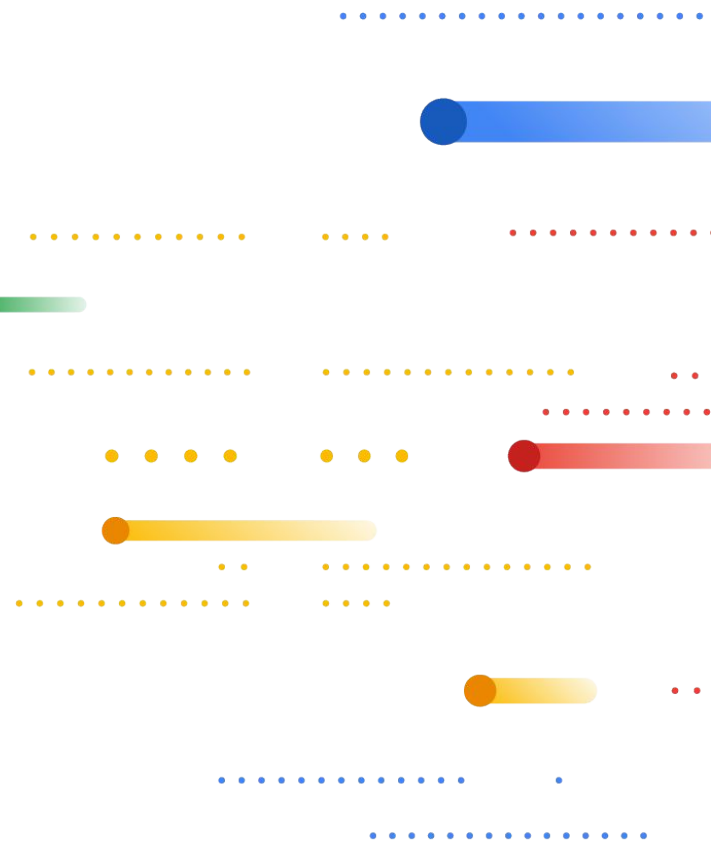


Grounded Language Learning (with neural nets)

Google Brain Toronto, Google AI

Jamie Kiros



This talk

Three parts:

- Grounding and Scope (high level, non-technical)
- Building blocks (current best practices, technical)
- Relationships to other research

We focus on telling a story, highlighting relevant work along the way

- We won't go into much detail on any specific paper
- Inspirations from cognitive science, linguistics and philosophy

I want to leave you with a lot to think about and get inspired!

Disclaimer

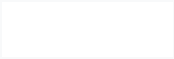
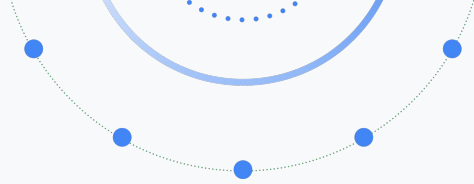
We focus exclusively on neural network models for grounding

- Other areas: semantic parsing, models of language acquisition, etc
- What we discuss is not new with deep learning!

This talk primarily focuses on language-vision integration

- Much of what we discuss generalizes to other modalities
- Vision is not the only way to ground language!

There is a lot more related work we don't cover - use these slides as a starting point and explore more

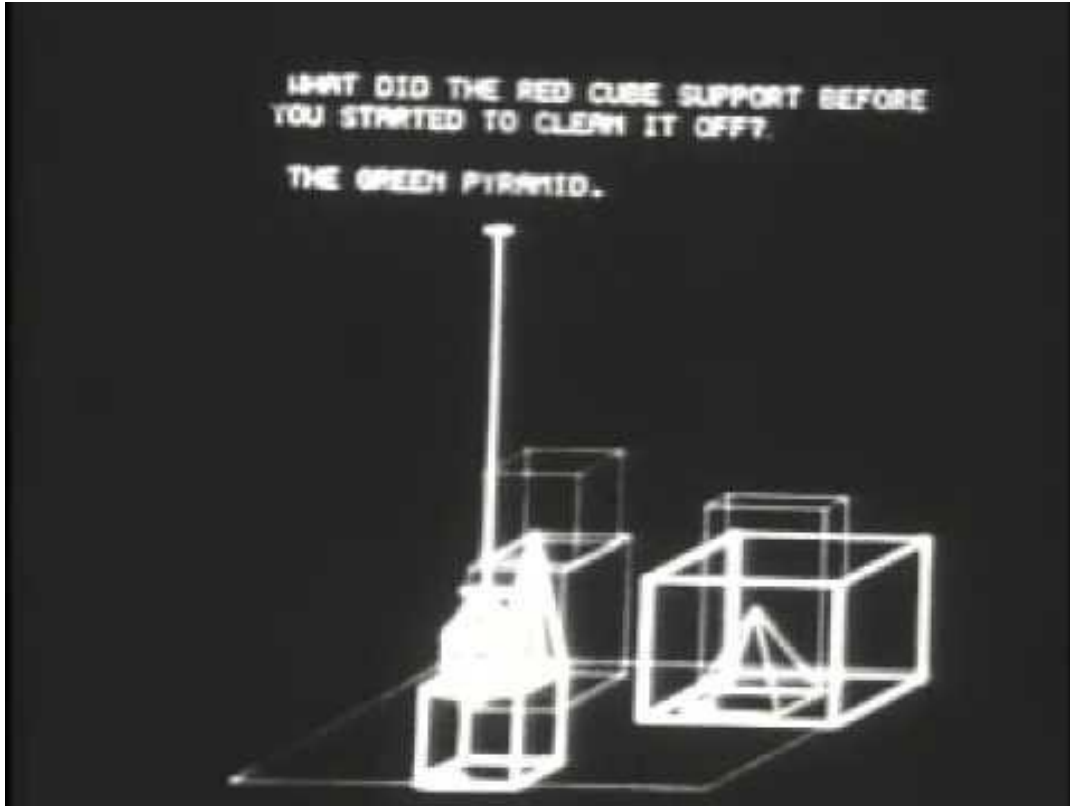


Part I: Grounding and Scope

Two approaches towards grounding natural language

SHRDLU (Winograd, 1972)

Very early attempt at natural language grounding



SHRDLU (Winograd, 1972)

The constructions (verbs, prepositions, question types etc) and vocabulary determined in advance

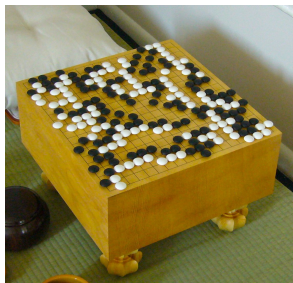
A parser written to map known constructions to 'action plans'

The effect of each 'action plan' on the visual environment programmed

Hard work to add a new linguistic capacity

Agent Scope

Domain of accessible perception to the agent



 AlphaGo

Training scope



Evaluation scope

→ “Intelligent”!

→ Not “Intelligent”!

Why NLP is hard (amongst many other reasons!)

Natural language is inherently tied to our world scope!

Embodiment: Match the perceptual scope at training with evaluation*

We add a lot of additional meaning to language from our own world embodiment!

Ambiguity and Pragmatics:

- “You have a green light”

What effect does agent scope have on pragmatics?

* This is a ML perspective on the definition from psychology “cognition depends on aspects of the agent’s body other than the brain”

Why NLP is hard (amongst many other reasons!)*

Meaning is context dependent

- John loves Mary
- John loves ice cream

Metaphoricity is the rule, not the exception

- Dave pushed the button
- Dave pushed the trainees
- Dave pushed the drugs

Meaning is not *in* language. Language *indicates* meaning!

Is embodiment necessary?

Can we learn everything from text alone? Why or why not?

- Infinite recursion of symbol lookup

If you can, is it efficient?

Weight

From Wikipedia, the free encyclopedia

This page is about the physical concept. In law, commerce, and in colloquial usage weight may also refer to [mass](#). For other uses see [weight \(disambiguation\)](#).

In [science](#) and [engineering](#), the **weight** of an object is related to the amount of [force](#) acting on the object, either due to [gravity](#) or to a reaction force that holds it in place.^{[1][2][3]}

Some standard textbooks^[4] define weight as a [vector](#) quantity, the gravitational force acting on the object. Others^{[5][6]} define weight as a scalar quantity, the magnitude of the gravitational force. Others^[7] define it as the magnitude of the reaction force exerted on a body by mechanisms that keep it in place: the weight is the quantity that is measured by, for example, a spring scale. Thus, in a state of [free fall](#), the weight would be zero. In this sense of weight, terrestrial objects can be weightless: ignoring [air resistance](#), the famous apple falling from the tree, on its way to meet the ground near [Isaac Newton](#), would be weightless.

Brittleness and Leverage

Brittleness: divergence(training scope, evaluation scope)

Leverage: Model's effect on observer from world scope.
More divergence -> more leveraging power



A woman is throwing a frisbee in a park.

Looks really impressive!

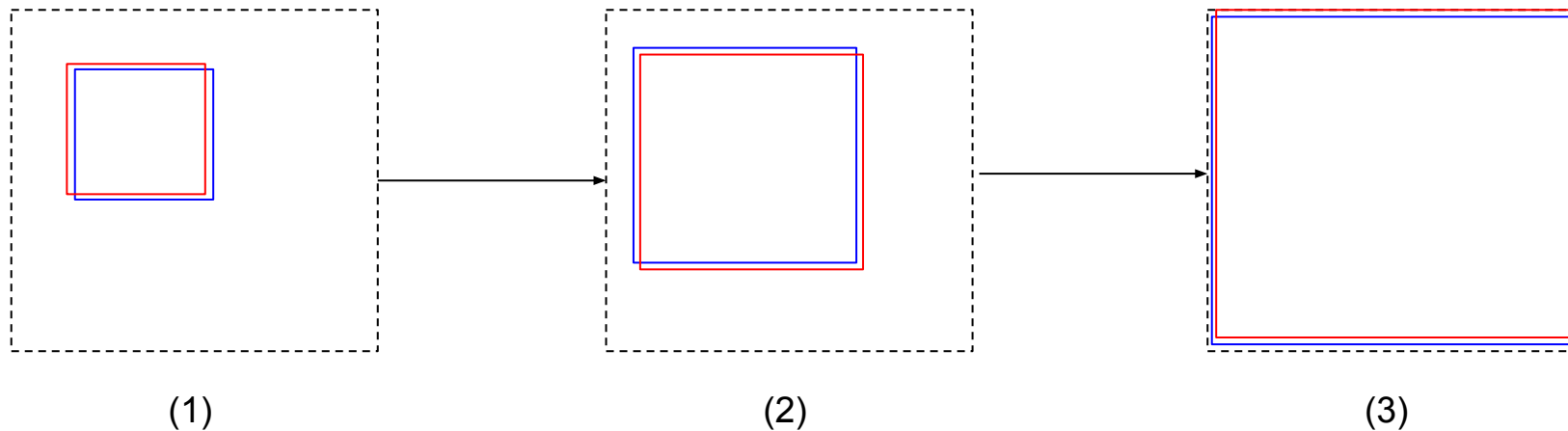


A large white bird standing in a forest.

Looks really silly...

Approach #1: Tie training and evaluation scopes*

Progress over time...



Training
scope

Evaluation
scope

World
scope

*This might be restricted to simulation initially

Recent Vision-Language Environments

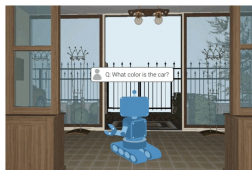
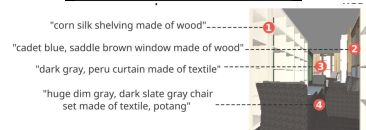
DeepMind Lab (Beattie et al, 2016)

ViZDoom (Kempka et al, 2016)

HoME (Brodeur et al, 2017)

Matterport3D / R2R (Anderson et al, 2018)

House3D (Das et al, 2017; Wu et al, 2018)



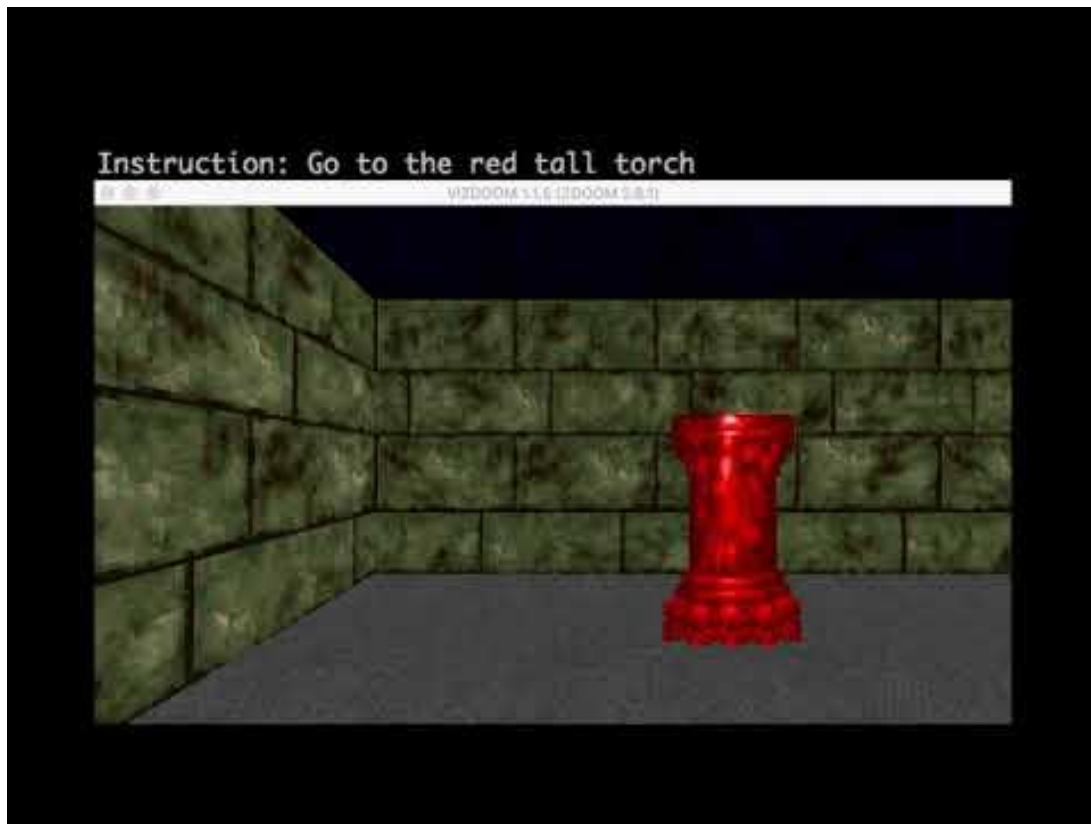
Simulated Language Learning (Hermann et al, 2017)

Agent must navigate an environment to find objects given instructions



ACTRCE (Wu et al, 2018)

Agent is evaluated on unseen test instructions with new attribute-object pairs



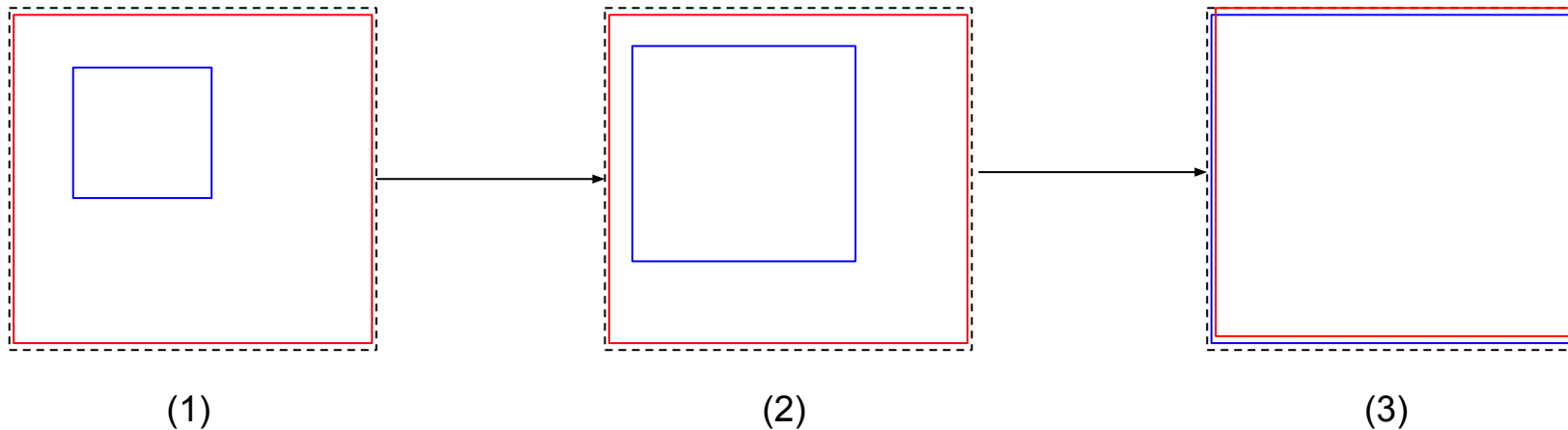
Embodied Question Answering (Das et al, 2018)

Agent must navigate an environment to answer questions



Approach #2: Grow training scope, evaluate in world scope

Progress over time...



(1)

(2)

(3)

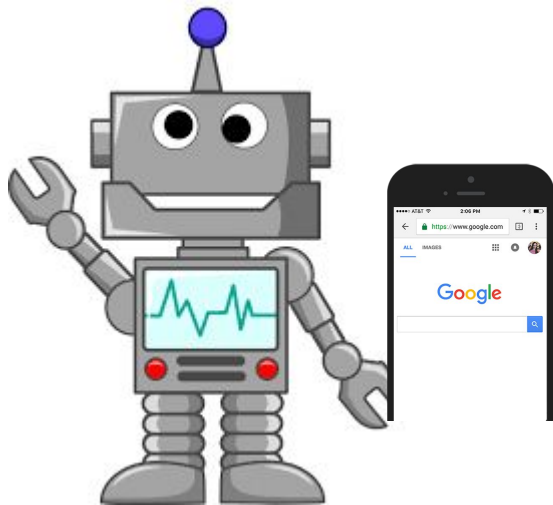
Training
scope

Evaluation
scope

World
scope

Search as Quasi-grounding

What if our agents had access to search?




cat

All Finance Images Videos News More Settings Tools


About 2,740,000,000 results (0.51 seconds)

Videos




CATS will make you LAUGH YOUR HEAD OFF - Funny CAT compilation

Tiger FunnyWorks
YouTube - May 31, 2017




This Little Puppy Is Very Determined To Play, And The Cat Is Not Having It - Digg

Digg - 21 hours ago



Cats are so funny you will die laughing - Funny cat compilation

Tiger Productions
YouTube - Dec 24, 2016



More Images

Cat

Animal

The domestic cat is a small, typically furry, carnivorous mammal. They are often called house cats when kept as indoor pets or simply cats when there is no need to distinguish them from other felids and felines.


[Wikipedia](#)

Lifespan: 2 – 16 years (In the wild)
Gestation period: 58 – 67 days
Family: Felidae
Mass: 3.6 – 4.5 kg (Adult)
Daily sleep: 12 – 16 hours


Did you know: Many cats are lactose intolerant and should not be given cow's milk. [petcha.com](#)

Breeds


View 20+ more




British Shorthair




Persian cat



Siamese cat




Maine Coon




Ragdoll

Top stories




Cat runs onto field during pitch at Reds game

MLB.com
9 hours ago



Taylor Swift Thinks Her Role In 'Cats' Already Started

Elle
20 hours ago



This \$1.5 Million Project Aims to Count All the Cats in Washington, DC

Live Science
21 hours ago

→ More for cat

Feedback

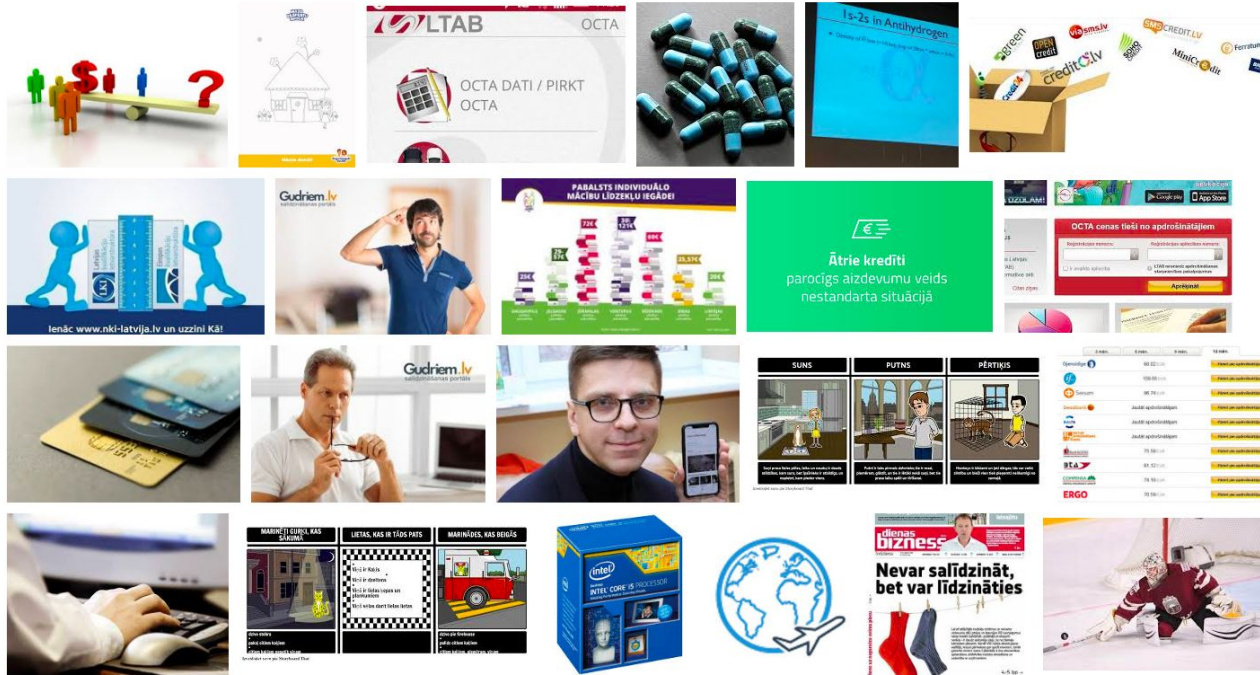
Image Dispersion (Kiela & Hill et al, 2014)

What's the meaning of "čūska"?



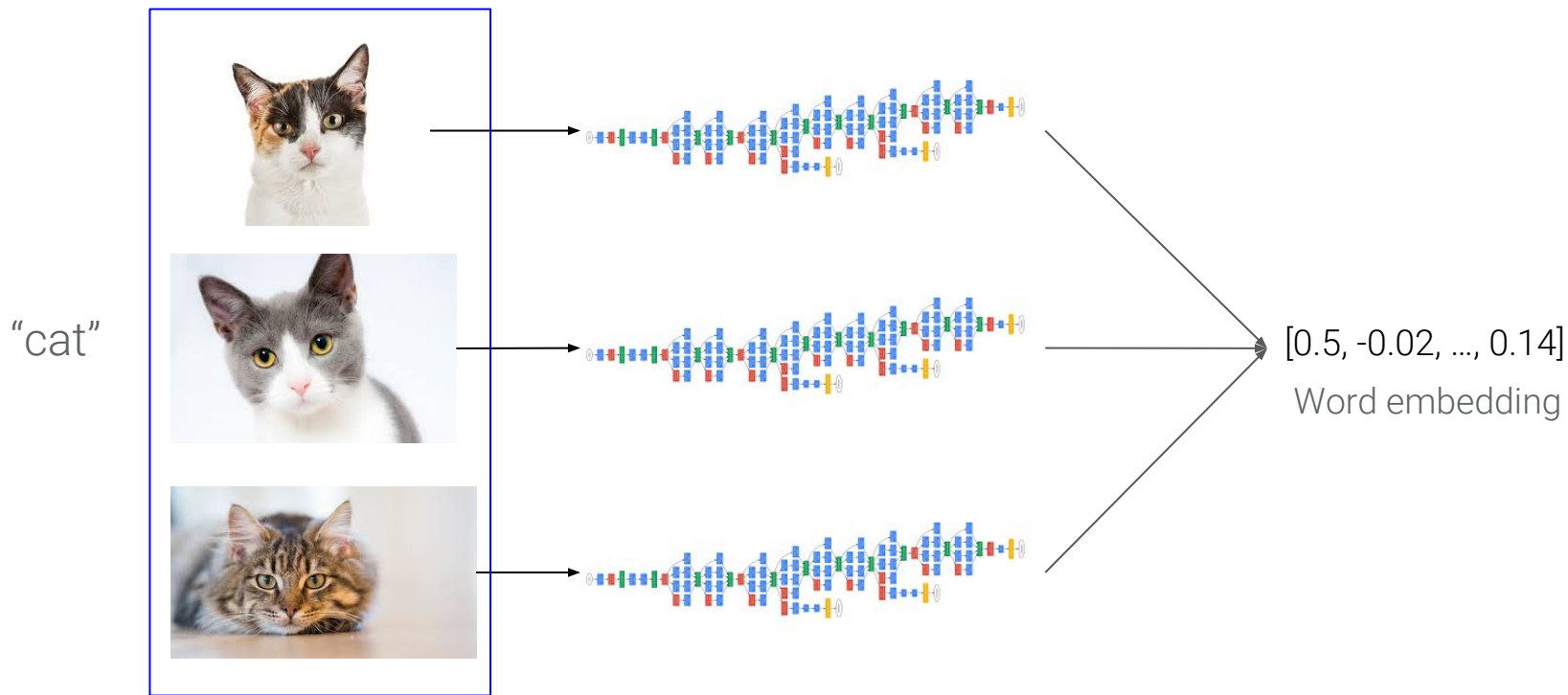
Image Dispersion (Kiela & Hill et al, 2014)

What's the meaning of "salīdzināt"?



Picturebook (Kiros & Chan et al, 2018)*

Grounding language through image search



Google Image Search

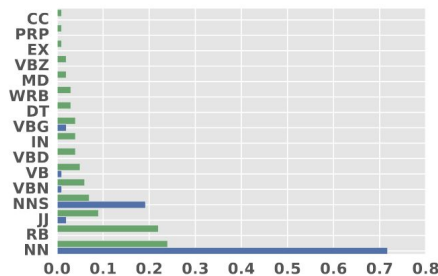
*We are not the first to do this! See paper for related work

Picturebook (Kiros & Chan et al, 2018)

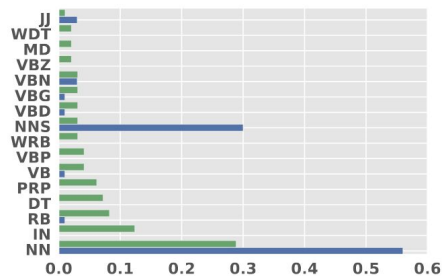
Abstraction vs concreteness of words

Rank	SNLI		MultiNLI		COCO		AG-News		DBpedia		Yelp		Amazon	
	ccorr	disp	ccorr	disp	ccorr	disp	ccorr	disp	ccorr	disp	ccorr	disp	ccorr	disp
top-1%	73	-41	39	-27	53	-22	60	-16	56	-30	47	-28	32	-17
top-10%	54	-39	48	-34	34	-23	52	-24	54	-32	49	-26	50	-30
all	35	-30	30	-27	21	-16	36	-17	39	-30	24	-20	33	-31

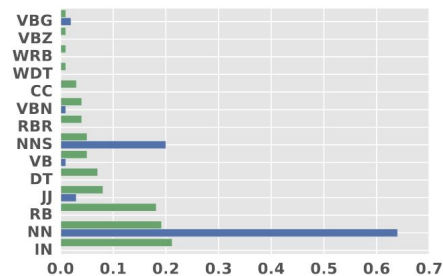
Correlations (x100) to human judgements and image dispersion



(a) SNLI



(b) MultiNLI



(c) AG-News

Part-of-speech analysis: Glove vs Picturebook preference

Navigating through StreetView (Mirowski et al, 2018)

Agent navigates to reach destination through Google StreetView



Playing games via YouTube (Aytar & Pfaff et al, 2018)

Learning to play hard exploration games through expert video



expert sequence



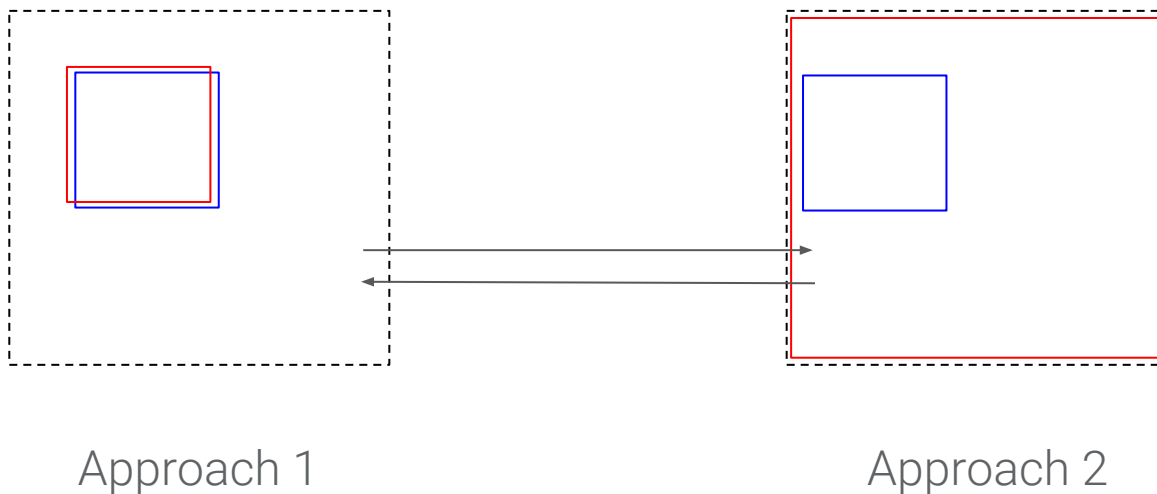
learnt agent

A strategy towards embodiment

Focus research on both approaches

Transfer representations across approaches

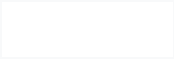
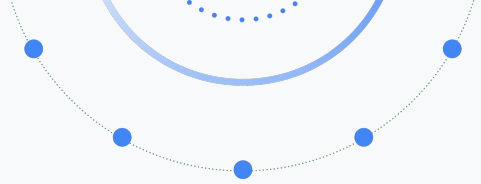
Add complexity and scope overtime



ACTRCE Unseen word generalization (Harris Chan)

Use pre-trained sentence embeddings to generalize beyond training words



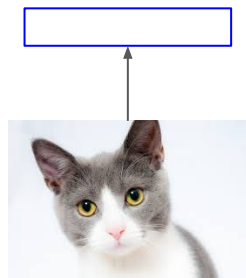


Part II: Building blocks

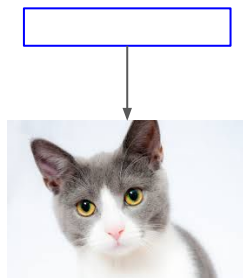
Current approaches for connecting multiple modalities

Five components in multimodal model building

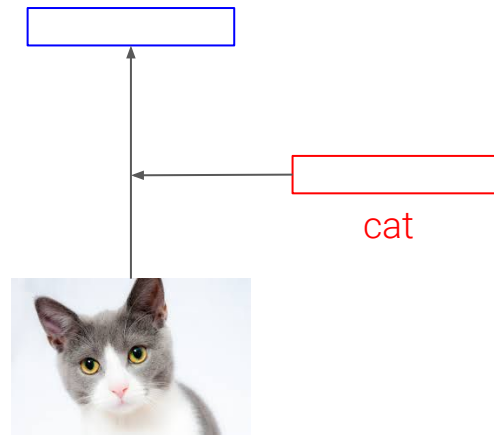
Encoding, Decoding, Interaction, Prediction, Objective



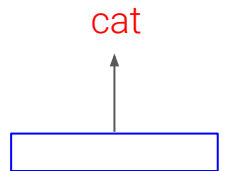
Encoding



Decoding



Interaction



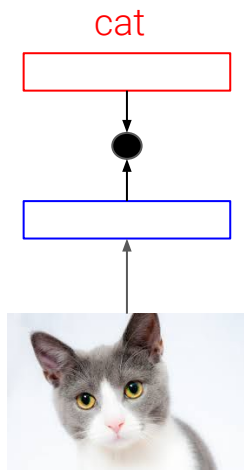
Prediction /
Control

$$L(P, T | X)$$

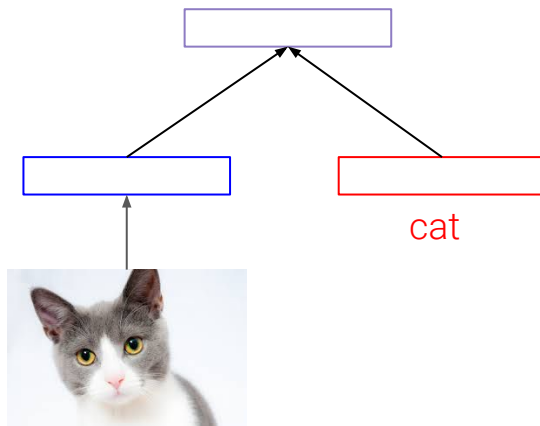
Objective / Loss

Three ways for modalities to interact

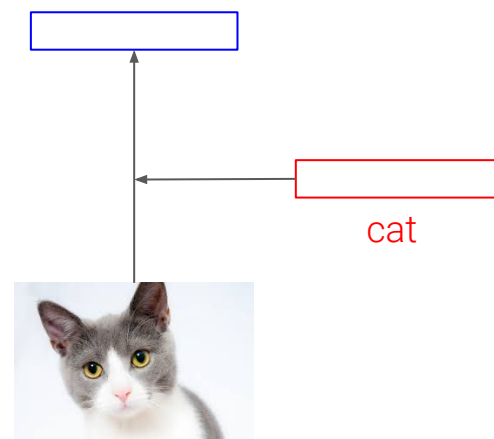
Matching, Fusion and Modulation



Matching
 $\text{score}(X,Y)$



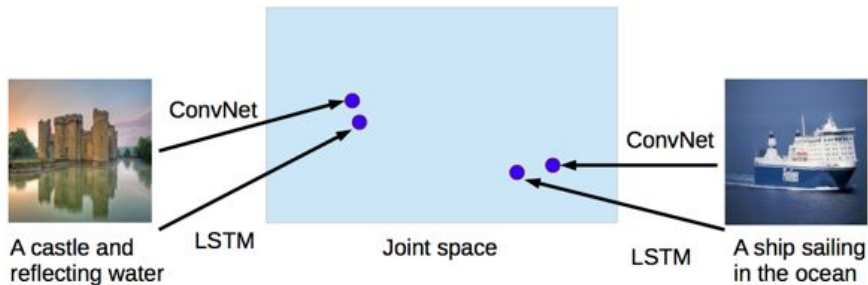
Fusion
 $\text{vec}(X,Y)$



Modulation
 $\text{vec}(X|Y)$

Matching

Learning a joint embedding space for retrieval



Two men and a woman smile at the camera .



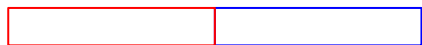
Women participate in a skit onstage .



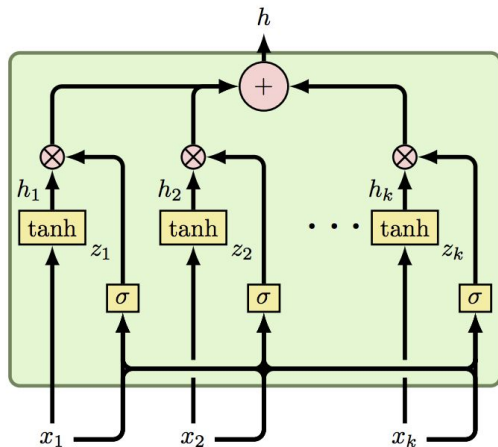
A man is doing tricks on a bicycle on ramps in front of a crowd .

Fusion (E.g: Arevalo et al 2017; Fukui et al 2016)

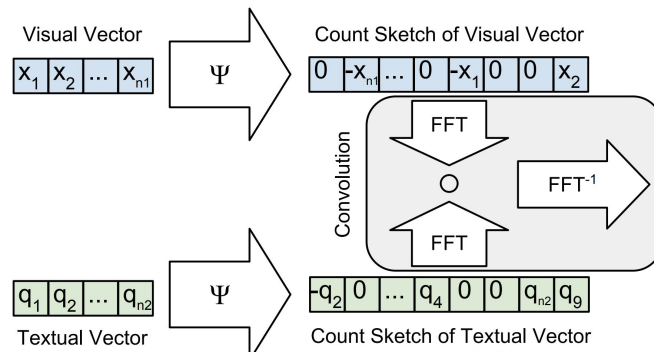
Combining representations of multiple modalities



Simple (add/concat)



Gating-based*



Compact Bilinear Pooling*

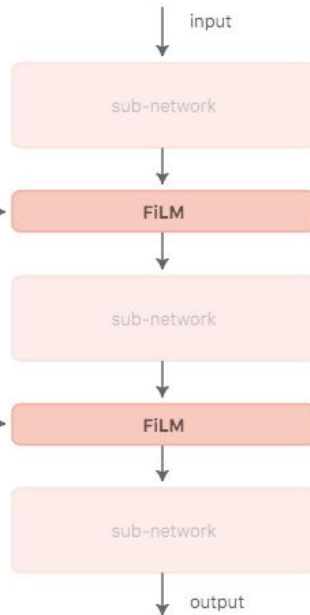
Feature-wise Linear Modulation (Perez et al, 2017)

A general approach to conditioning / modulation

The **FiLM generator** processes the conditioning information and produces parameters that describe how the target network should alter its computation.



Here, the **FiLM-ed network's** computation is conditioned by two FiLM layers.

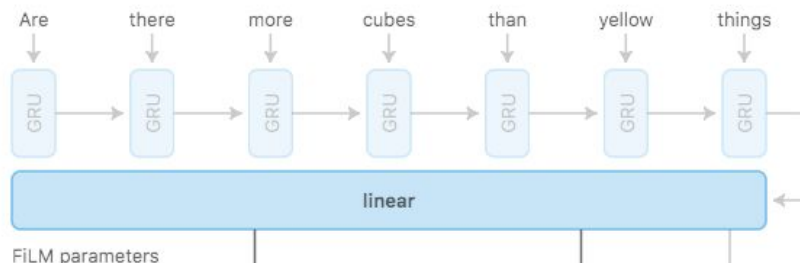


$$\text{FiLM}(\mathbf{x}) = \gamma(\mathbf{z}) \odot \mathbf{x} + \beta(\mathbf{z}).$$

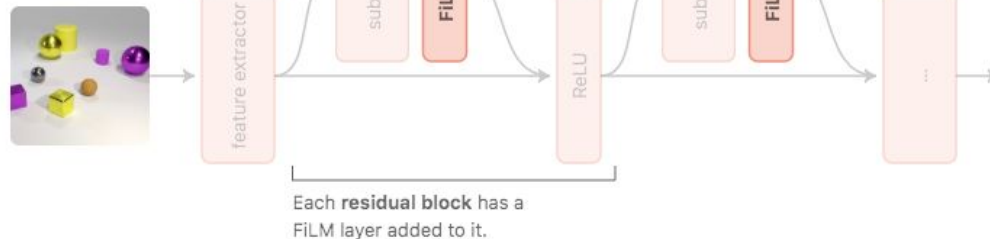
Visual Question-Answering on CLEVR (Perez et al, 2017)

Answering questions about images

The linguistic pipeline acts as the FiLM generator.



FiLM layers in each residual block modulate the visual pipeline.



RL with text-based instructions (Chaplot et al, 2018)

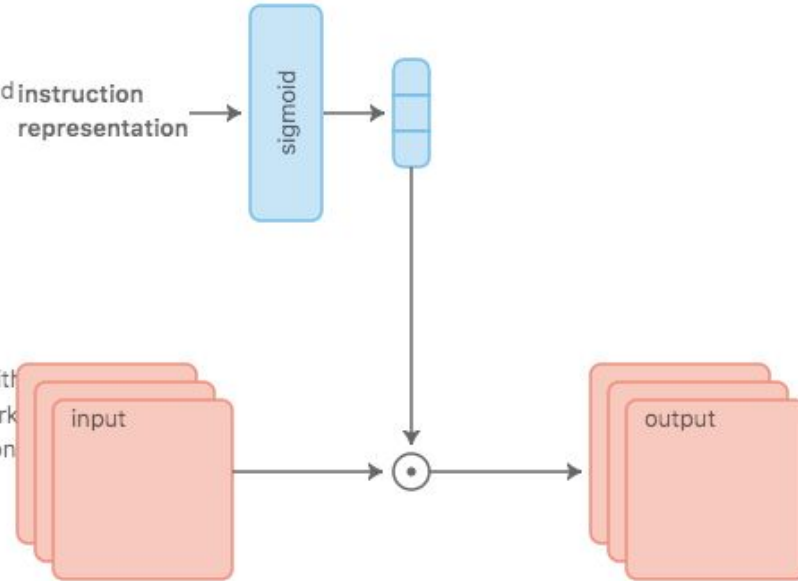
Modulating visual processing with language instructions

Chaplot et al. use **sigmoidal gating**

as a multimodal fusion mechanism.

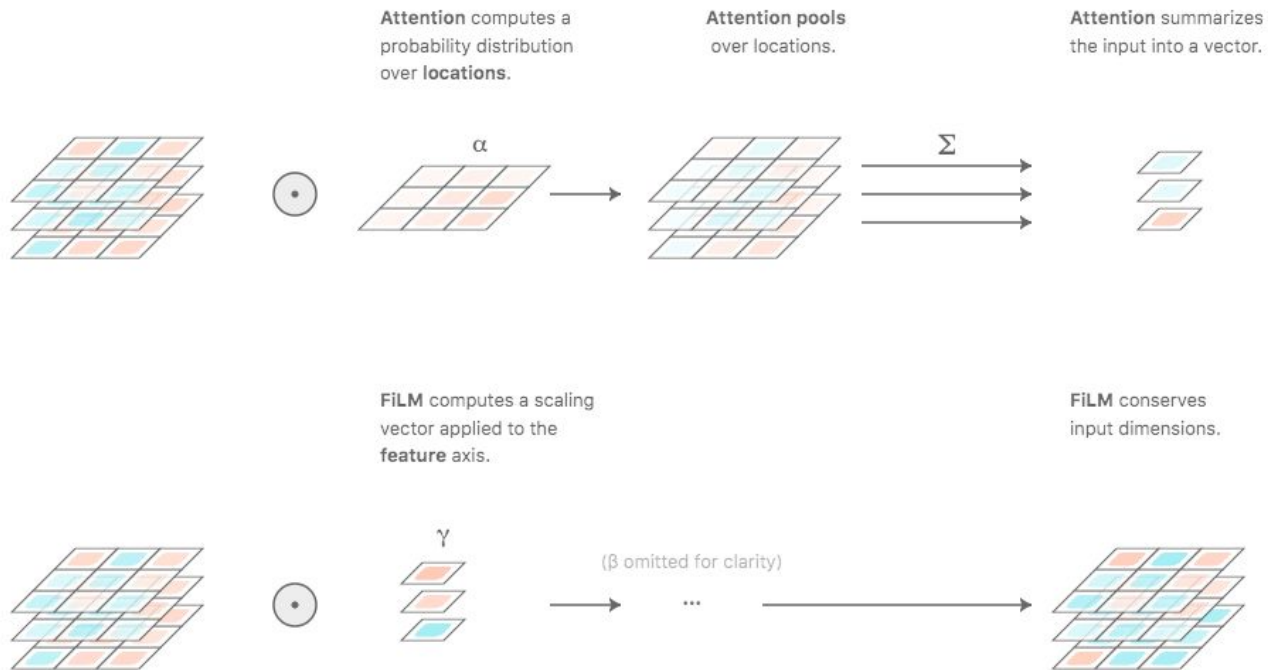
An instruction representation is mapped to a scaling vector via a sigmoid layer.

The scaling vector is then multiplied with the input feature maps. A policy network uses the result to decide the next action.



Feature-wise Linear Modulation and Attention

How FiLM and attention are related:



“Translation” in a general sense

Translating one type of modality into another



To English

“A man that is sitting in front of a suitcase”



To Category

Category 127
(Male Human)

“Last week, Kigali raised the possibility of military retaliation after shells...”

To French

“La semaine dernière, Kigali a soulevé la possibilité de représailles militaires après avoir débarqué des coquilles...”

“Can you give our readers some details on this?”

To German

“Können Sie unseren Lesern einige Details dazu geben?”

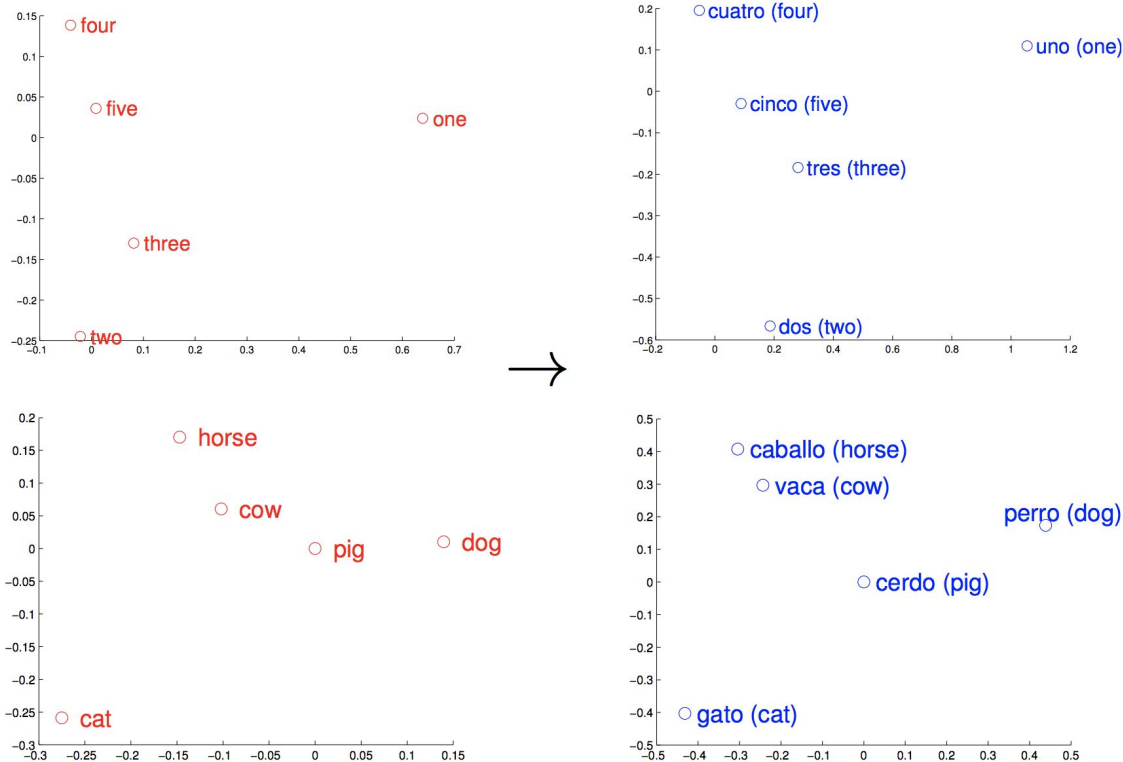
The above represents a triumph of either apathy or civility

To Parse

“S NP DT JJS /NP VP VBZ NP NP DT NN /NP PP IN NP NP NN /NP CC NP NN /NP /NP /PP /NP /VP . /S”

“Translation” as Mappings in Vector Space

Exploiting geometric structure in language embeddings



Summary

Five generic component types:

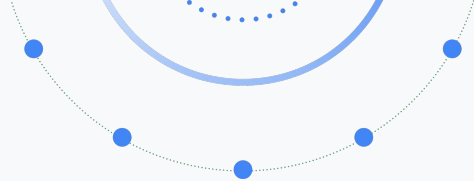
- Encoding, Decoding, Interaction, Prediction/Control, Objective

Three generic types of interaction layers:

- Scoring, Fusion, Modulation

“Translating” between modalities as vector space mappings

These ideas allow for building expressive multimodal models. What’s missing?



Part III: Grounding and Related Research

Contextualization, Multi-Aptness, Specificity, Relevance Realization, Language Generation

Grounding and related research

We will discuss how grounding relates to:

- Contextualization
- Multi-apt representations
- Relevance Realization
- Specificity
- Natural Language Generation and Dialogue

Unifying the above ideas will allow us to discover new research directions!

Think about how to use the building blocks to implement models.

Perspective shifting (Barsalou)

What do the following concepts have in common?

Spouse, Children, Pets, Works of Art, Explosive material, Toxic material

Perspective shifting (Barsalou)

What do the following concepts have in common?

Spouse, Children, Pets, Works of Art, Explosive material, Toxic material

“Things you pay attention to in a fire”

Perspective shifting (Barsalou)

What do the following concepts have in common?

Spouse, Children, Pets, Works of Art, Explosive material, Toxic material

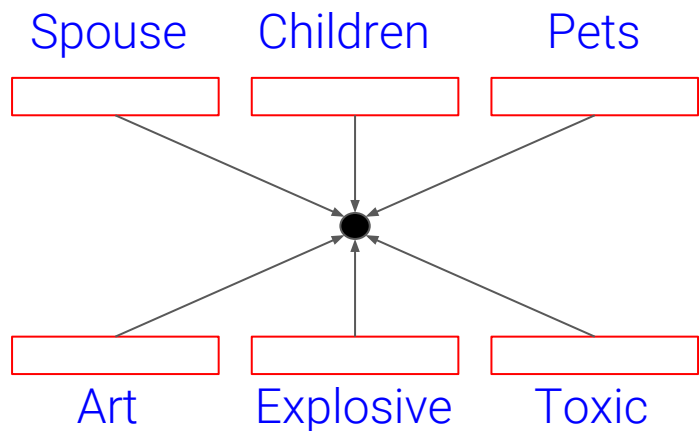
“Things you pay attention to in a fire”

How would you build a program to model this behaviour?

Do the underlying representations need to change with new information?

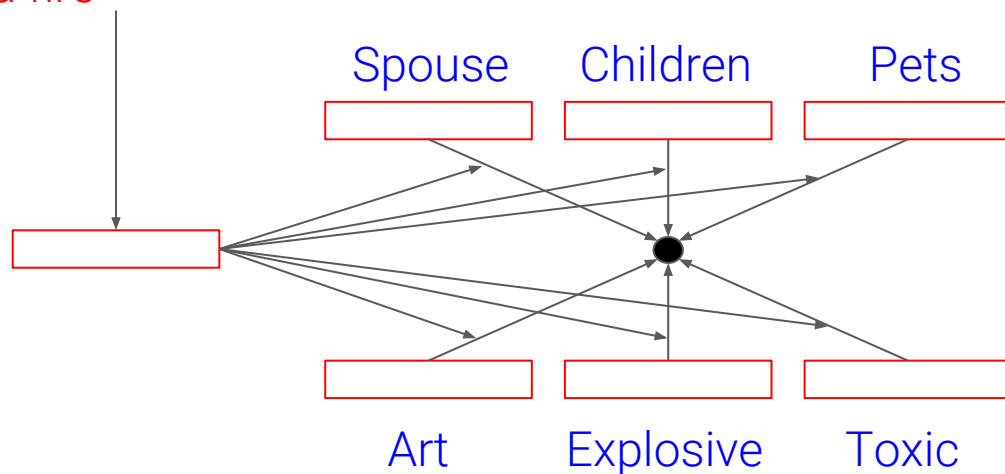
Exercise!

Lets use our building blocks to construct a model



Low score

“Things you pay attention to in a fire”



High score

Contextualization through Cross-Situational Learning

Learning meaning through multiple uncertain exposures



Divritenis atrodas zālē.



Persona, kas brauc ar skrituļdēlu.



Divas ēdoš meitenes, brauca ar divriteni.

Which words relate to “bike”?

Contextualization through Cross-Situational Learning

Learning meaning through multiple uncertain exposures



Divritenis atrodas zālē.



Persona, kas brauc ar skrituļdēlu.



Divas ēdoš meitenes, brauca ar divriteni.

Which words relate to “bike”?

Sense Disambiguation and Grounding

“The crane was so massive it blocked the sun.”

Which sense of “crane” is this referring to?*

Sense Disambiguation and Grounding

“The crane was so massive it blocked the sun.”

Which sense of “crane” is this referring to?*



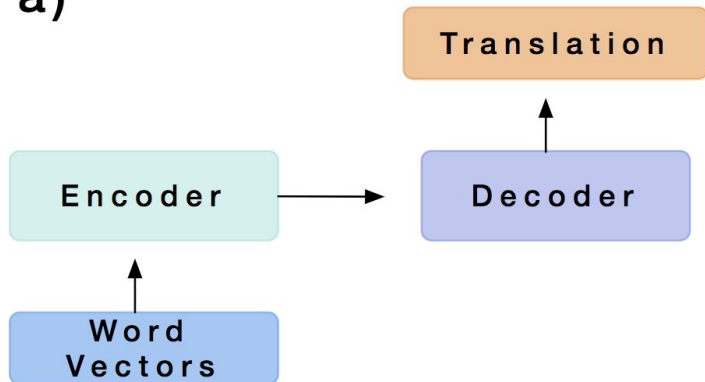
*Example from May et al, 2012.



Contextualized reps. (McCann et al, 2017; Peters et al, 2018; etc)

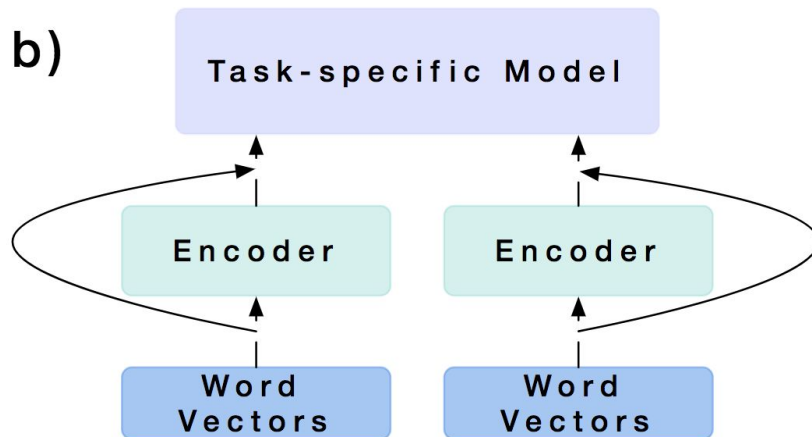
Harness large corpus to learn contextualized word representations

a)



Training

b)



Inference

Winograd Schemas (Winograd, Levesque, et al)

The trophy doesn't fit into the brown suitcase because it's too [small/large].

What is too [small/large]?

The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.

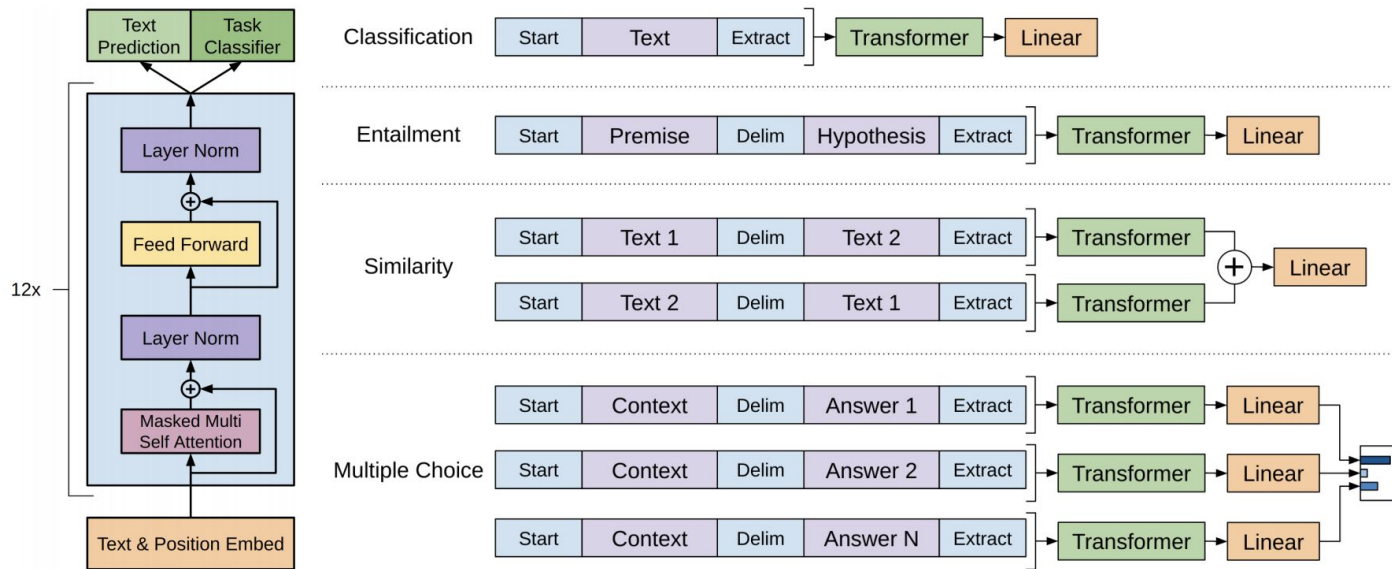
Who [feared/advocated] violence?

The man couldn't lift his son because he was so [weak/heavy].

Who was [weak/heavy]?

Multi-apt representations (E.g: Radford et al, 2018)

Multi-aptness allows for flexible modulation across many contexts / tasks



Other examples: CoVe, ELMo, ULMFit, InferSent, Skip-Thoughts, etc

InferLite (Kiros & Chan, 2018) [to appear soon]

Very fast training and inference of multi-apt (generic) sentence embeddings

- Simplest versions can be trained in < 30 min on 1 GPU. No recurrence

Works as well as InferSent (Conneau et al, 2017) on average

- Obtained much stronger results on semantic similarity tasks

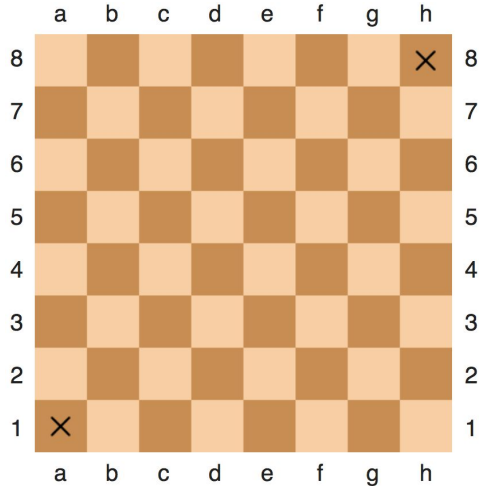
Can learn binary codes for semantic hashing as part of the model

- Combined with fast inference allows for sparse encoding of billions of sentences with minor performance degradation

Relevance Realization (Vervaeke et al, 2012)

Our ability to “zone in” on what’s relevant at any given time

Act in a combinatorially-large search space without “scoring” all states.

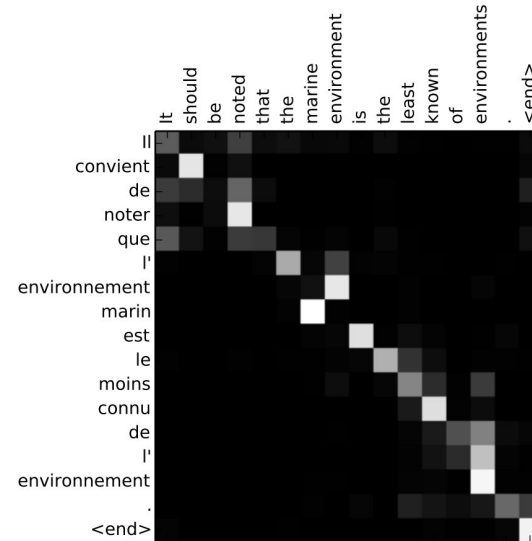
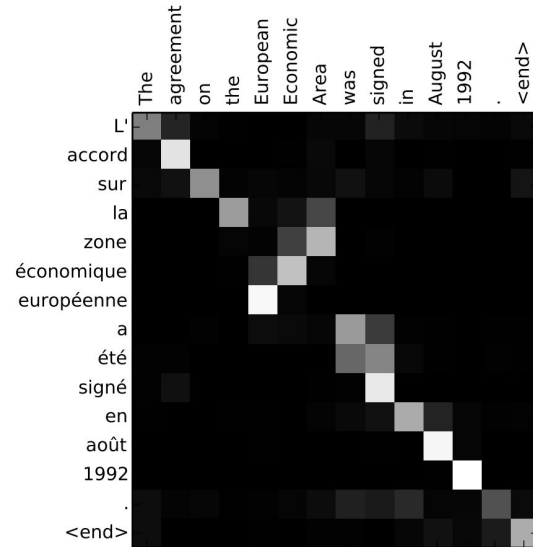


Mutilated Chessboard Problem (Max Black, 1946)

Dynamic Representations \leftrightarrow Modulation \leftrightarrow Relevance

Modulation allows for dynamic representation updates

Update representations based on relevance, through optimization



Visual Question Answering

Answering questions about images and video



(Antol et al, 2015; Tapaswi et al, 2016; etc)

How would you realize relevance in these problems?

Does it appear to be rainy?
Does this person have 20/20 vision?



Q: How does E.T. show his happiness that he is finally returning home?

A: His heart lights up



Q: Why do Joy and Jack get married that first night they meet in Las Vegas?

A: They are both vulnerable and totally drunk



Q: Why does Forrest undertake a three-year marathon?

A: Because he is upset that Jenny left him



Q: How does Patrick start winning Kat over?

A: By getting personal information about her likes and dislikes

Aligning Books and Movies (Kiros & Zhu et al, 2015)

Matching book passages with movie clips

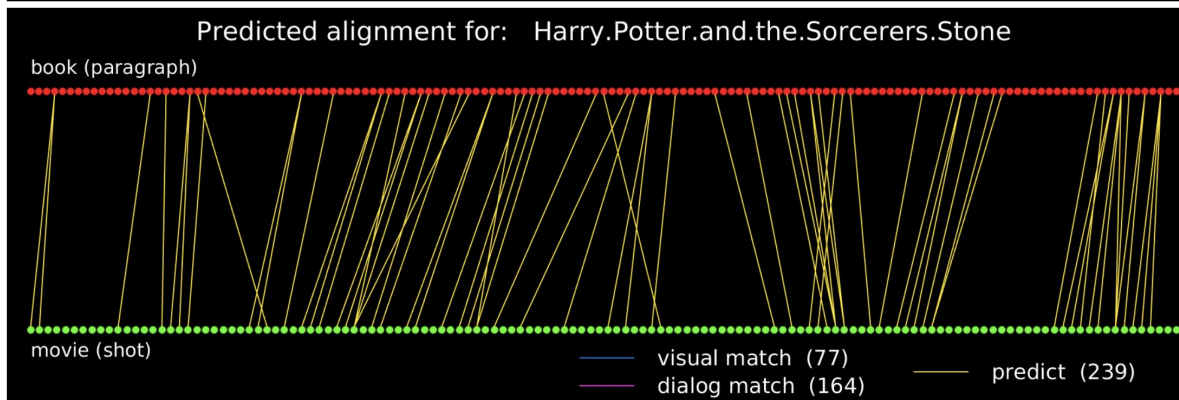
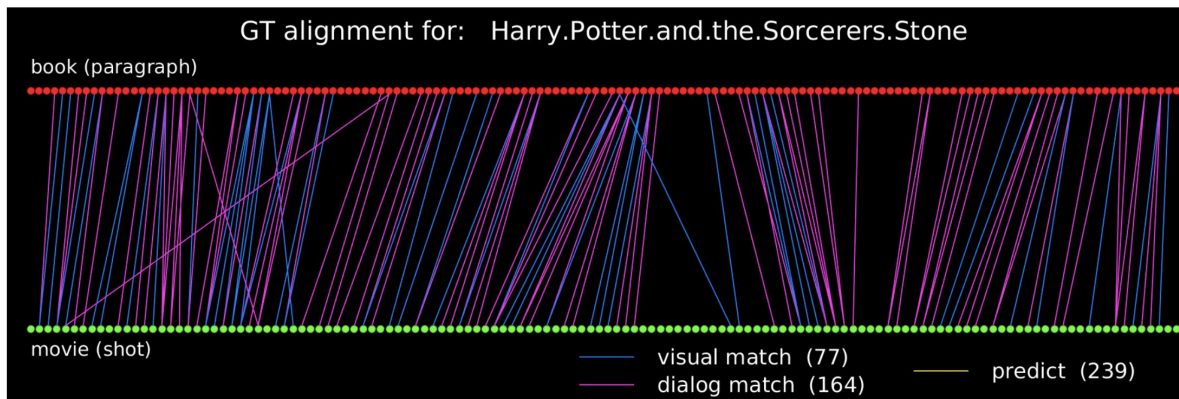


Image Specificity (Jas & Parikh, 2015)

Variance in human generated image descriptions:

Specificity = 0.89



There is a lot of snow on the mountain.
There is a snow covered mountain.
A snow covered mountain.
A mountain with snow.
A snowy mountain.

Specificity = 0.59



Children play racing games in an arcade.
A group of kids playing games.
A few kids playing arcade games.
some kids in an arcade.
Kids are playing racing games.

Specificity = 0.37



A house with a porch.
There is a railing around the porch
of the house.
House with really green grass.
A view of a small white and blue house.
a house shown from outside.

Specificity = 0.11




People waiting at an airport.
The interior of a building with a sloped roof.
the inside of airport.
A decadent room with people walking around.
A large bowling rink.



Tasks ordered by specificity of the conditionals

High to low specificity:

- Autoencoding
- Reversing or sorting sequences
- Machine Translation
- Image captioning
- Dialogue
- Writing prompts / story completion
- Unconditional language generation



More need for
human evaluation

Key idea: Increase specification through grounding (E.g. Dialogue)

High specificity tasks are amenable to automated metrics.

GuessWhat?! (De Vries et al, 2017)

Finding object references through grounded dialogue



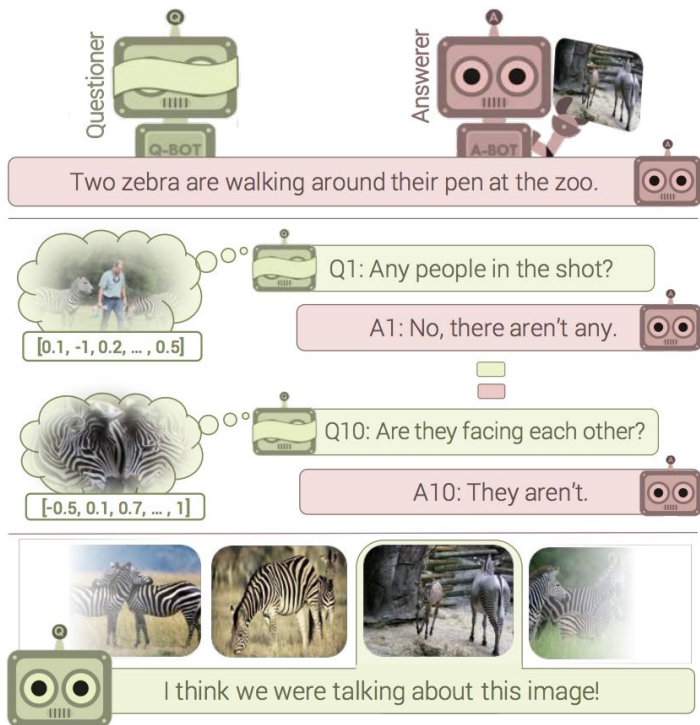
- Is it a person? *No*
- Is it an item being worn or held? *Yes*
- Is it a snowboard? *Yes*
- Is it the red one? *No*
- Is it the one being held by the person in blue? *Yes*



- Is it a cow? *Yes*
- Is it the big cow in the middle? *No*
- Is the cow on the left? *No*
- On the right ? *Yes*
- First cow near us? *Yes*

Cooperative Visual Dialog Agents (Das et al, 2017)

Image guessing game between two agents

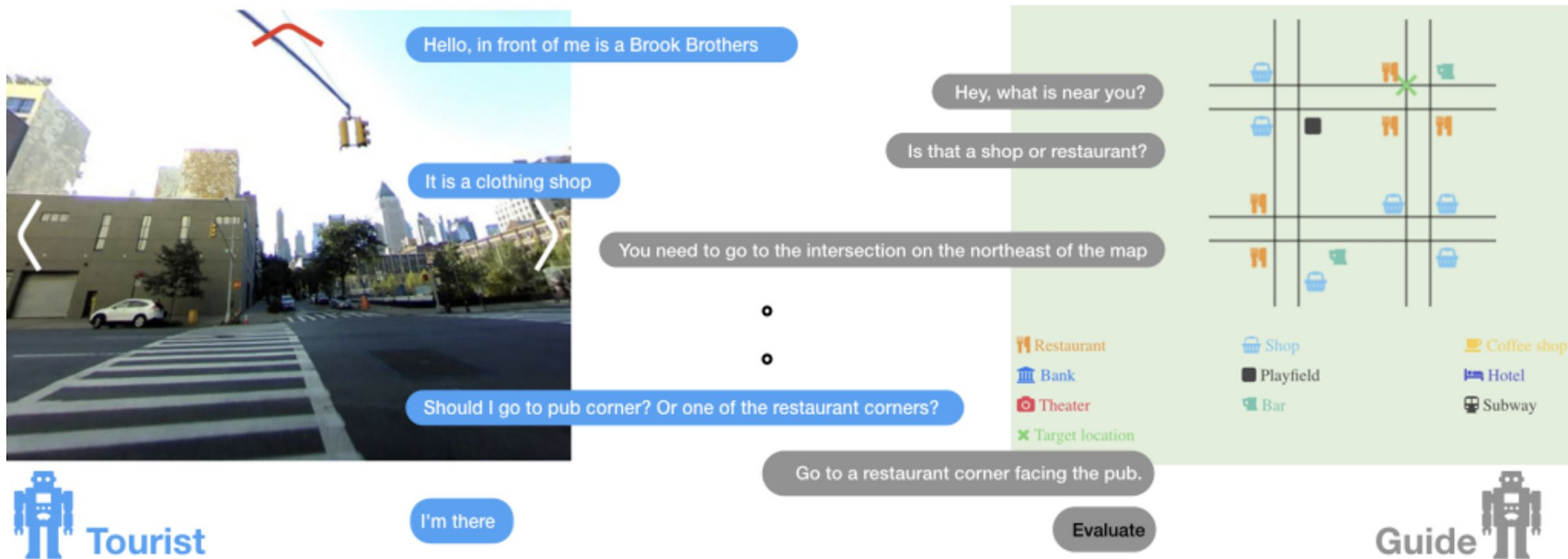


RL-full-QAf

Are people male or female or both?	I think 1 is male, can't see other 1 ...
What are they wearing?	They are all wearing snow pants and jackets
Is it snowing?	It does not appear to be snowing at moment
Can you see any trees?	Yes, there are lot of trees in background
Can you see any buildings in background?	No, I can not
Can you see any trees?	Yes, there are lot of trees in background
Can you see any buildings in background?	No, I do not see any buildings
Can you see any trees?	Yes, there are lot of trees in background
Can you see any buildings in background?	No, I do not see any buildings
Does it look like they are in park?	It does not appear to be

Talk the Walk (De Vries et al, 2018)

Navigating New York City through Grounded Dialogue



Capacity vs Specificity in natural language generation

An oracle allows you to generate perfect text with a copy mechanism

- High Specificity, low model capacity requirement

Unconditional models require large capacity to generate “reasonable” text

- Low specificity, high model capacity requirement

A mechanism for relevance realization allows for dynamic specificity adjustment

Visual Storytelling (Huang et al, 2016)

Ground storytelling through images: increased specificity



+*Viterbi* This is a picture of a family. This is a picture of a cake. This is a picture of a dog. This is a picture of a beach. This is a picture of a beach.

+*Greedy* The family gathered together for a meal. The food was delicious. The dog was excited to be there. The dog was enjoying the water. The dog was happy to be in the water.

-*Dups* The family gathered together for a meal. The food was delicious. The dog was excited to be there. The kids were playing in the water. The boat was a little too much to drink.

+*Grounded* The family got together for a cookout. They had a lot of delicious food. The dog was happy to be there. They had a great time on the beach. They even had a swim in the water.

Table 5: Example stories generated by baselines.

Hierarchical Neural Story Generation (Fan et al, 2018)

Generate a story given a writing prompt

Example Prompt 2: The scientists have discovered something terrible .

The scientist stood there, a little dazed as he stared.

“What is it?” He asked.

“This...this...Thing...This is a virus. A chemical that can destroy entire planet and it is a very small, complex, chemical that could destroy any planet.” The scientist replied. His lab assistant looked down at the tablet.

“I’ve just discovered it. I can’t believe it. It looks like it’s made of some sort of chemical that’s very dangerous.”

“ Well, there’s a virus on the ground. It’s very effective...” “ I can’t believe what it is.” he said, pointing to the scientist .

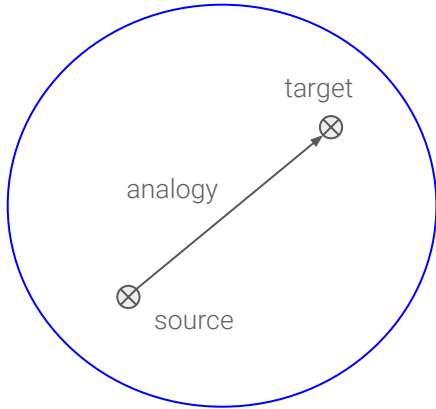
“ We don’t know what this thing is. We haven’t seen anything like it . We can’t even see anything like this. ” Dr. Jones stared at the scientist for a moment.

“What do you mean what does it do ?”

“It...It ’s a monster.”

Generation as Inference, Inference as Analogy Making

Infer a vector that when decoded, represents the intended text.
Analogy making as a form of inference.



Generated story about image
Model: Romantic Novels

"He was a shirtless man in the back of his mind, and I let out a curse as he leaned over to kiss me on the shoulder."

He wanted to strangle me, considering the beautiful boy I'd become wearing his boxers."

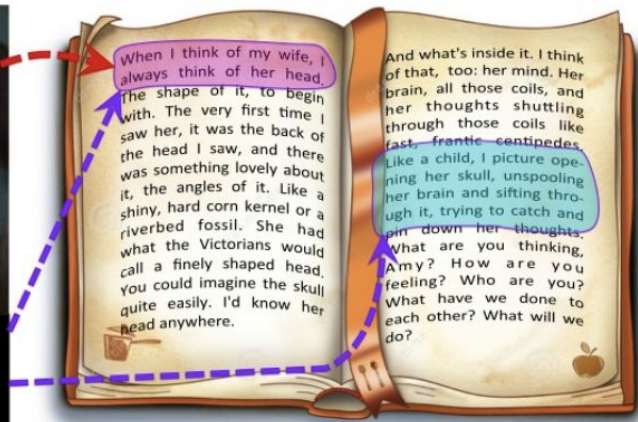
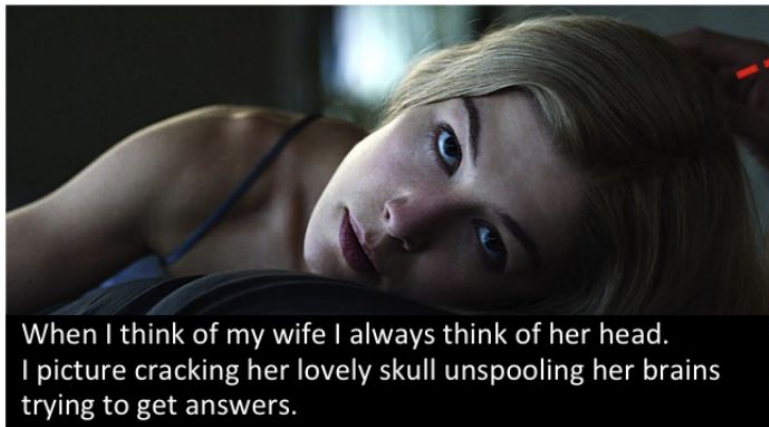
Putting all the pieces together

- There are seemingly infinite notions of similarity and contexts in language
- Grounding allows for embodied contexts
- Representations that allow for easy context modulation are multi-apt
- Relevance realization: our ability to “zone in” on relevant information
- RR with multi-apt representations allow for flexible context encoding
- Grounding and context fixing can increase specificity in tasks like dialogue
- Realizing relevance -> increased specificity -> more controllable text gen

- Overall: we would like methods that can harness contexts, realize relevance and dynamically adapt its representations over time

A “moonshot” project

Generating books conditioned on a movie / metadata / plot / dialogue etc



All of the ideas we discussed would be required!

(We tried this in ~2015 with Sanja Fidler and it was a disaster)



Thanks!

Special thanks to:

Astrid Berg
Erin Grant (Berkeley)
Felix Hill (DeepMind)
William Chan (Google Brain)
Geoff Hinton (Google Brain)

for discussion and feedback on this talk!