CMPT 983

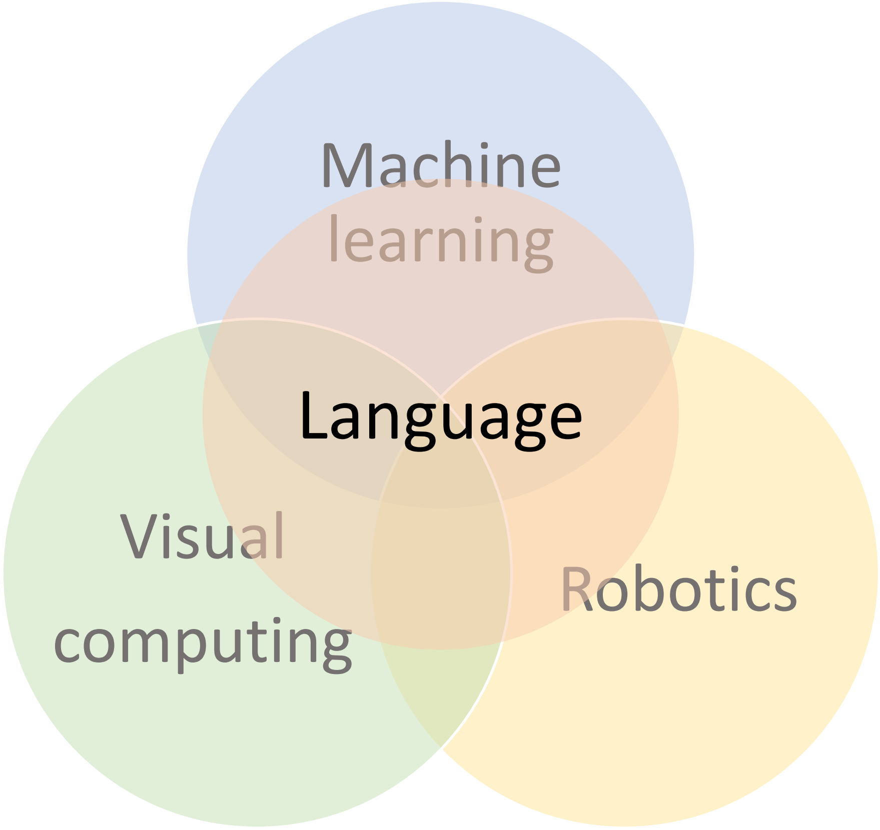
Grounded Natural Language Understanding

April 15, 2021

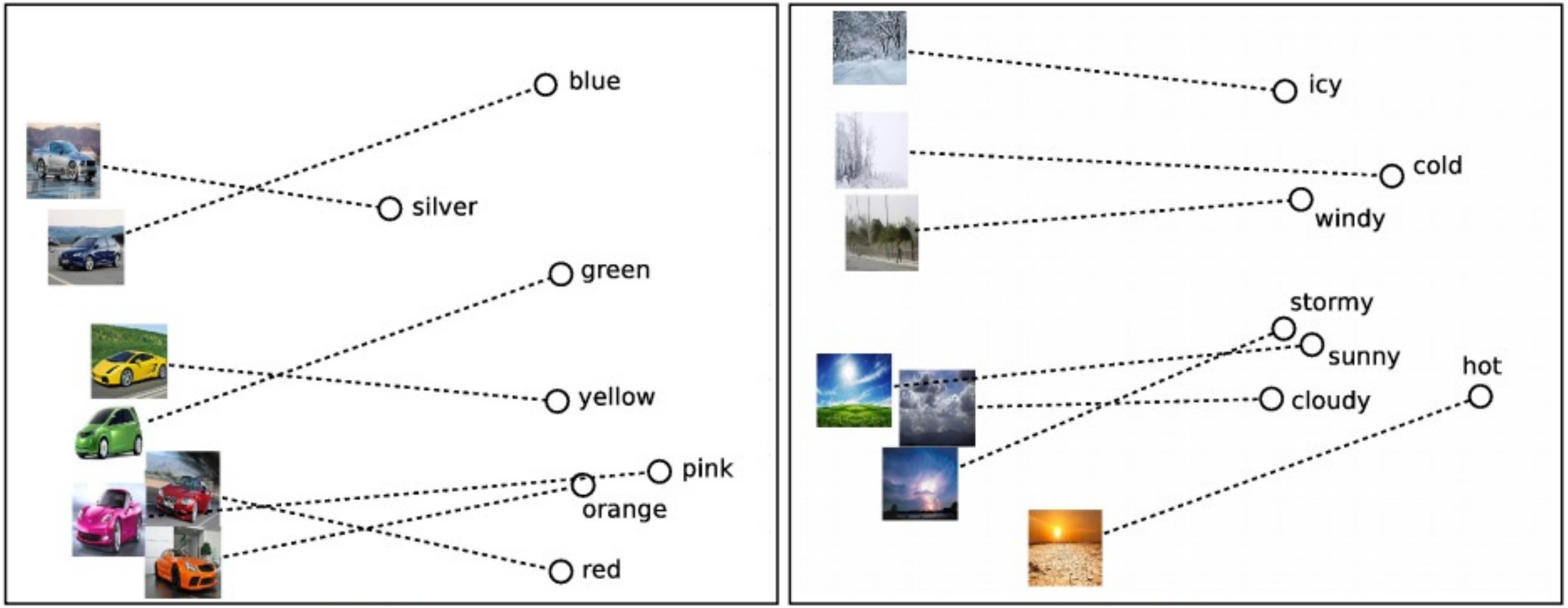
Conclusion

Grounded natural language understanding

- Lightening tour of topics at the intersection of language and machine learning, visual computing and robotics



Multimodal Embeddings



“Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models”
[Kiros, Salakhutdinov, Zemel TACL 2015]

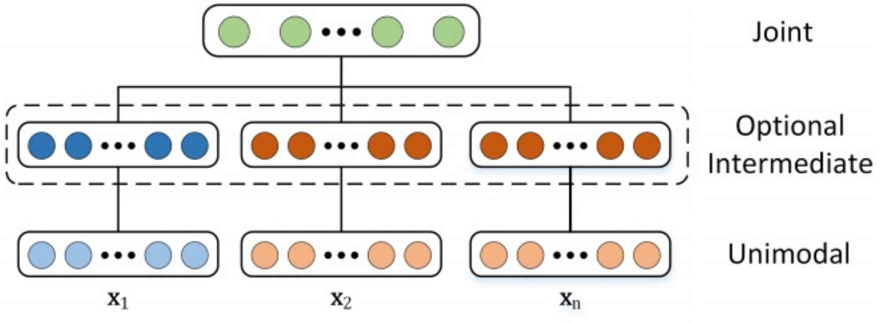
Multimodal representations

- Joint vs Coordinated representations
  - Joint: Autoencoder + Fusion (e.g. concat)
  - Coordinated: CCA, joint embeddings

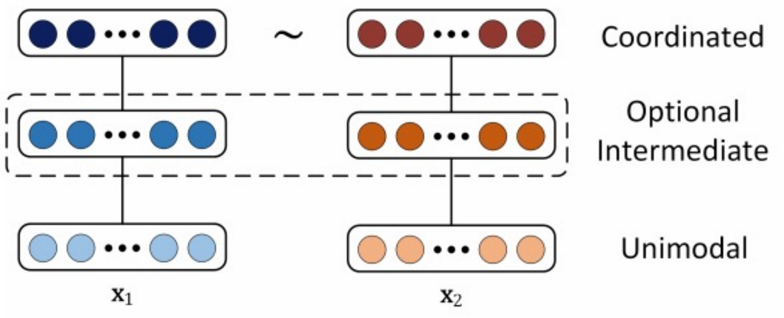
Correct label (more similar) Other labels (less similar)

$$\mathcal{L}_C(\mathbf{W}, \mathbf{U}, I_i, y_i) = \sum max\{0, \alpha - \underbrace{D(\Psi(I_i), \mathbf{u}_{y_i})}_{\text{Correct label (more similar)}} + \underbrace{D(\Psi(I_i), \mathbf{u}_{y_c})}_{\text{Other labels (less similar)}}\}$$

$$\mathbf{x}_m = f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$



$$f_1(\mathbf{x}_1) \quad f_2(\mathbf{x}_2)$$



- Useful for retrieval, translation

Attention

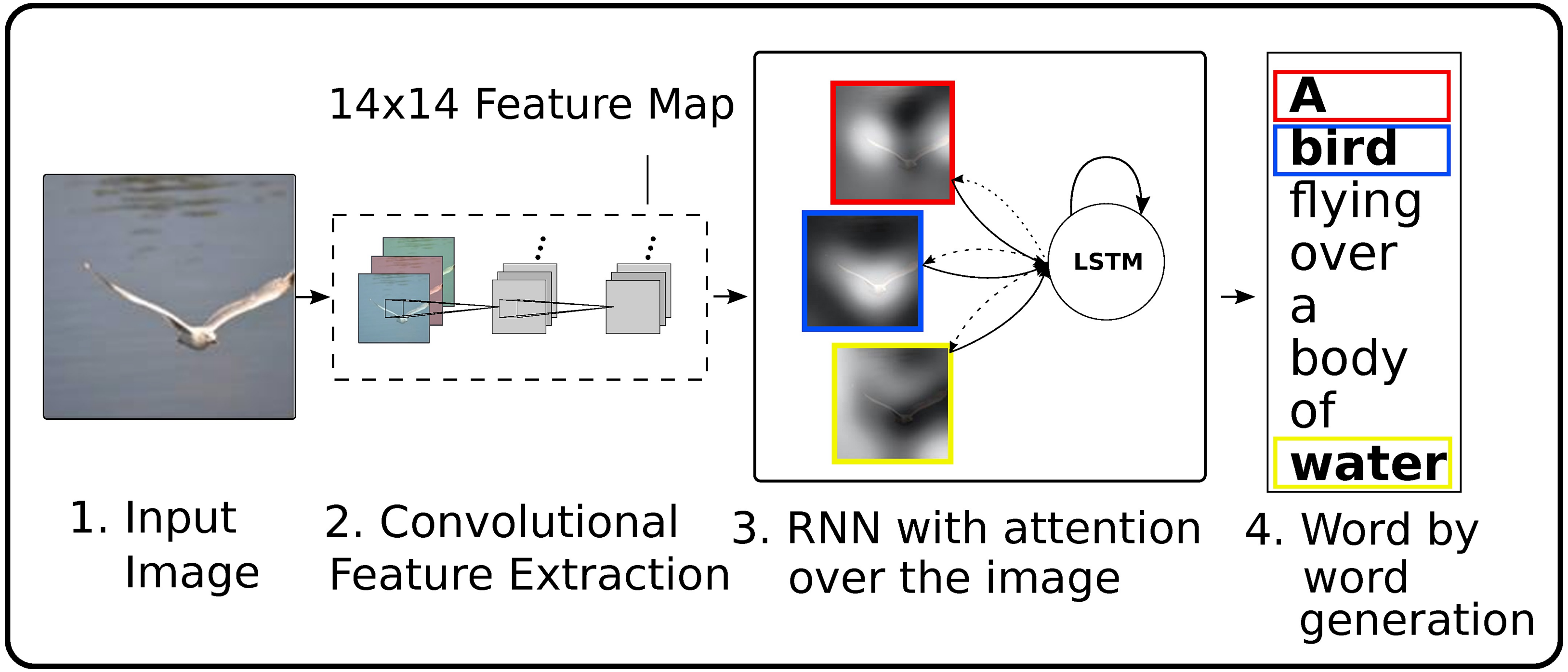
- Not every part of the input given the task context

Exit the bathroom. **Turn left and exit the room using the door on the left.** Wait there.

Slide credit: Stefan Lee

Attention

- Used for many vision and language tasks
- Including captioning and understanding referring expressions
- Representation that weighs different parts of the input differently



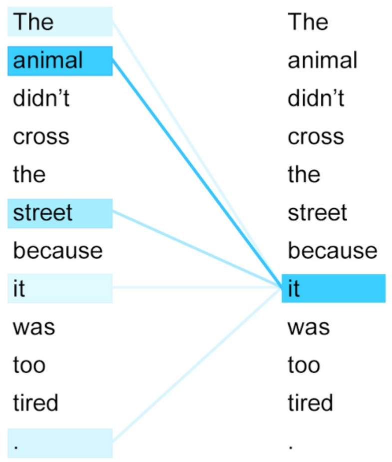
Show, Attend and Tell, Xu et al. ICML 2015

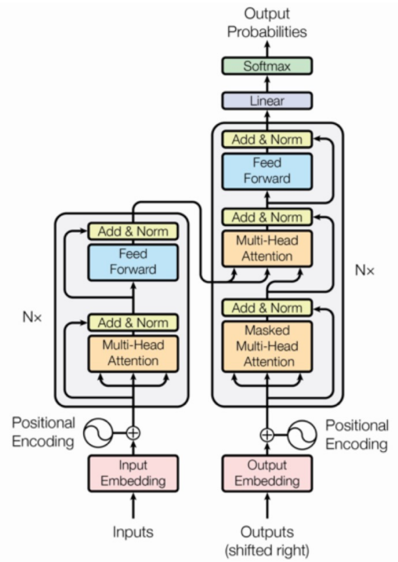
Attention

- Mathematically: weighted sum $\hat{\boldsymbol{v}} = \sum_{i=1}^{k} \alpha_i \, \boldsymbol{v}_i$
- Types of attention
  - Different ways to compute weight / similarity
  - Hard vs Soft
- Query-key-value view of attention
- Self-attention and transformers

Attention function, f

$$a_i = g(\boldsymbol{k}_i, \boldsymbol{q})$$
$$\boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{a})$$
$$\hat{\boldsymbol{c}} = \sum_{i=1}^{k} \alpha_i \, \boldsymbol{v}_i$$





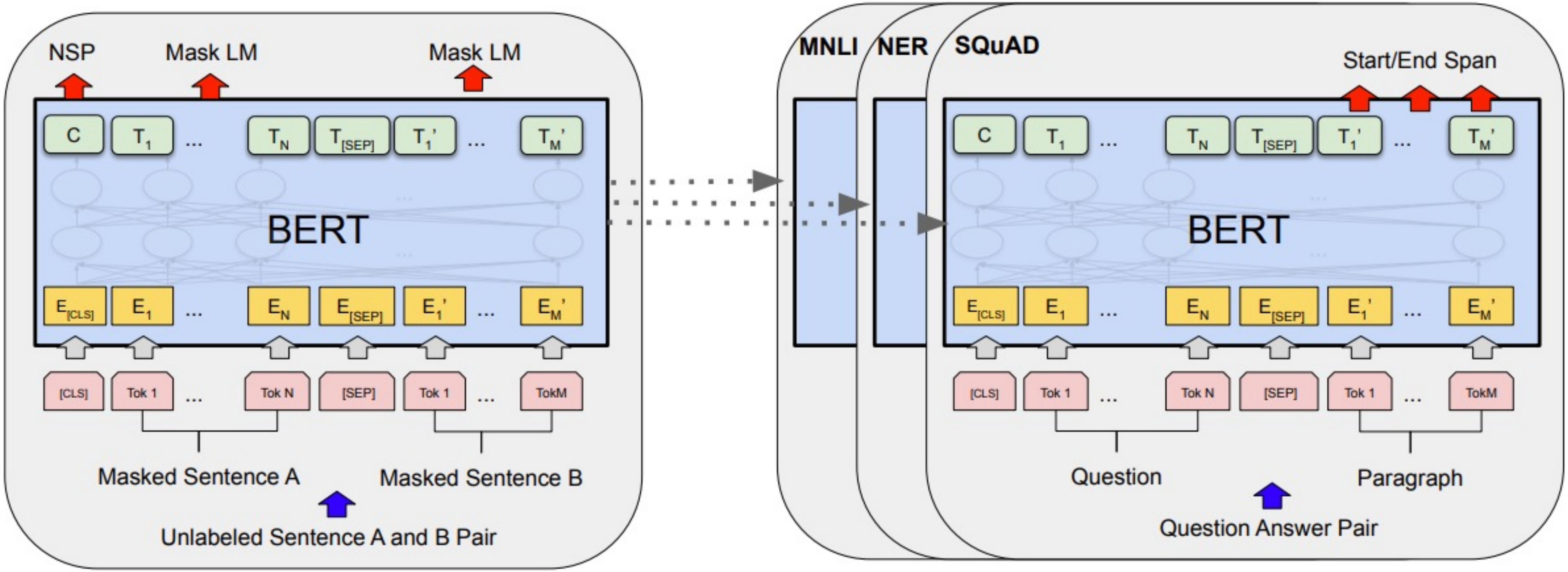
- Scaled dot-product attention:

$$g(c_i, z) = z^\top c_i / \sqrt{d}$$

Pretraining

Big pile of unannotated data!
Lots of resources to train!

Task specific
Small amount of annotated data
Start with pre-trained model

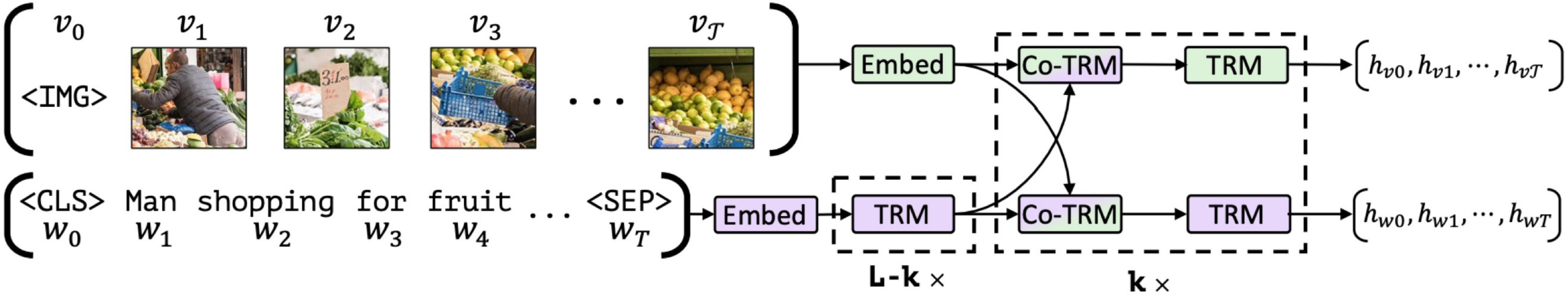


Pre-training

Fine-Tuning

BERT, Devlin et al, 2018

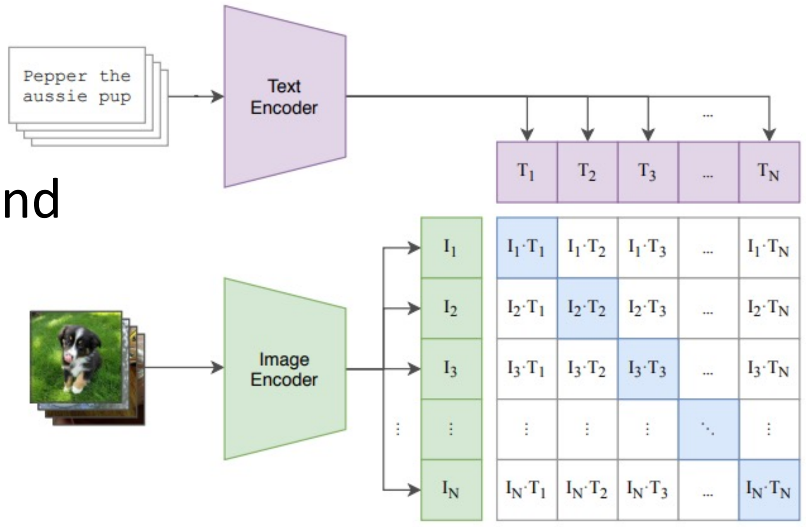
Pretraining and masked multimodal models



ViLBERT, Lu et al, NeurIPS 2019

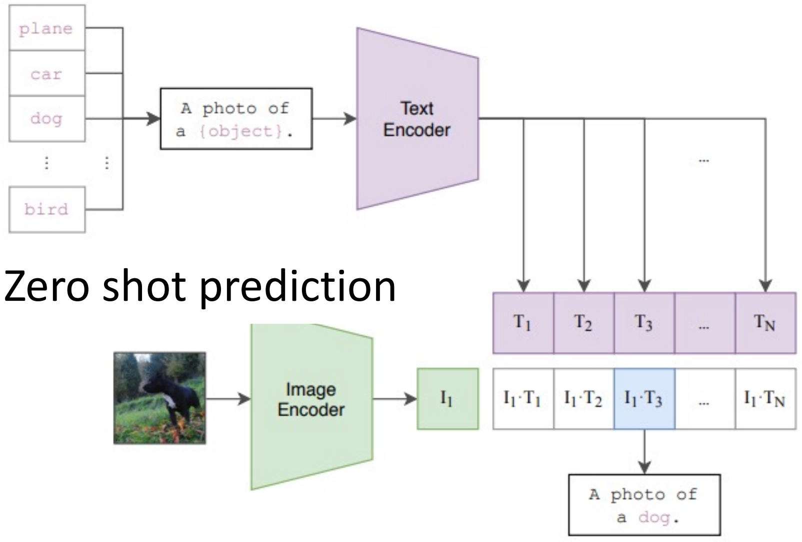
Contrastive pretraining

Does the image and text pair match?



Create classifier by generating captions and encoding

Zero shot prediction



CLIP, Radford et al, 2021

Structure and compositionality

- Compositionality

Compositional Generalization

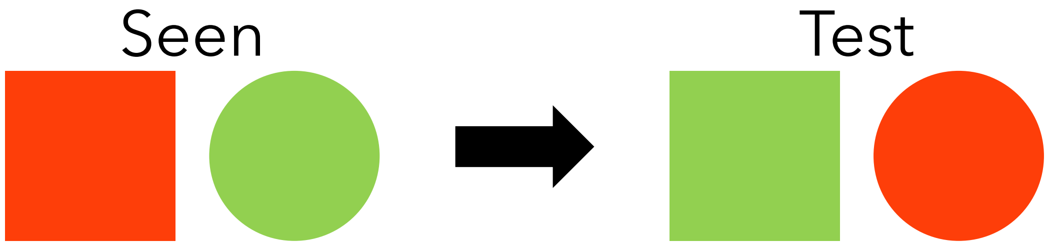
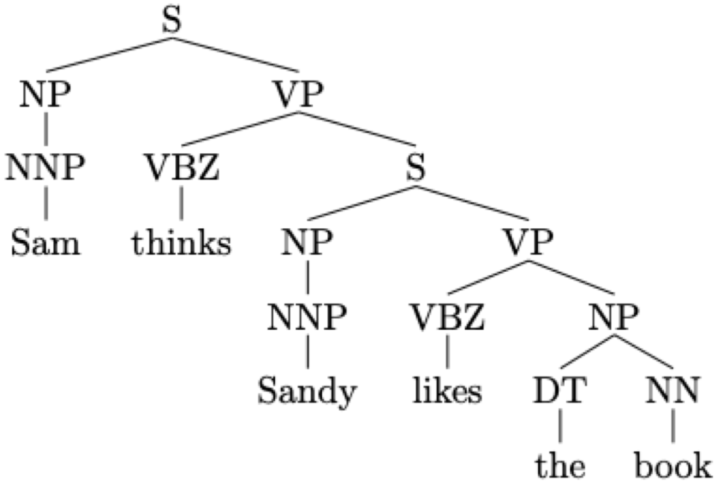


Image credit: Stefan Lee

- Structured representations for compositionality

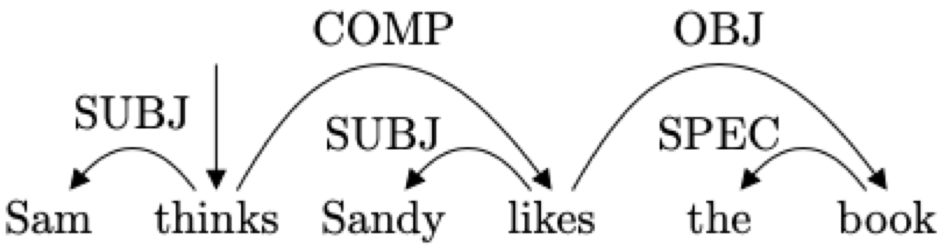
Constituency Parse Tree

Hierarchical



Dependency Parse

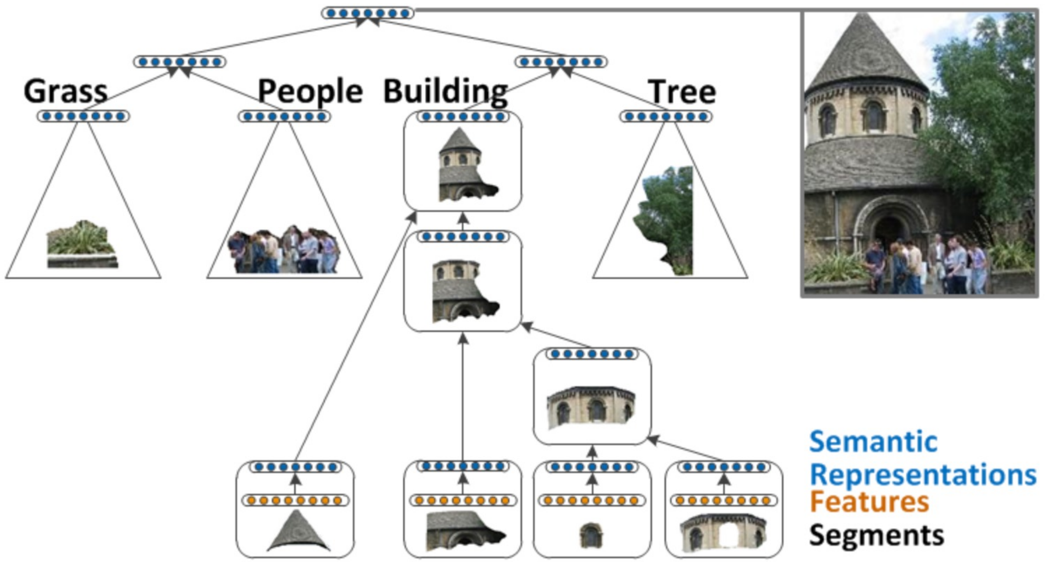
Relational



Structured representation of images

Scene Parse Tree

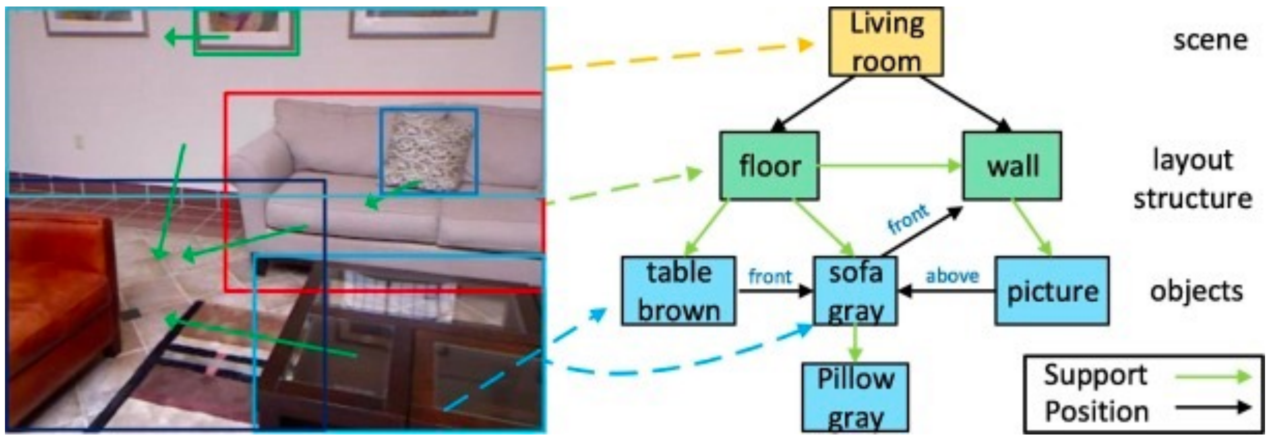
Hierarchical



Socher, Lin, Ng, and Manning, "Parsing Natural Scenes and Natural Language with Recursive Neural Networks", ICML 2011

Scene Graph

Relational



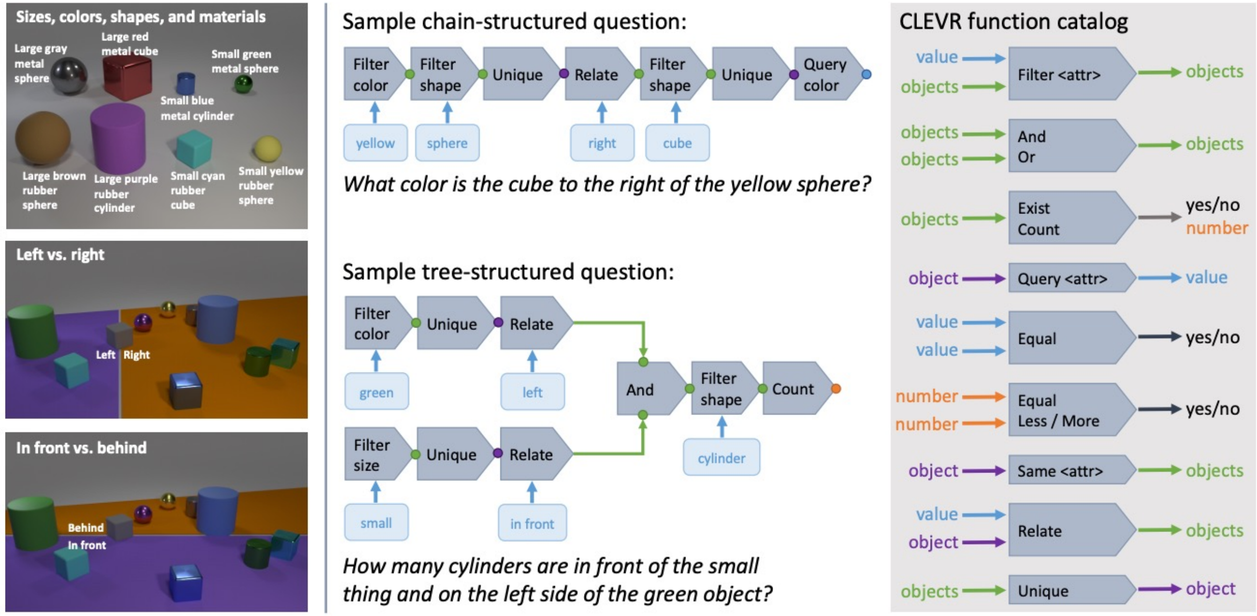
Yang, Liao, Ackermann, and Rosenhahn, "On support relations and semantic scene graphs", ISPRS Journal of Photogrammetry and Remote Sensing, 2017

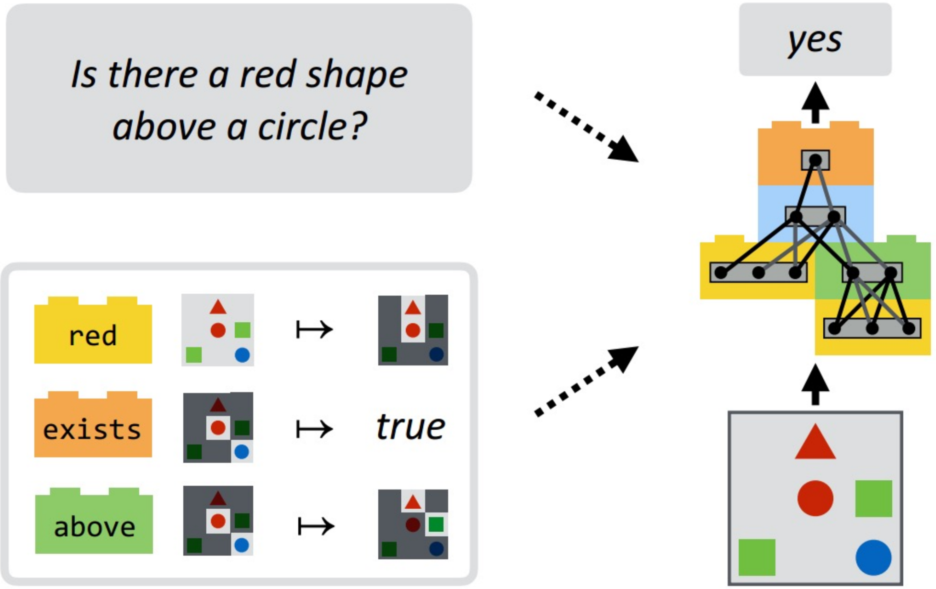
Semantic parsing

- Parse natural language into programs
- Use in VQA

Shape and attributes

Programs: formed from composable modules





Relations

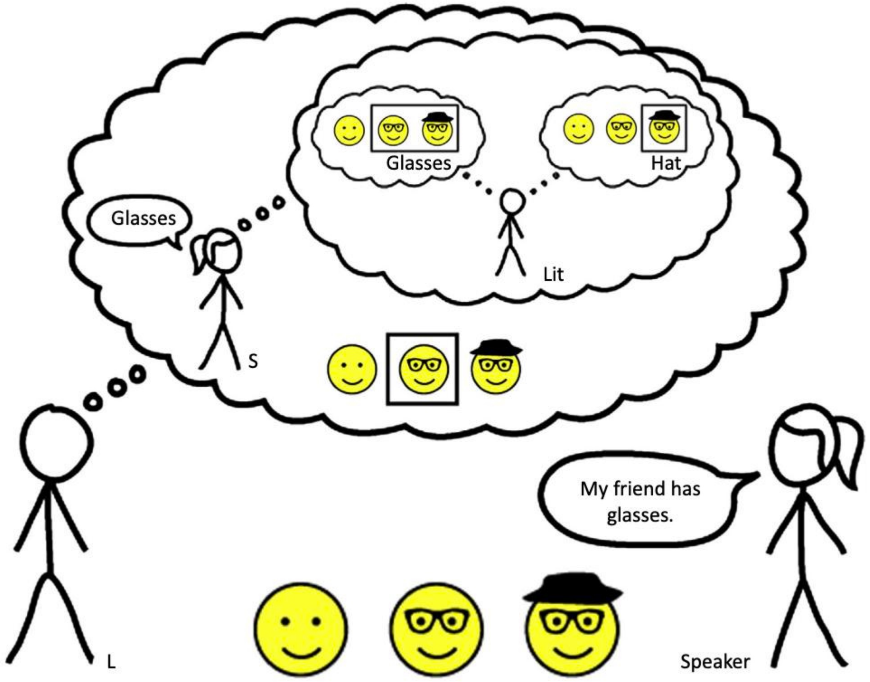
Generated language

Neural module networks, Andreas et al, CVPR 2016

CLEVR dataset, Johnson et al, 2017

Speaker-listener models

- Need to model other party
- Rational Speech Acts (RSA)
- Used in referring expression generation + comprehension
- Looked at ShapeGlot and emergent communications

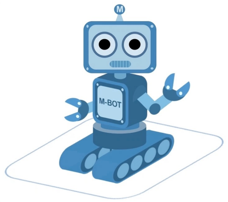


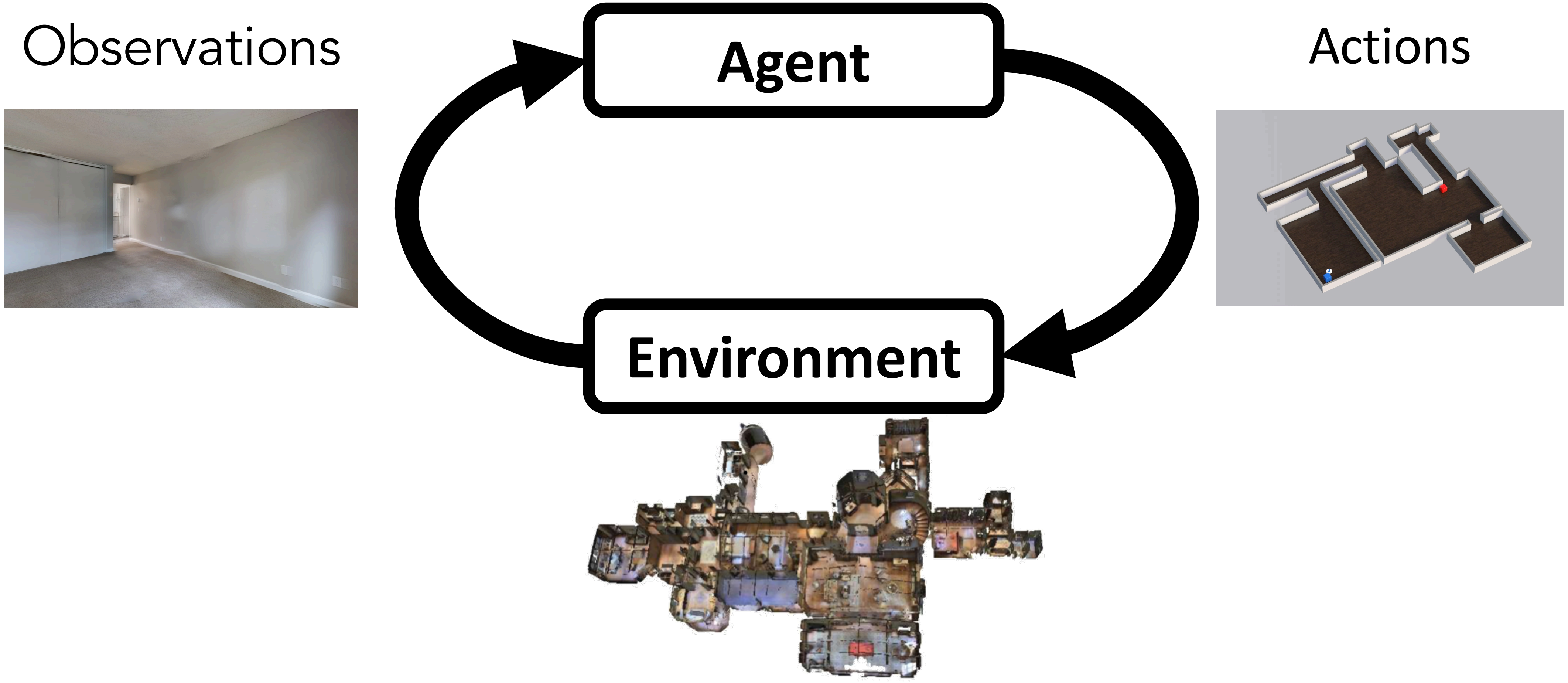
Goodman and Frank, 2016

Instruction following

- How to train agent to follow instructions?
- Can the agent learn language through interacting with the environment?

Exit the bedroom. Turn left down the hall and stop in the kitchen.





Instruction following (RoboNLP)

- Quick review of imitation learning and reinforcement learning
- Visual language navigation
- Instruction following with manipulation and interaction

ALFRED, Shridhar et al, CVPR 2020

- Lots of challenges:
  - Data, task specification, accurate simulation

Interactive language learning

- Language learning with feedback
  - Human or the environment
- Model weights are adjusted based on feedback

Text conditioned content generation

- Review of generative models

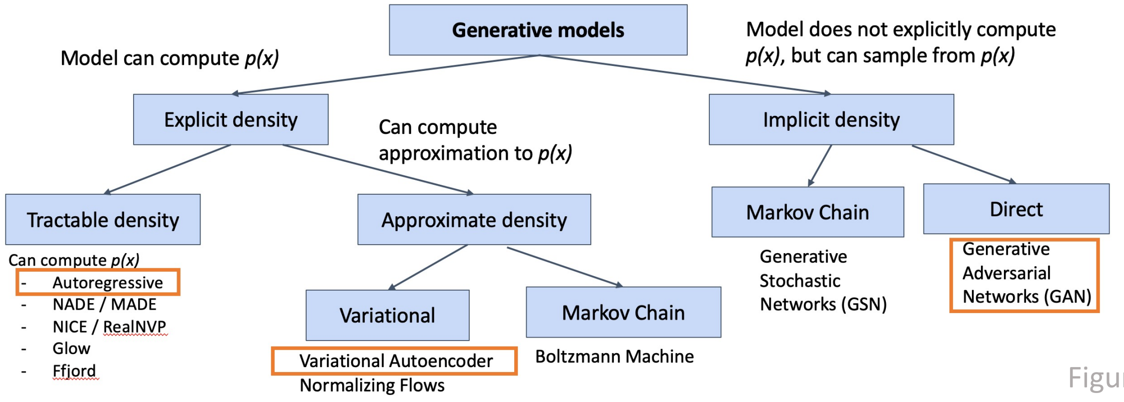
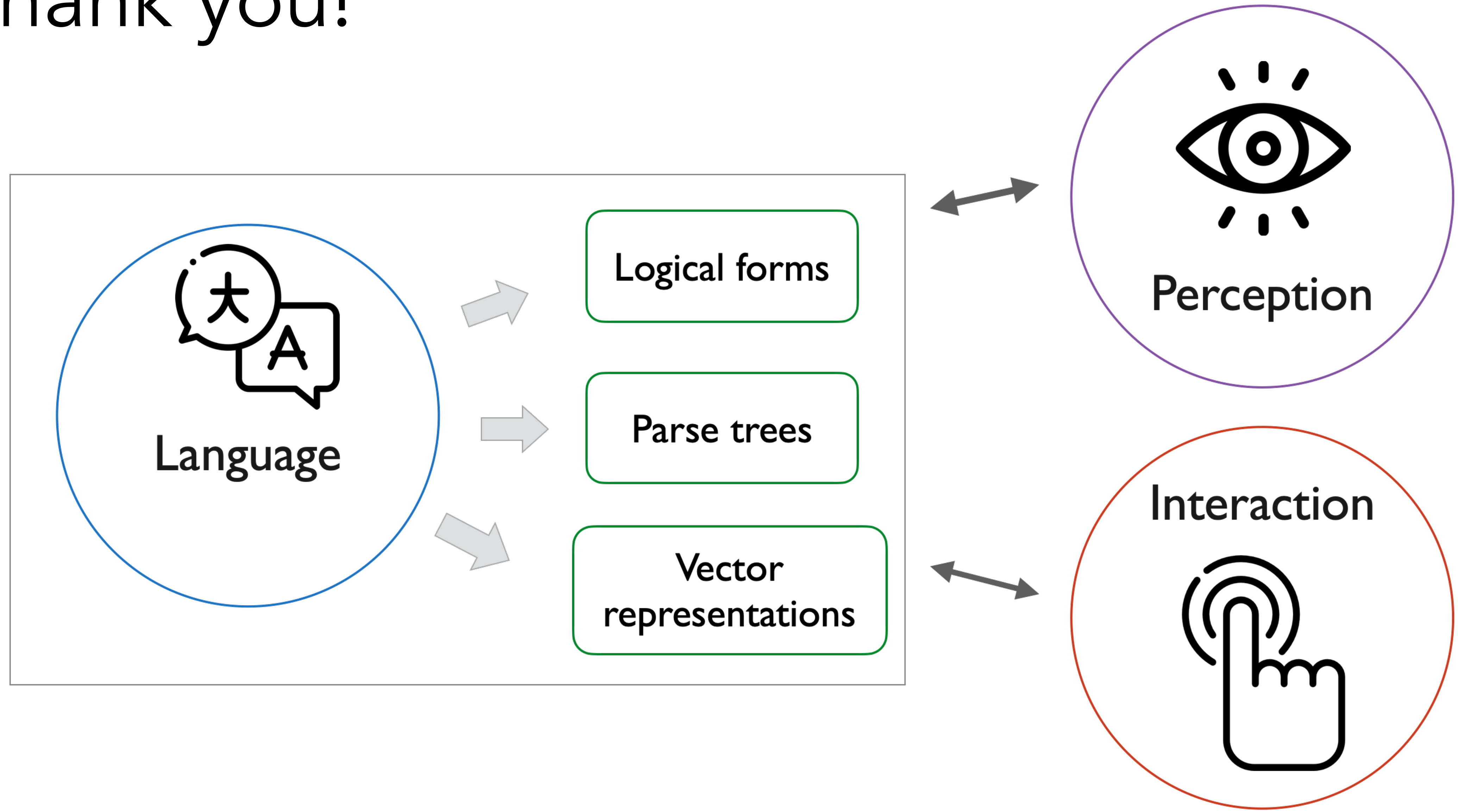


Figure copyright and adapted from Ian Goodfellow, Tutorial on Generative Adversarial Networks, 2017.

- Examples of text-to-image generation with
  - GANs (GAN+CLS+INT, StackGAN++)
  - VAE+Autoregressive (DALL-E – like VQ-VAE but text conditioned)

- Text to 3D is underexplored

Thank you!


Language
Logical forms
Parse trees
Vector representations
Perception
Interaction