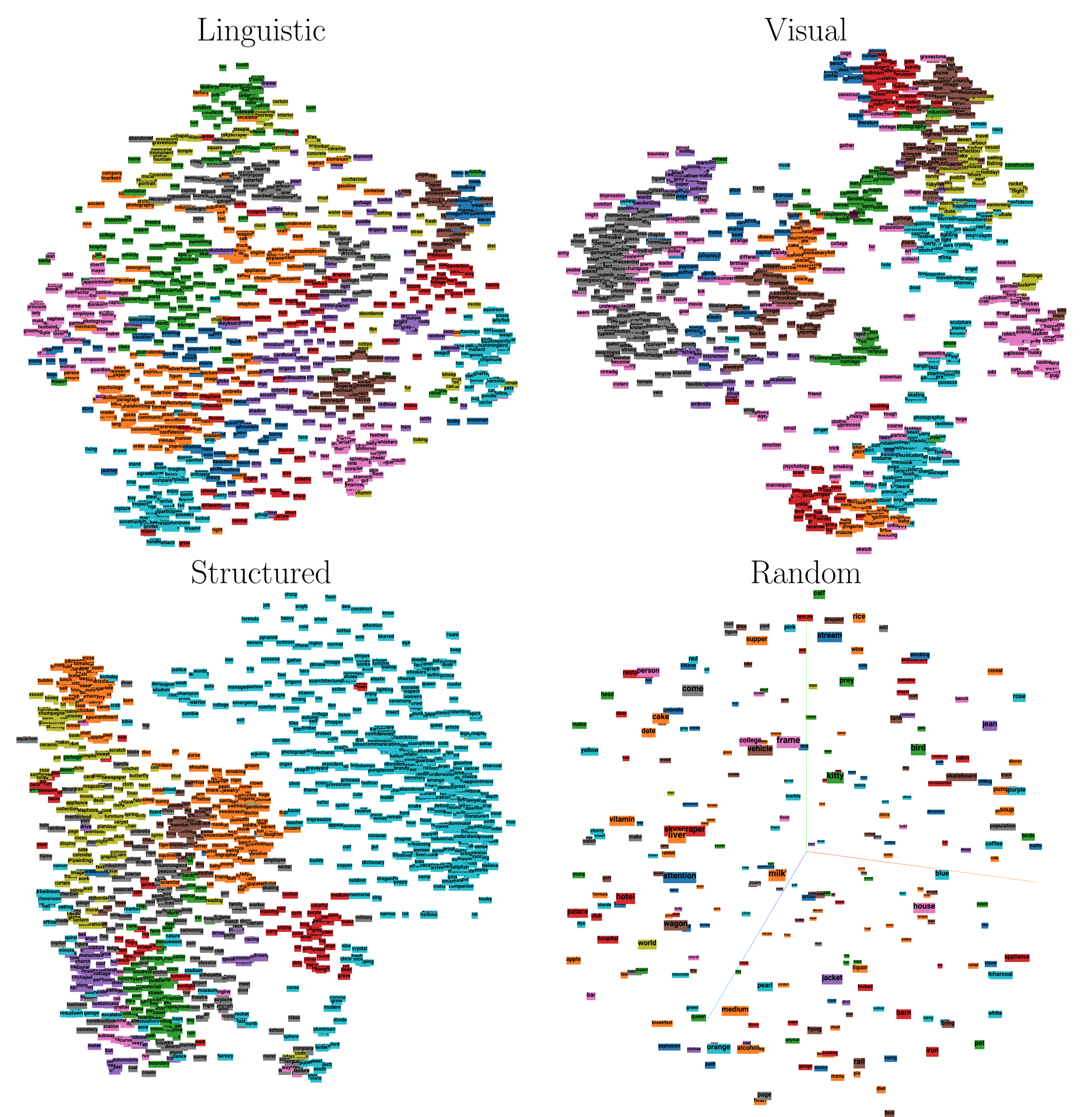
Linguistic models: **Skip Gram Negative Sampling**, **Structured SGNS**

Text Corpora: **Wikipedia**, **Common Crawl**, **Wiki News**

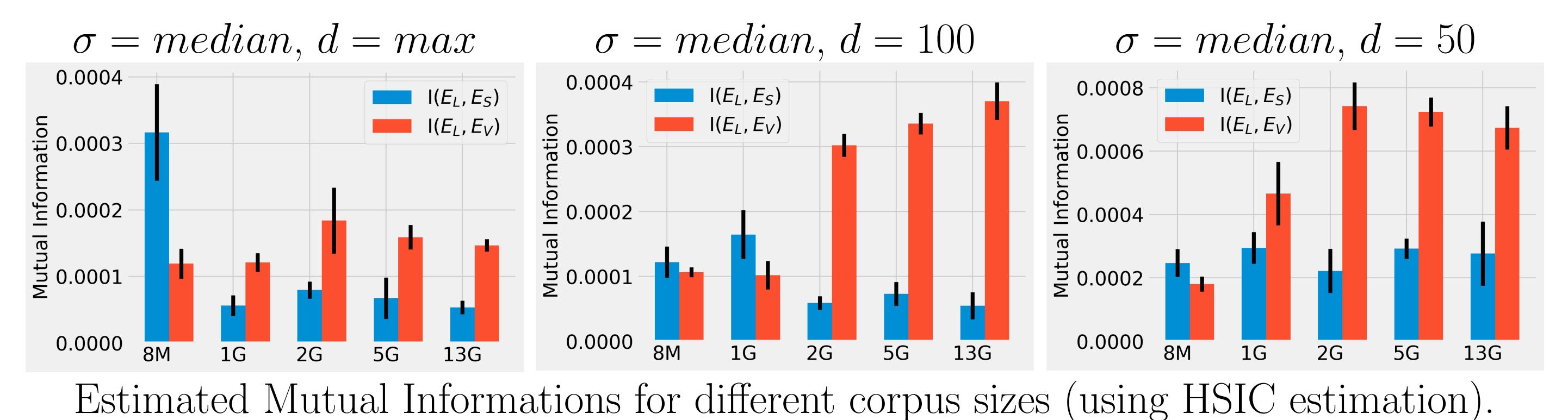
6. Performance and Efficiency



7. Structural Analysis of Model-Concepts



8. Independence Analysis



9. Conclusion

- The **source** of images affect the performance.
- The **number of images** in ordered sources stabilizes at around 10-20 images.
- Visual information is complementary** for smaller corpora, but this effect does not scale with corpus size.
- There is no direct indication of the impact of **low level** visual features.
- Structured model** achieves comparable performance in an **economic** way, using orders of magnitude less resources than visual models. It conveys more divergent information. Its clusters represent concrete concepts, in-between visual and linguistic domains.