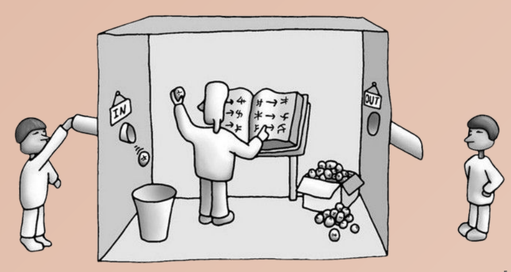
Exploring Visually Grounded Word Embeddings

Milan Miletić, Ryan Ott, Adrian Sauter

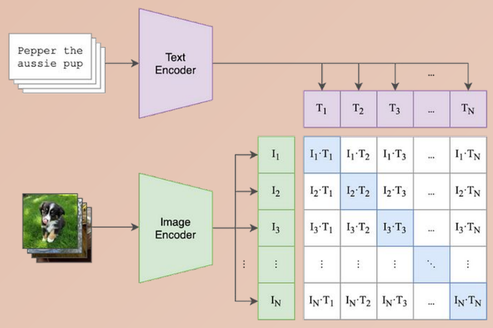
Context

RED
YELLOW
GREEN
BLUE

Stroop task [3]



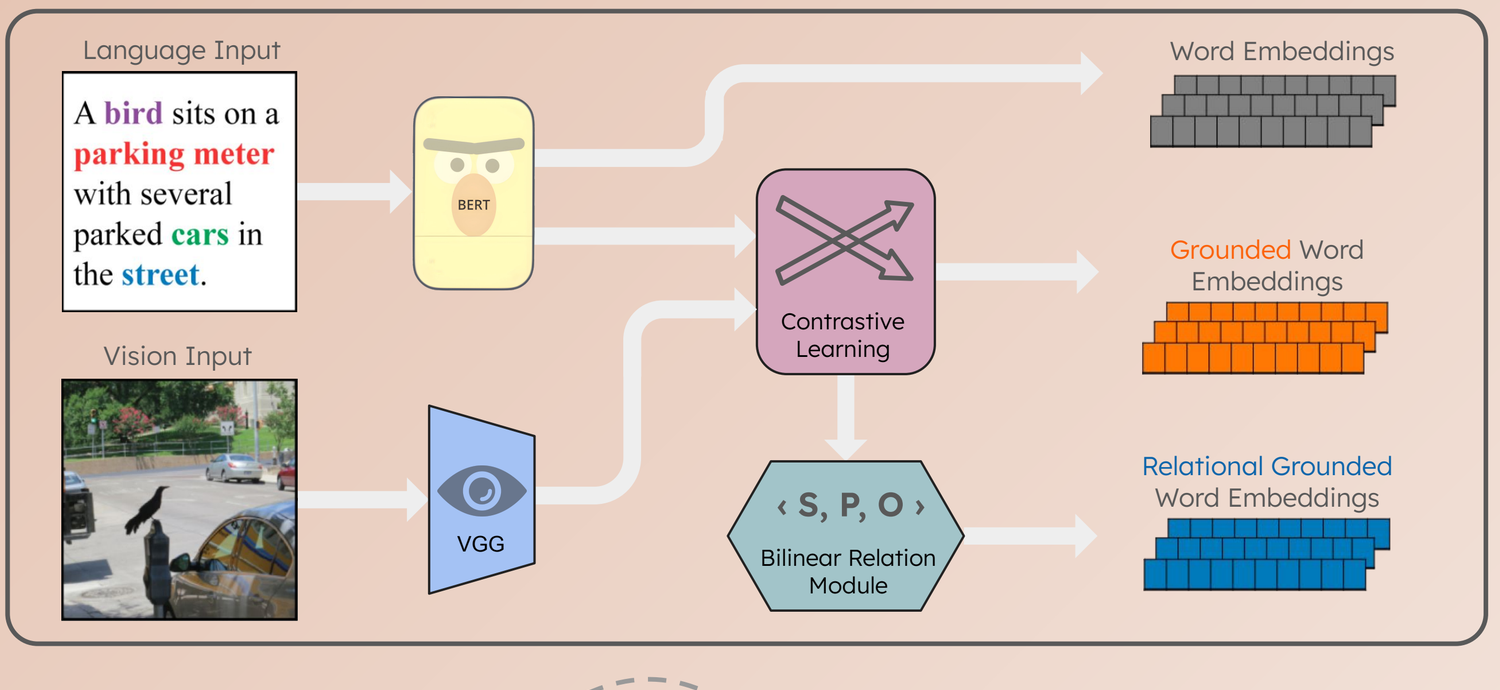
Chinese room argument [2]



openai/CLIP [4]

Paper

"Explainable Semantic Space by Grounding Language to Vision with Cross-Modal Contrastive Learning" (Zhang et al., 2021) [1]



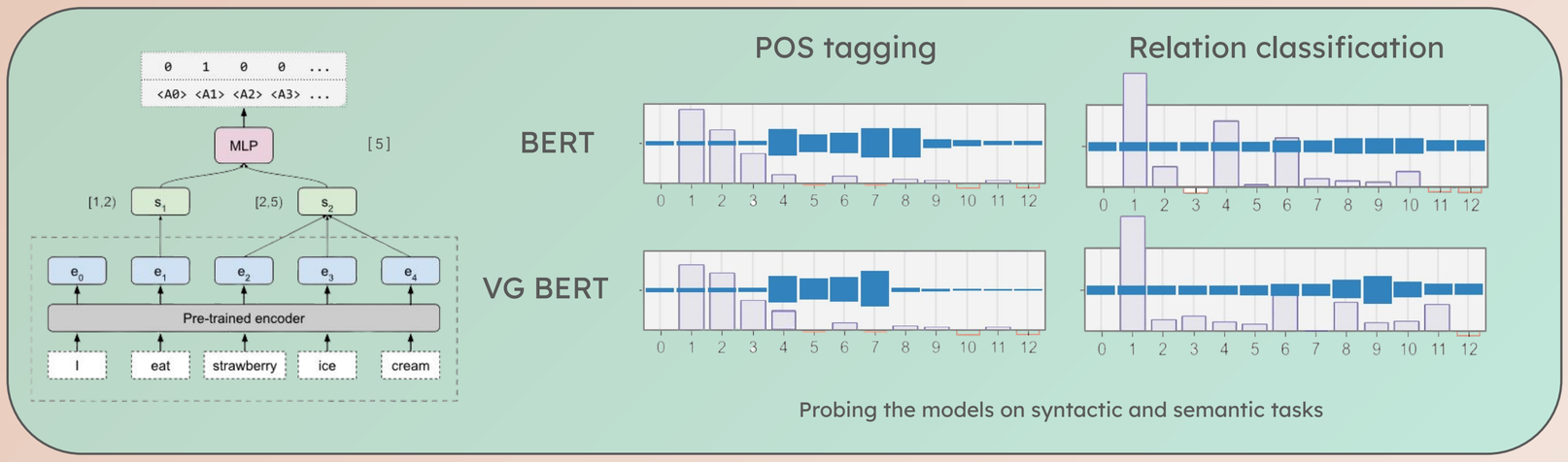
Research Questions

1. How is information propagated through the VG encoder?
2. Does VG equally impact concrete and abstract concepts?
3. Does VG help with resolving lexical ambiguities?
4. Are the grounded embeddings clustered better?

Results

	Probing		Lexical Ambiguity	Clustering
	POS	Rel. ↑	Cosine sim. ↓	Silhouette coef. ↑
BERT	0.96	0.49	0.56 ± 0.13	0.02
VG BERT	0.96	0.52	0.34 ± 0.15	0.31

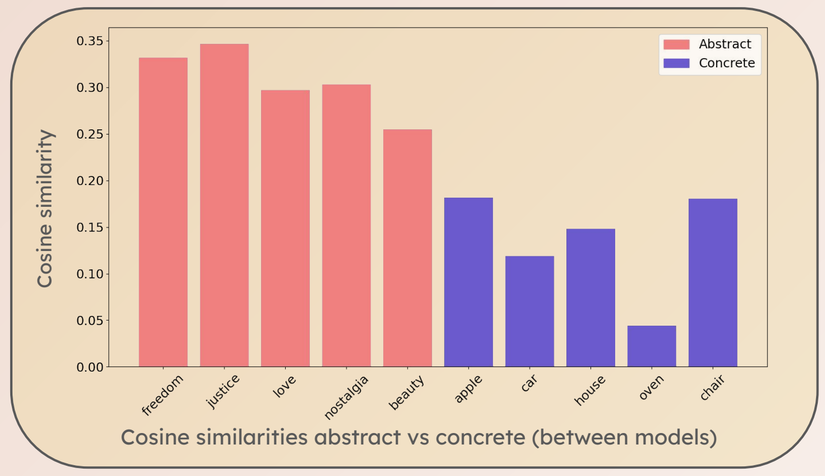
Probing



[5]

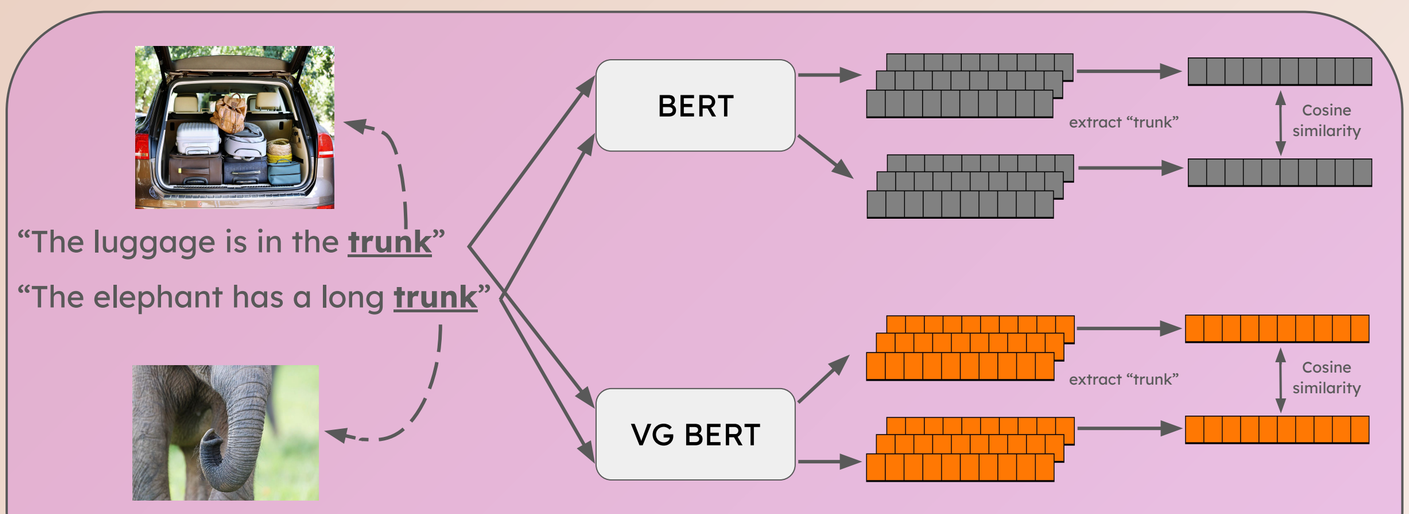
Probing the models on syntactic and semantic tasks

Abstract vs. Concrete



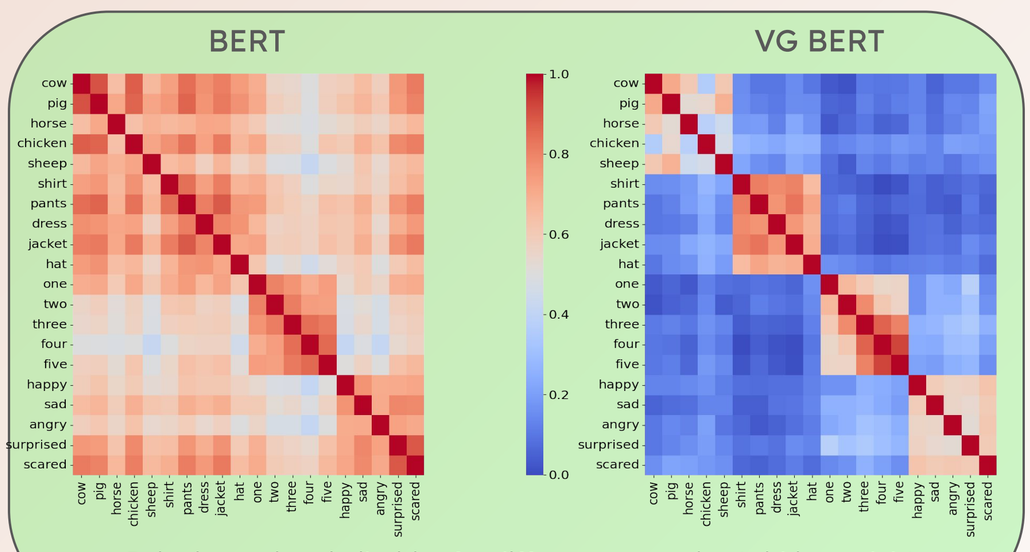
Cosine similarities abstract vs concrete (between models)

Lexical ambiguity



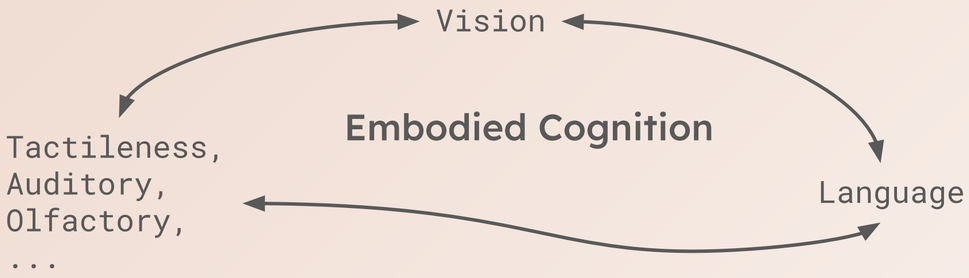
Testing the disambiguation abilities

Semantic Clustering



Pairwise cosine similarities for different categories (within models)

Discussion / Future Work



Conclusion

VG BERT still adheres to the classical NLP pipeline - 1.

VG more strongly affects concrete concepts - 2.

Contrastive Learning

Better homonym distinction in VG embeddings - 3.

VG embeddings better clustered into categories - 4.

References

[1] Zhang et al. Explainable Semantic Space by Grounding Language to Vision with Cross-Modal Contrastive Learning. 2021
[2] John R Searle et al. Minds, brains, and programs. The Turing Test: Verbal Behaviour as the Hallmark of Intelligence. 1980.
[3] J Ridley Stroop. Studies of interference in serial verbal reactions. 1935.
[4] Radford & Kim et al. Learning transferable visual models from natural language supervision. 2021
[5] Tenney et al. BERT Rediscovers the Classical NLP Pipeline. 2019