Multimodal Human-inspired Language Learning

Graham Neubig
(Work w/Alexander Hauptmann, Yonatan Bisk,
Xinyu Wang, Po Yao Huang, Liangke Gui, Hao Zhu,
Junxian He, Juncheng Billy Li, Paul Michel, Rosaline Su)

Carnegie Mellon University
Language Technologies Institute

Scene

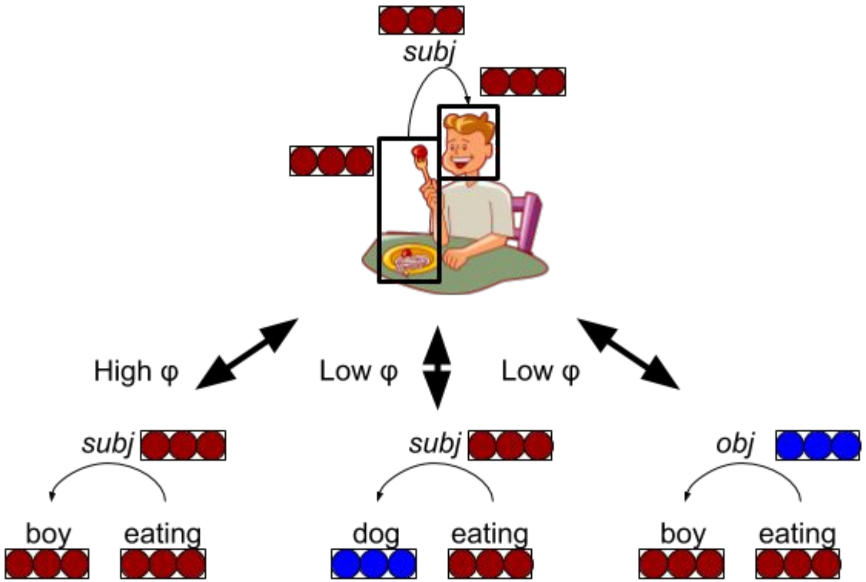
big boy eating pasta

Learner

Courtesy of Sébastien GARNIER: https://www.flickr.com/photos/sebcreation/1428973477/

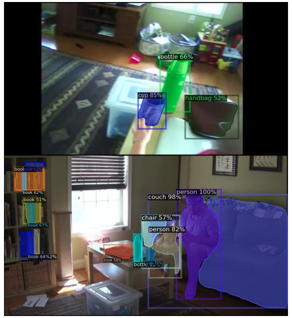
Approach

- Align structure from the visual and verbal domains for better underlying language understanding



Workflow

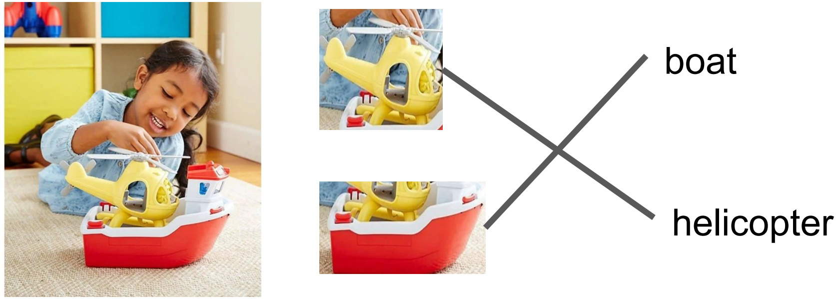
1. Visual concept recognition
   (assume some pre-linguistic ability)



CLIPPING

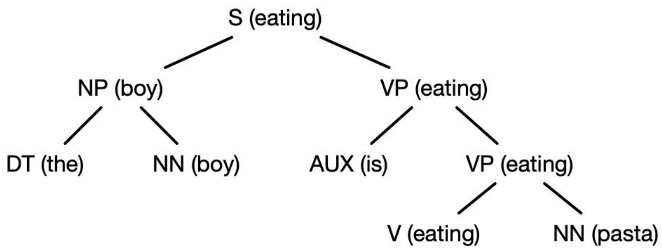
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

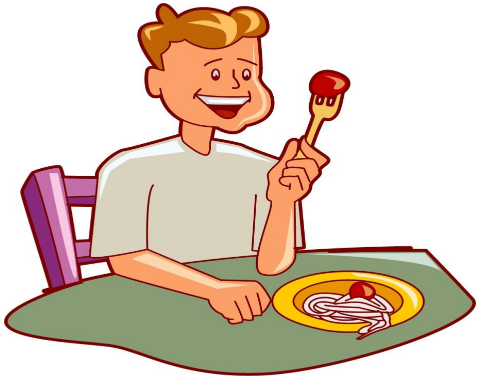
2. Visual-verbal grounding of atomic units



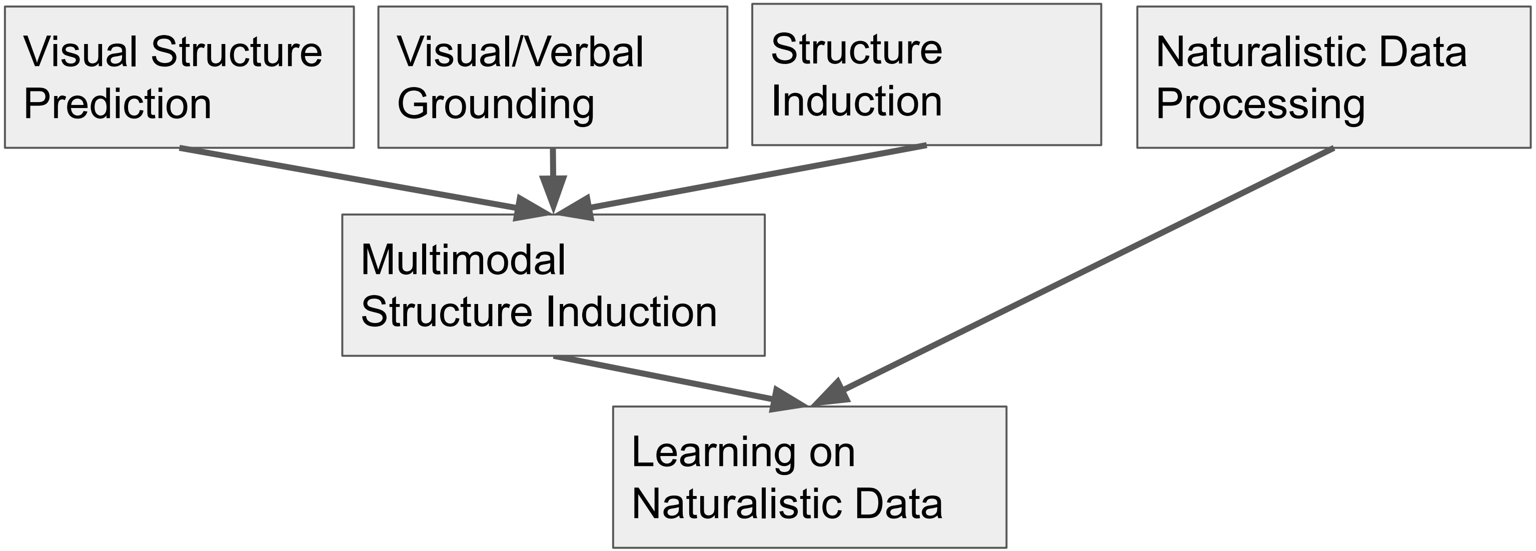
A yellow helicopter on a red rescue boat

3. Linguistic structure induction with multimodal constraints

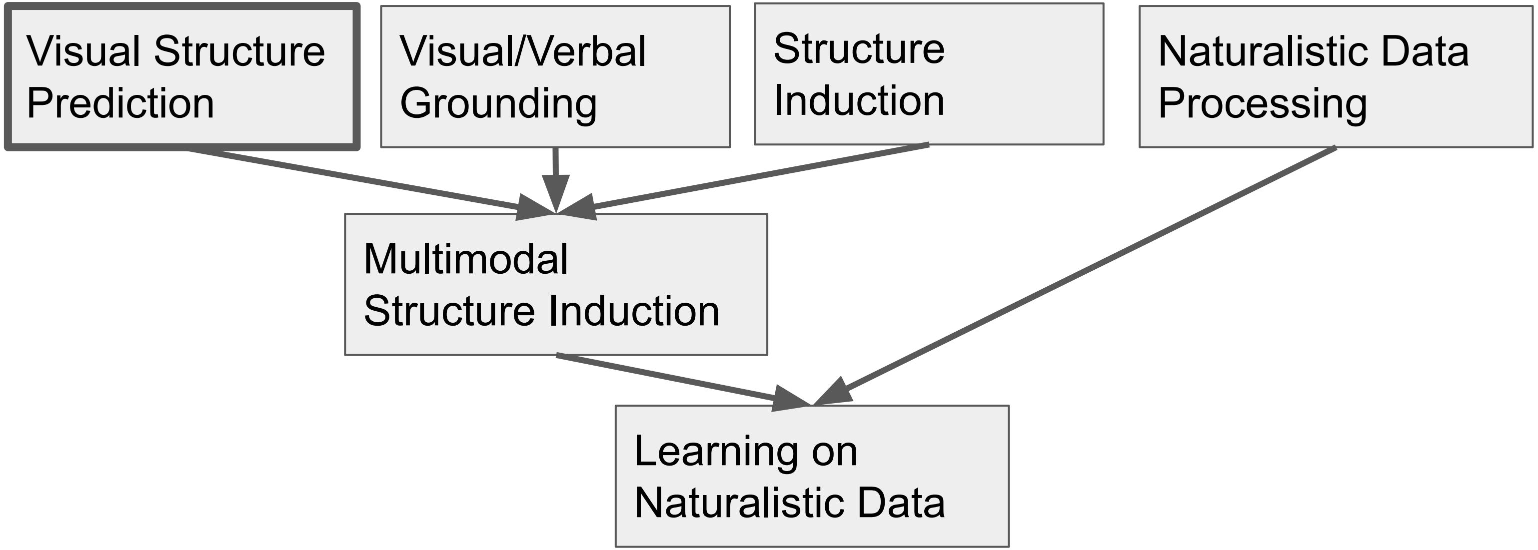




Underlying Components



Visual Structure Prediction



Visual Semantic Frames: ImSitu Dataset (http://imsitu.org/)

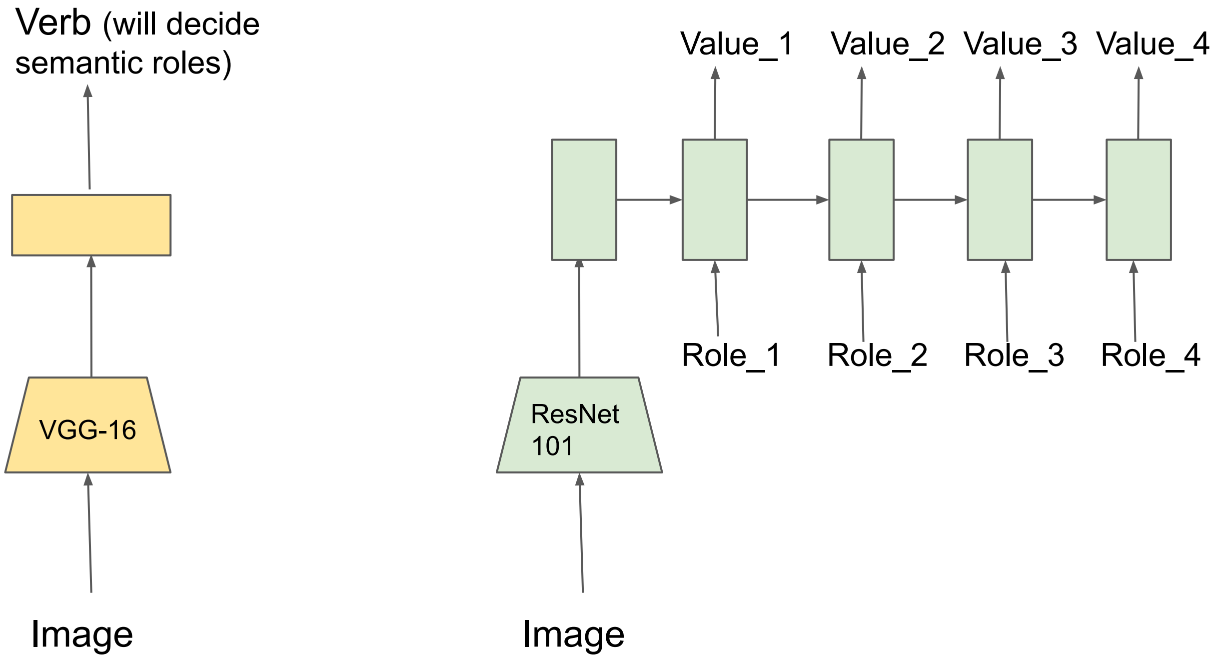
CLIPPING			
ROLE	**VALUE**	**ROLE**	**VALUE**
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

The ImSitu dataset contains annotations of 1) the main activity (e.g. clipping) 2) the participating objects and the roles they play (e.g. the man is clipping the sheep)

It contains over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations

.

Recognition Model [Yatskar+ 2017]



Example predictions of unseen images

Top-3 predicted frames, including verbs and predicted role values

Verb	**Place**	**Agent**		
skiing	Ski slope	skier		
Verb	**Place**	**Agent**		
ascending	mountain	person		
Verb	**Source**	**Place**	**Tool**	**Agent**
descending	mountain	outdoors	ski	person

Example predictions on unseen images

Top-3 predicted frames, including verbs and predicted role values

Verb	**item**	**destination**	**place**	**agent**	
stuffing	food	mouth	room	agent	
Verb	**food**	**container**	**tool**	**place**	**agent**
eating	sandwich	NULL	hand	inside	man
Verb	**container**	**theme**	**place**	**agent**	
cramming	mouth	food	NULL	man	

Incorrect predictions also often make sense

top-1 predicted frame:

Verb: buttering

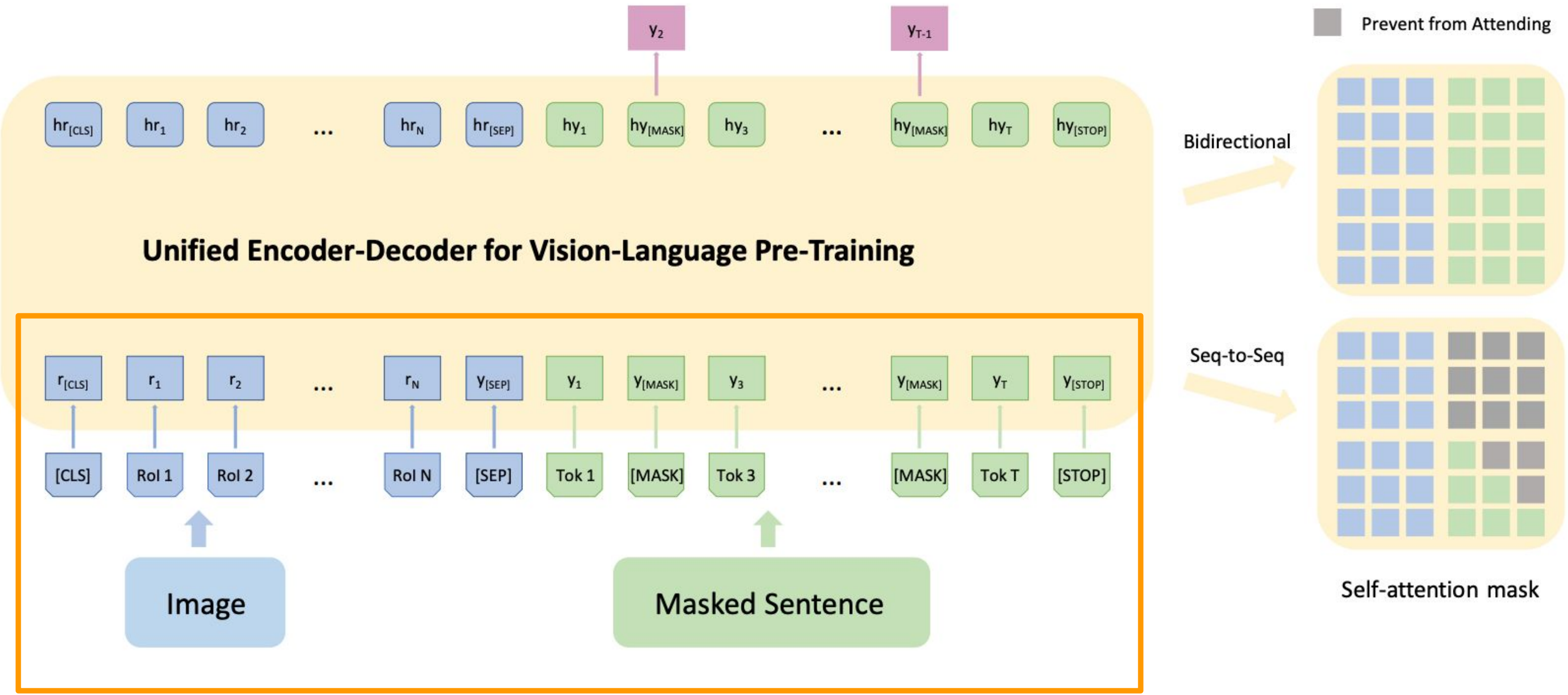
Item: bread

Tool: knife

Place: kitchen

Agent: girl

Proposed Model: Vision-language Pre-training



Recognition Model using pre-trained VLP

Pre-training on Conceptual Captions

~3M image-text pairs

Bidirectional & Seq2seq Objectives

Fine-tuning on imSitu training set

~75k images and labels

Seq2seq objective only

Transform semantic labels into text

Decoding on imSitu dev/test set

~25k images respectively

Continuously predicting [MASK] token to generate sentence

Note: During inference time, the image regions are first encoded along with [CLS] and [SEP] token. Then the model is fed in a [MASK] token and predict what it is. After prediction, another [MASK] token is appended and the process is repeated until [STOP] is chosen.

Visual Semantic Frame Prediction Results

Quantitative

	Dev Set (verb accuracy)	Test Set (verb accuracy)
Baseline	32.2%	32.3%
Fine-tuned VLP	36.9%	36.8%

Qualitative

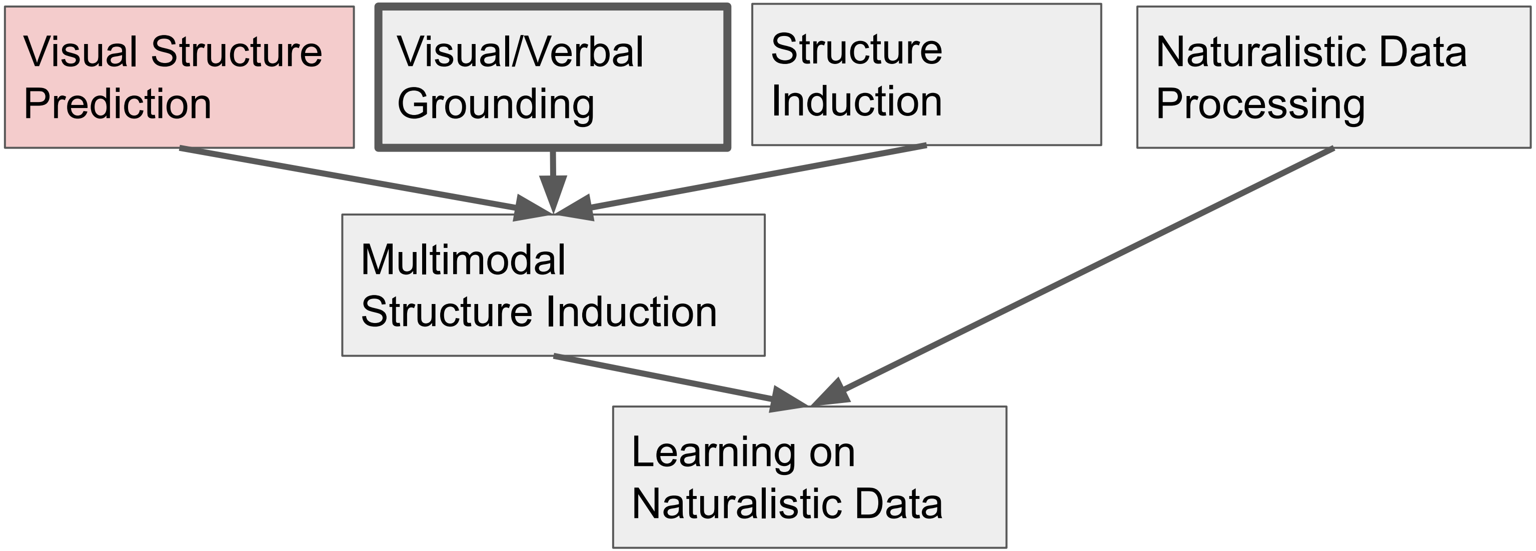
Generated: the verb is flapping . bird flapped its wing at outdoors .

Ground truth:
Verb: flapping
Agent: bird
Bodypart: wing
Place: outdoors

Generated: the verb is marching . soldier marches at street .

Ground truth:
Verb: parading
Agent: soldier
Place: street

Visual/Verbal Grounding

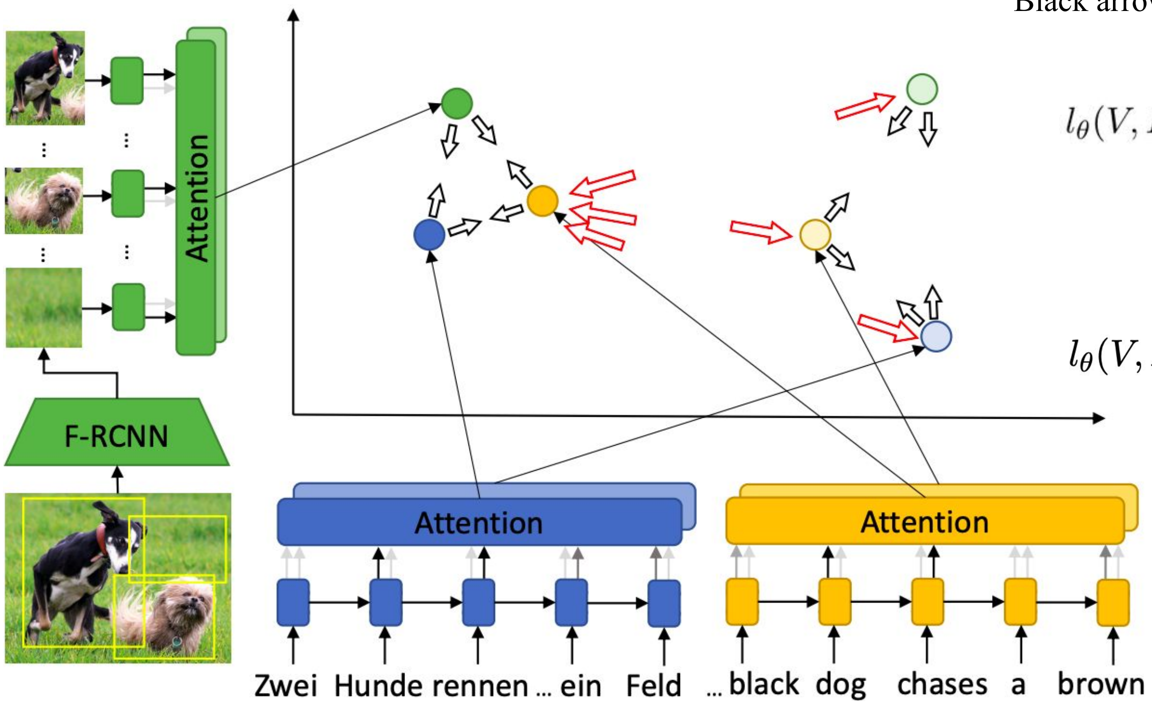


Grounded Multimodal Learning

Children learn in a multimodal environment. We investigate human-like learning in the following perspectives:

- Association of new information to previous (past) knowledge
- Generalization of learned knowledge to unseen (future) concepts
    - Zero-shot compositionality of the learned concepts
        - Blue + Dog -> Blue dog !?

Model



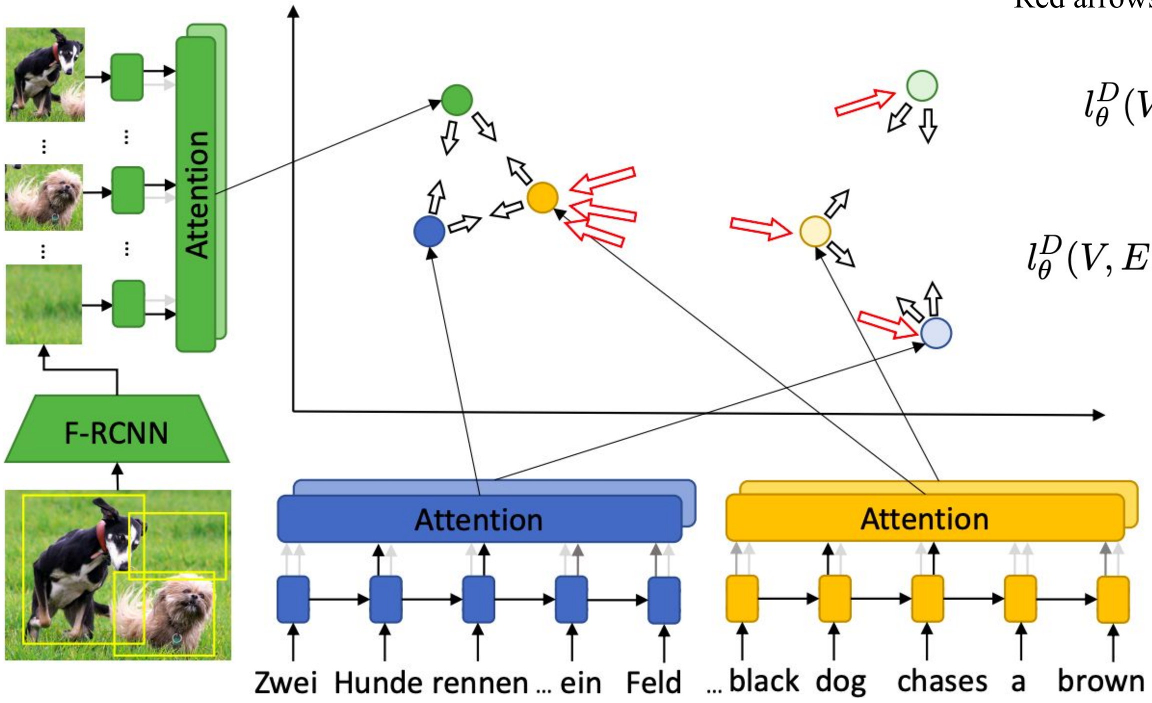
Black arrows: triple correlation objective:

$$l_\theta(V,E) = \sum_p \left[\alpha - s(v_p, e_p) + s(v_p, \hat{e}_p)\right]_+ + \sum_q \left[\alpha - s(v_q, e_q) + s(\hat{v}_q, e_q)\right]_+$$

$$l_\theta(V,E,G) = l_\theta(V,G) + l_\theta(V,E) + \gamma l_\theta(G,E)$$

The proposed model with multi-head attention for associating visual objects and words in the joint multilingual multimodal embedding space.

Model



Red arrows: attention diversity objective:

$$l_\theta^D(V,E) = \sum_p \sum_k \sum_r [\alpha_D - s(v_p^k, e_p^{k\neq r})]_+$$

$$l_\theta^D(V,E,G) = l_\theta^D(V,V) + l_\theta^D(G,G) + l_\theta^D(E,E) + l_\theta^D(V,E) + l_\theta^D(V,G) + l_\theta^D(G,E),$$

The proposed model with multi-head attention for associating visual objects and words in the joint multilingual multimodal embedding space.

Experiments

Dataset: Multi30K (Multilingual version of Flickr30K)

Language: English, German

Attention heads: 3; **Embedding space dim:** 512

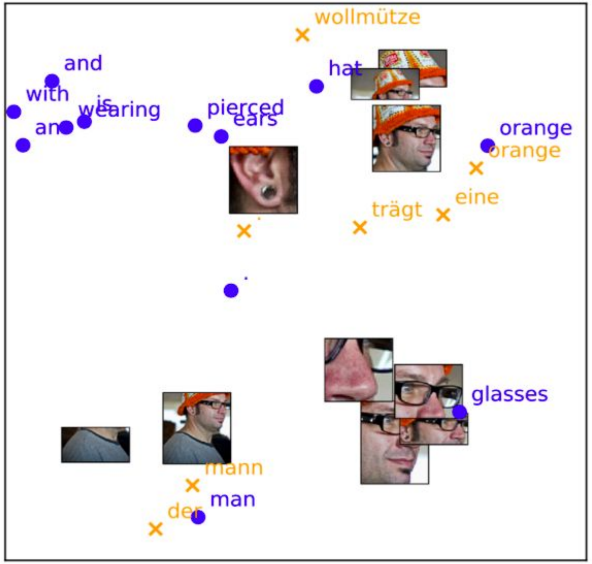
Tasks:

- English-Image matching (Metric: Recall at k)
- German-Image matching (Metric: Recall at k)

Qualitative Results

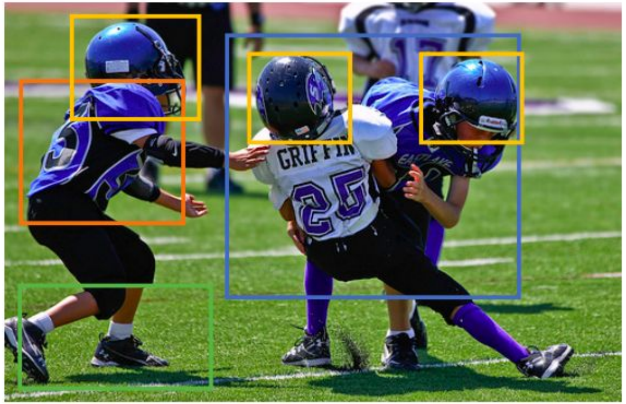
t-SNE Visualization of the multilingual visual-semantic embedding space

The man with pierced ears is wearing glasses and an orange hat
Der mann trägt eine orange wollmütze .



Qualitative Results

Grounded fine-grained multilingual word-visual object alignements



Three children in football uniforms of two different teams are playing football on a football field .
3 kinder am sportplatz , zwei im blauen dress , einer im schwarz-weißen mit blauen schutzhelmen rangeln .

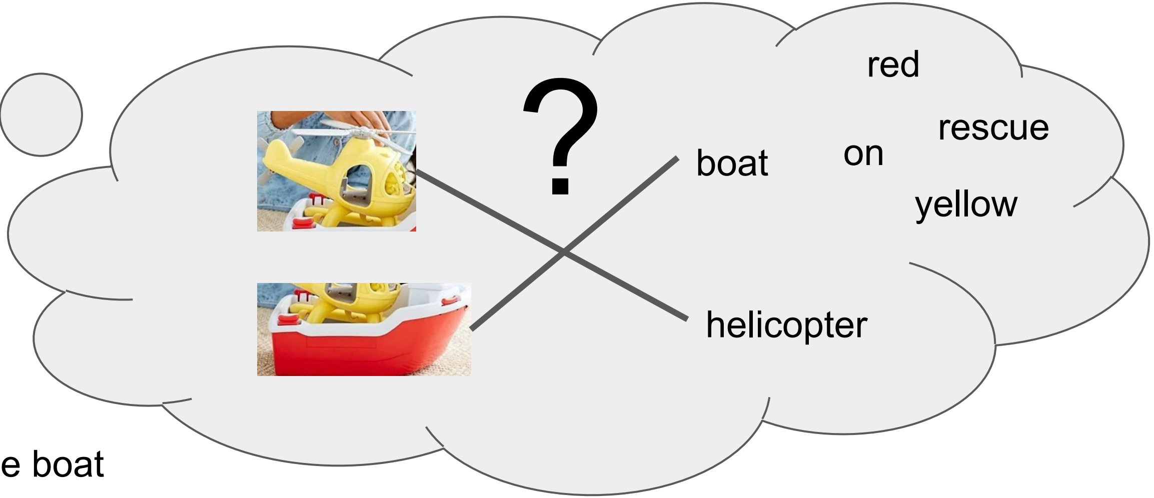
A woman midair vaulting over a bar .
Die frau springt über die stange auf die matte .

Grounded Visual-Verbal Relation Acquisition

Goal:

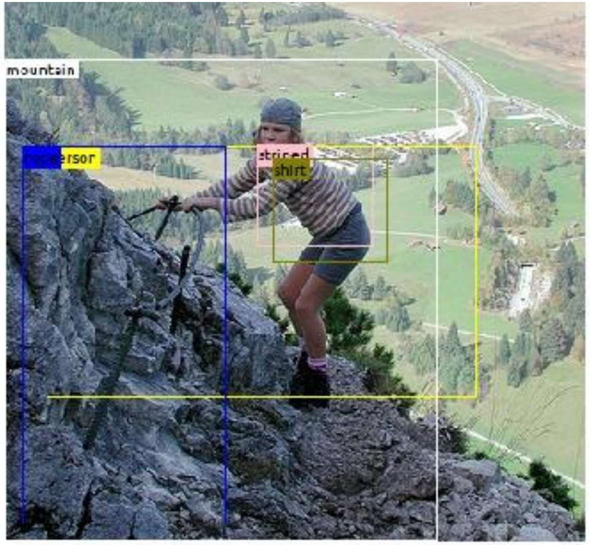
- Learning a model which associates words and visual objects to investigate and mimic the multimodal learning process of humans.

A yellow helicopter on a red rescue boat



Grounded Visual-Verbal Relation Acquisition

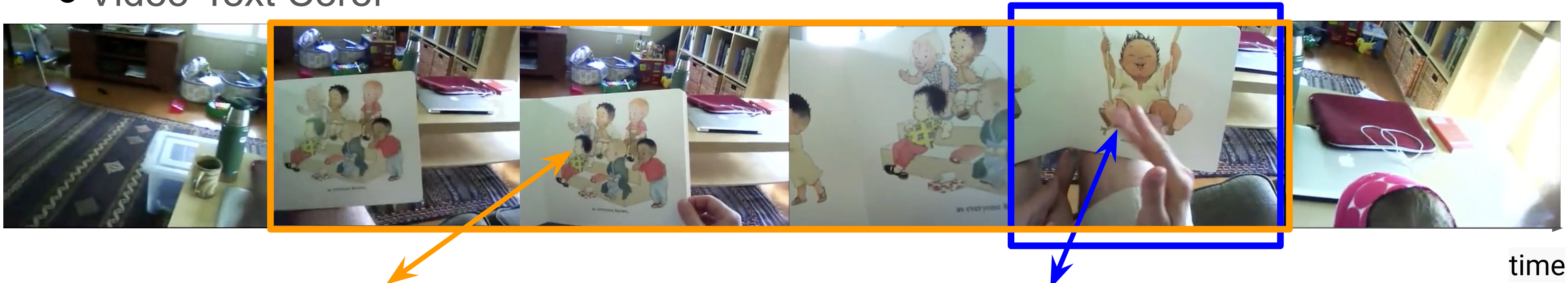
- Use guidance of co-occurring visual scenes and verbal descriptions.
- Encode and align visual-textual pairs into the shared multilingual multimodal representations where semantically correlated word tokens and visual objects are close to each other.
- Completed:
    - Image-text association and grounding
- What's New:
    - Video-text association and grounding



the person has a striped shirt on and is holding on to a rope on a mountain .

New work in progress: Video-Text Coref

- Target
  - Temporal localization (finding video segments/clips associated with the text mentions/descriptions)
  - Spatial + Temporal localization (future plan)
  - Learning visual-semantic embeddings for video-text coref
- Video-Text Coref



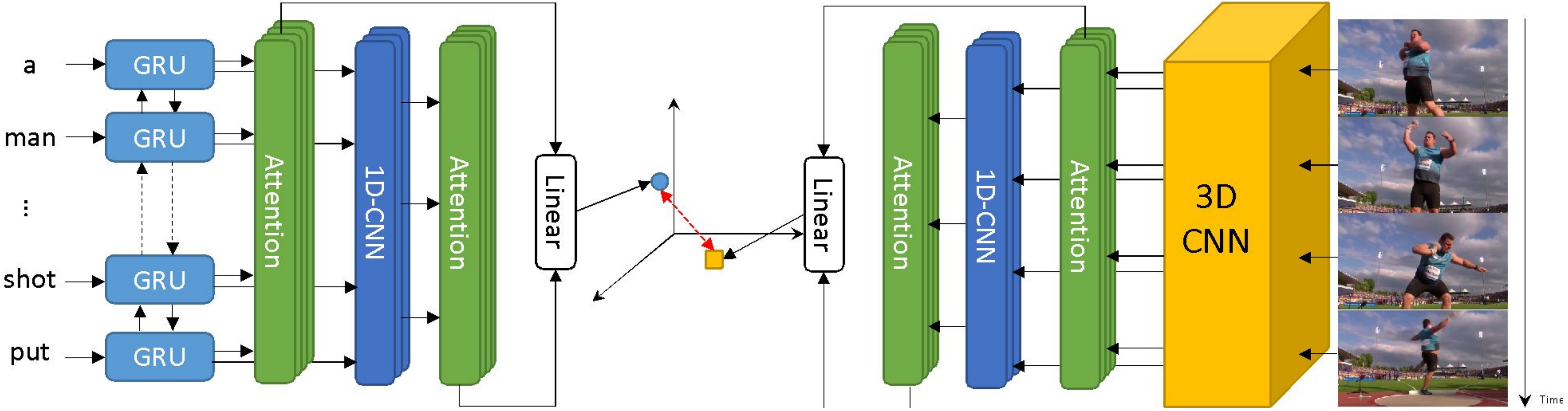
And both of these babies as everyone knows. Yeah…. Had ten little fingers and ten little toes. ….

Video-Text Coref

- Target
  - Temporal localization
    - (finding video segments/clips associated with the text mentions/descriptions) (in progress)
  - Spatial + Temporal localization (future plan)
- Challenge
  - Lack of video-text data and reliable annotation.
- Solution: *Cross-modal Transferring Pre-Training*
  - We propose to use an image-to-video generator to generate ``pseudo videos'' from well-annotated image-text data for (pre-)training video-text coref models.
    - GAN-based Generator (MOCO-GAN) or A simple Augment-and-concatenate generator

Video-Text Coref

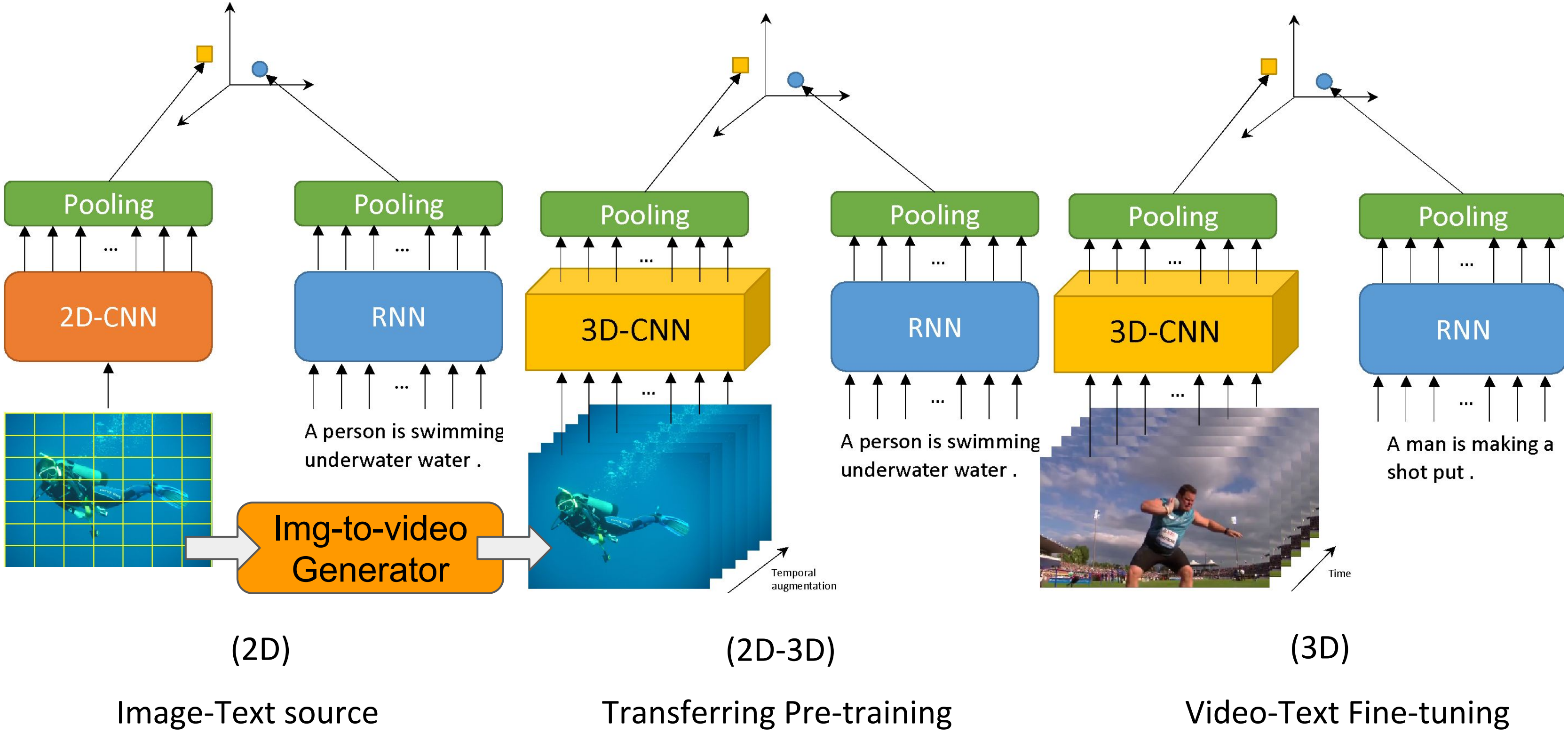
- Model: Hierarchical Multi-head Attention Network



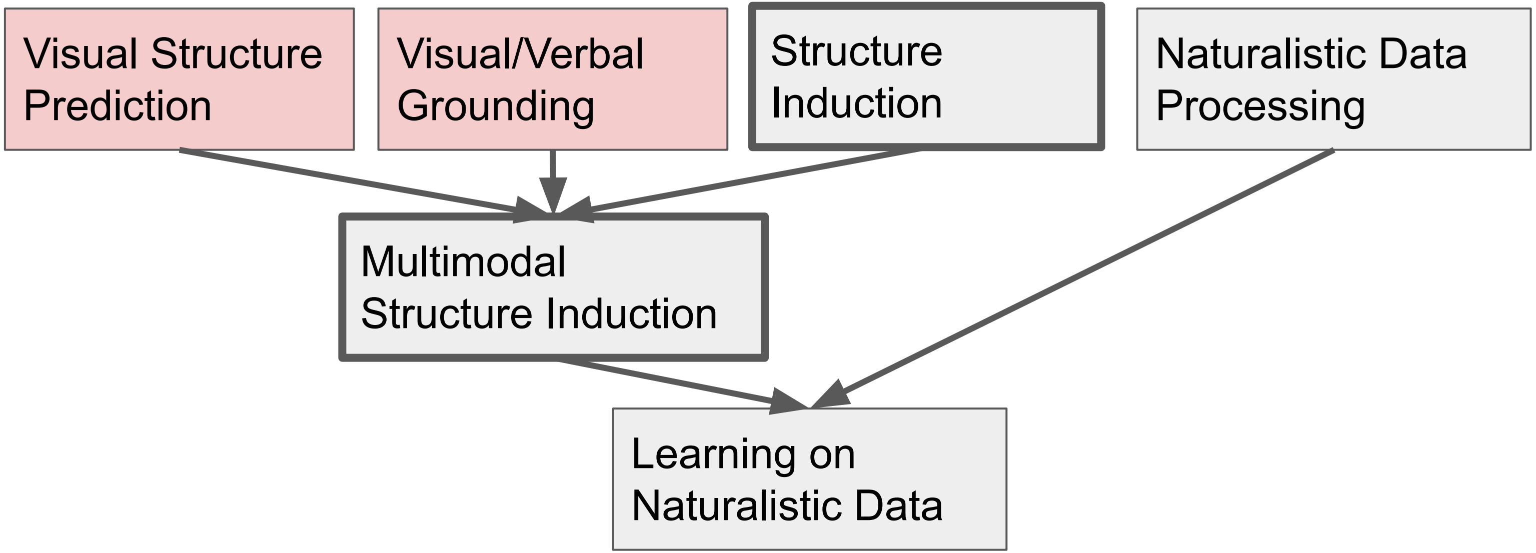
- For encoding spatial-temporal info in videos:
  - 3D CNN (spatial + temporal) encoder
  - (Dilated) 1D-CNN to capture long-term temporal dependencies and increase temporal receptive fields

Video-Text Coref

Cross-Modal Transferring Pre-Training



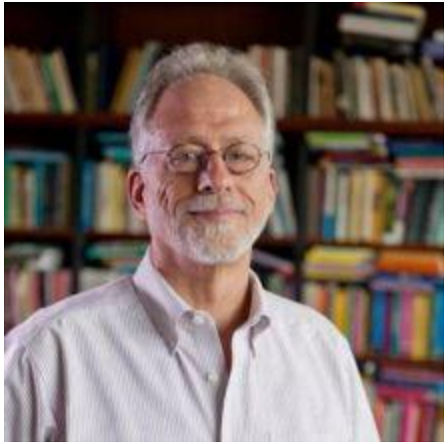
Structure Induction



Two Theories of Human Learning

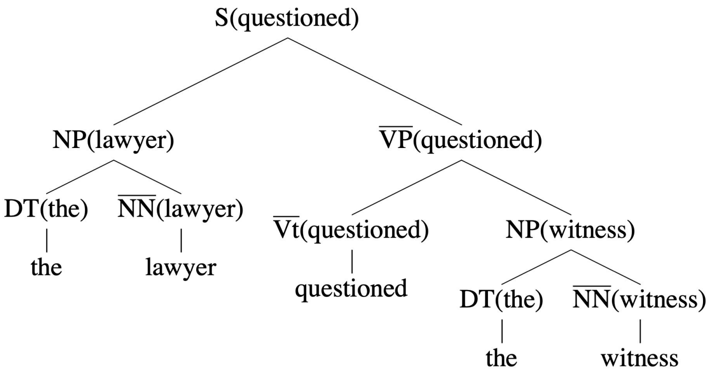
<u>Universal Grammar</u>
(e.g. Chomsky)

<u>Language Acquisition w/ Templates</u> (e.g. Tomasello)

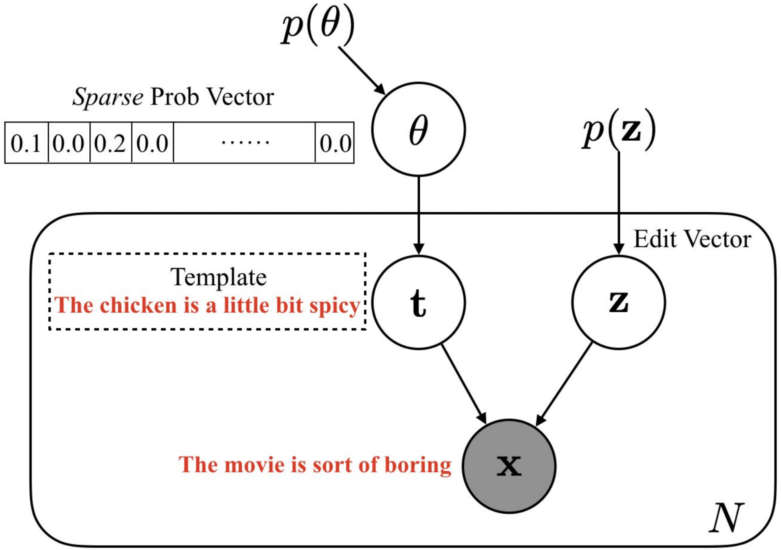


Two Approaches to Latent Structure Learning

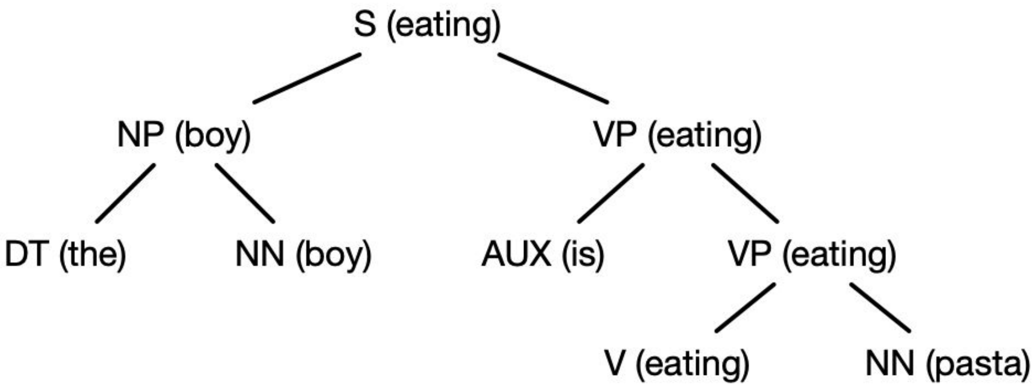
Latent Tree Learning



Latent Template Learning



Latent Tree Formalism: Lexicalized PCFG
(review from last PI meeting)



Context free grammar, where each phrase is associated with a head word

- More powerful formalism than simple PCFG
- Gives us both phrase structure and dependencies between words (important for multimodal grounding!)

Probabilistic Model of Lexicalized PCFG

A lexicalized CFG takes the following form:

Left-headed rule: left child inherits the head word from parent.

Right-headed rule: right child inherits the head word from parent.

$$
\begin{array}{ll}
S \rightarrow A[\alpha], & P(A, \alpha \mid S) \\
A[\alpha] \rightarrow B[\alpha]C[\beta], & P(B, C, \beta, \curvearrowleft \mid A, \alpha) \\
A[\alpha] \rightarrow B[\beta]C[\alpha], & P(B, C, \beta, \curvearrowright \mid A, \alpha) \\
T[\alpha] \rightarrow \alpha, & 1
\end{array}
$$

where $A \in \mathcal{N}, B, C \in \mathcal{N} \cup \mathcal{P}, T \in \mathcal{P}, \alpha, \beta \in \Sigma$.

Left-headed

Right-headed

Latent Tree Learning: Probability Factorization

$$P(A, \alpha \mid S) = P(A \mid S)P(\alpha \mid A)$$

$$P(B, C \mid A, \alpha, \curvearrowleft) \propto \exp\,[\boldsymbol{u}_A; \boldsymbol{\alpha}]^T \boldsymbol{w}_{BC\curvearrowleft}$$

$$P(B, \curvearrowleft \mid A, \alpha) \propto \exp f_{\mathtt{MLP}}([\boldsymbol{u}_A; \boldsymbol{\alpha}])^T \boldsymbol{w}_{B\curvearrowleft}$$

Take the left-headed rule as an example, we parameterize the probabilities using dot products of representations.

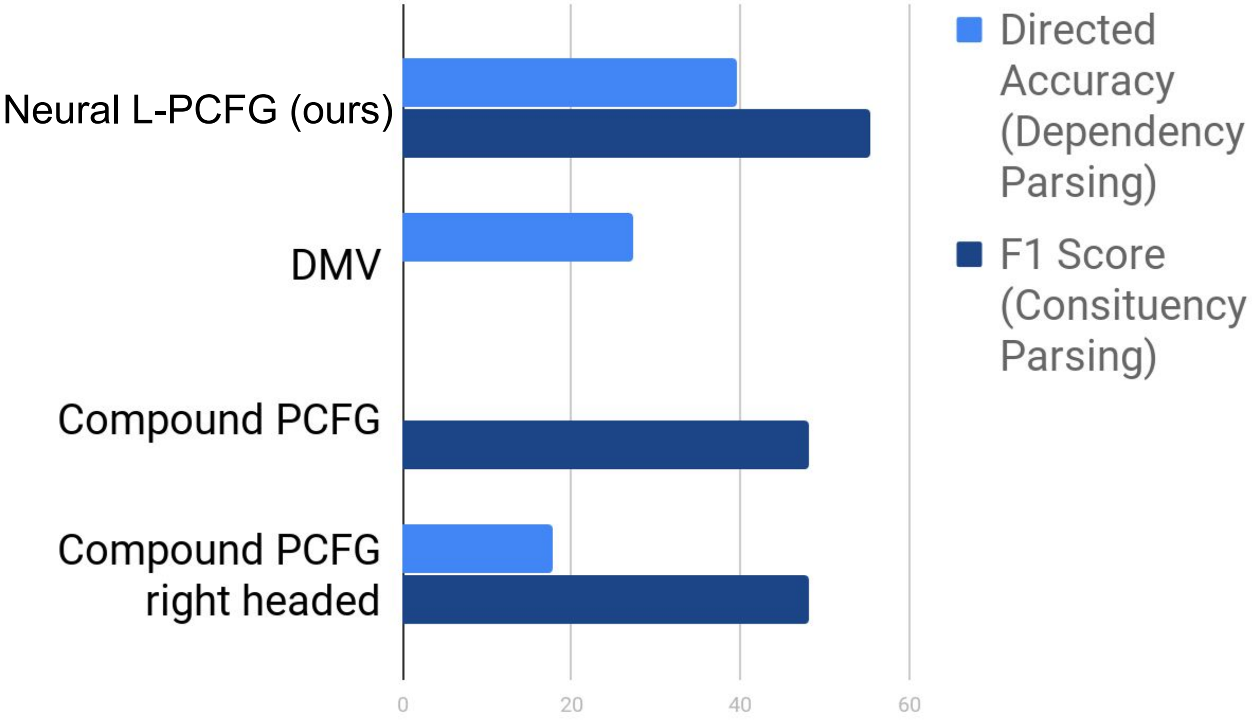
Latent Tree Learning: Experimental Setup

- **Data:**
  - Penn Treebank (Marcus et al., 1993), dependencies created using universal dependency rules from Stanford Core NLP (Manning et al., 2014)
  - MSCOCO (Lin et al., 2014)

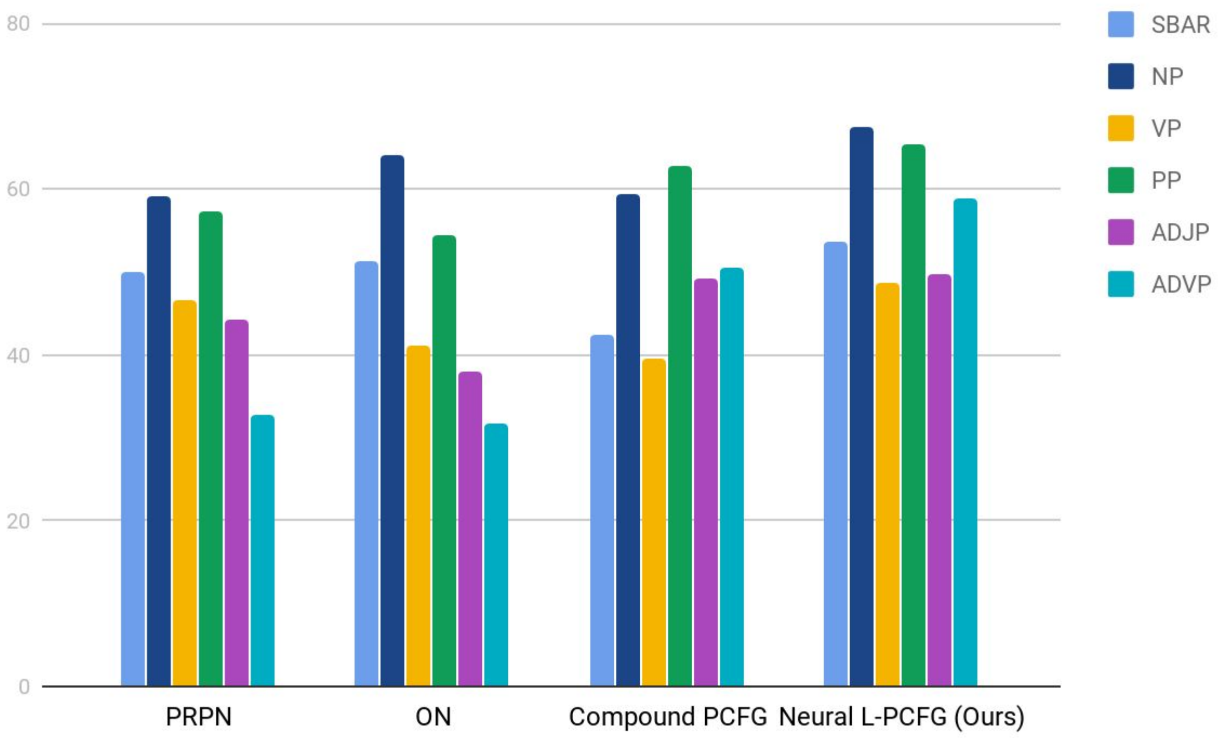
Latent Tree Learning: Baselines

- **Baselines:**
  - DMV (Klein and Manning, 2004): generative model of dependency structures.
  - Compound PCFG (Kim et al., 2019): neural model to parameterize probabilistic context-free grammar using sentence-by-sentence parameters and variational training.
  - Compound PCFG w/ right-headed rule: takes predictions of Compound PCFG and choose the head of right child as the head of the parent.
  - ON-LSTM (Shen et al., 2019) and PRPN (Shen et al., 2018): two unsupervised constituency parsing models
  - VGNSL (Shi et al., 2019): unsupervised constituency parsing model with image information

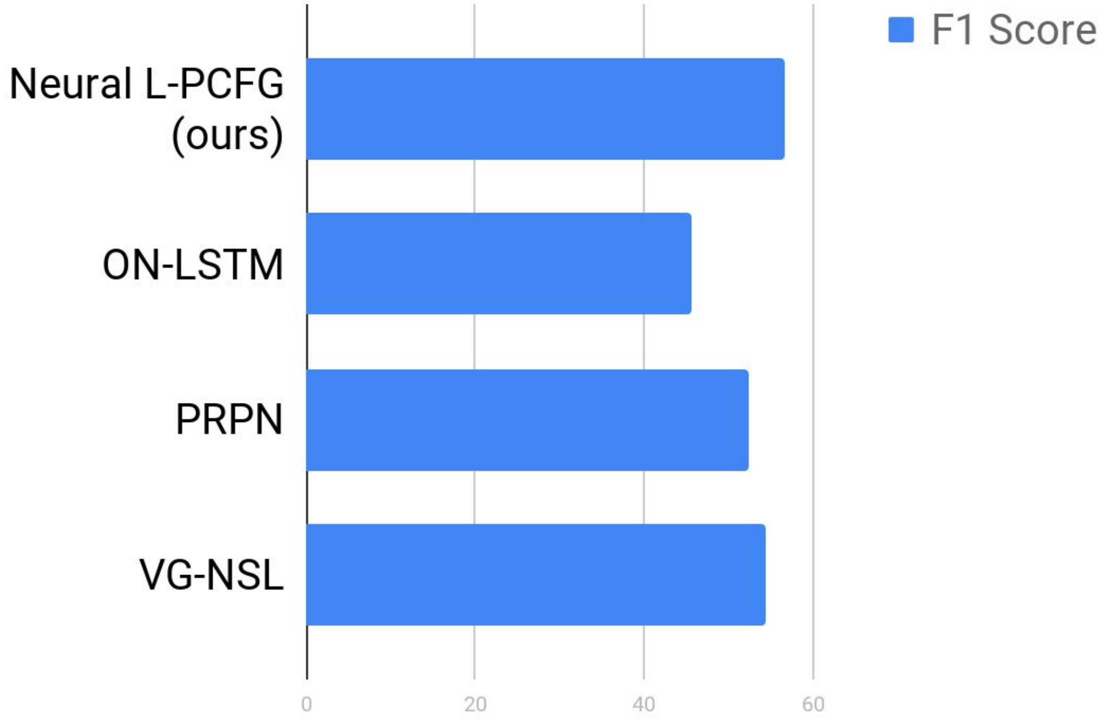
Latent Tree Learning: PTB Results



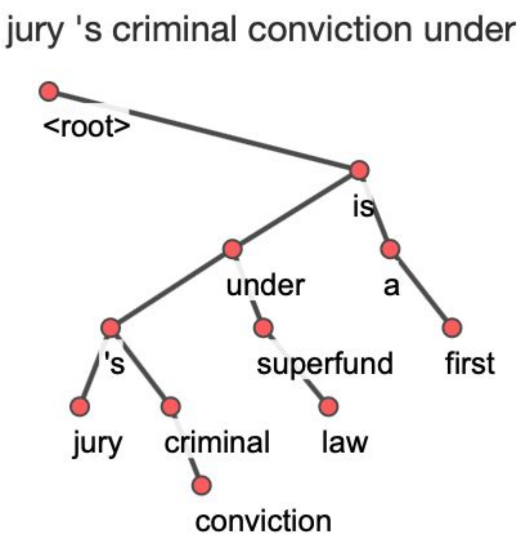
Latent Tree Learning: PTB Label-Level Recall

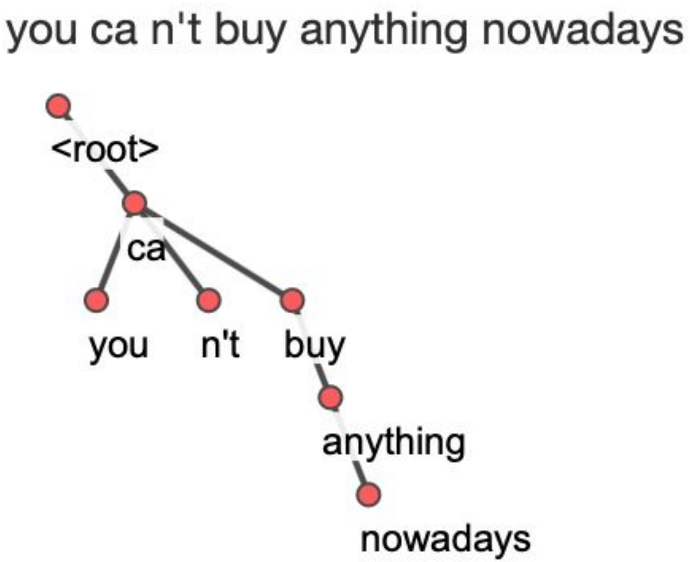


Latent Tree Learning: MSCOCO Results



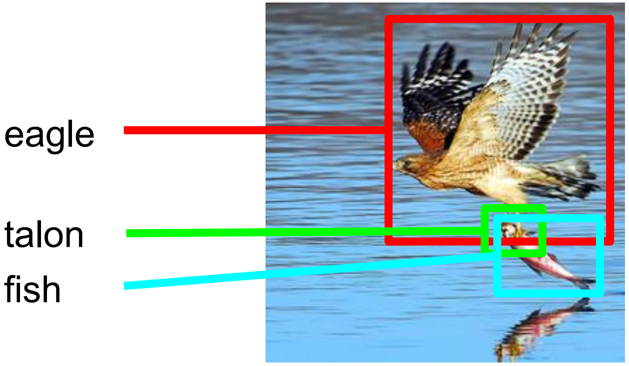
Latent Tree Learning: Visualization



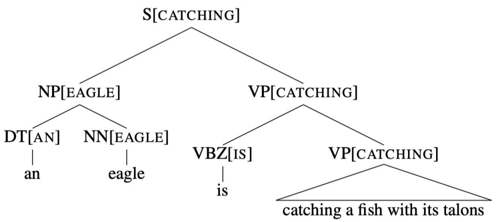


[Visualization Website]

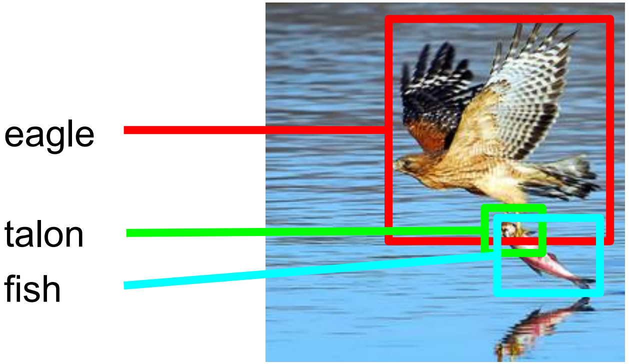
Latent Multimodal Tree Learning



An eagle is catching a fish with its talons.

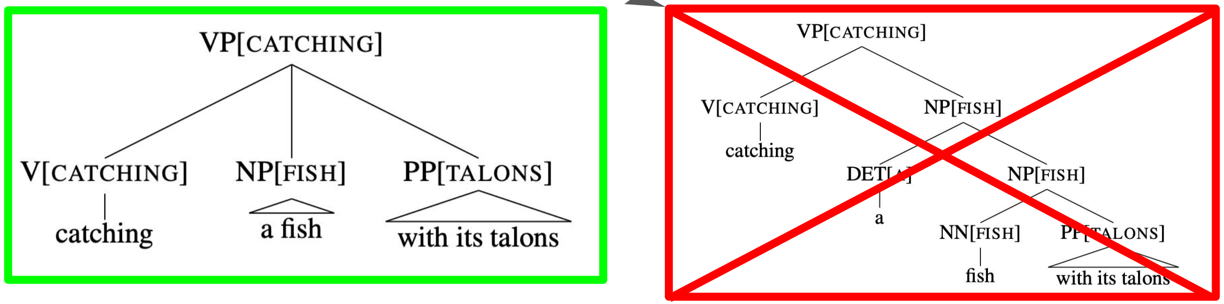


Latent Multimodal Tree Learning



Visual Semantic Role Labelling

action	catch
agent	eagle
patient	fish
instrument	talon



An eagle is catching a fish with its talons.

Latent Multimodal Tree Learning Constraints

Visual information conveys the relationships between objects within the image

Situation recognition: **activity, participants, roles of participants**

Alignment: **activity** vs. **predicate**, **participants** vs. **arguments**

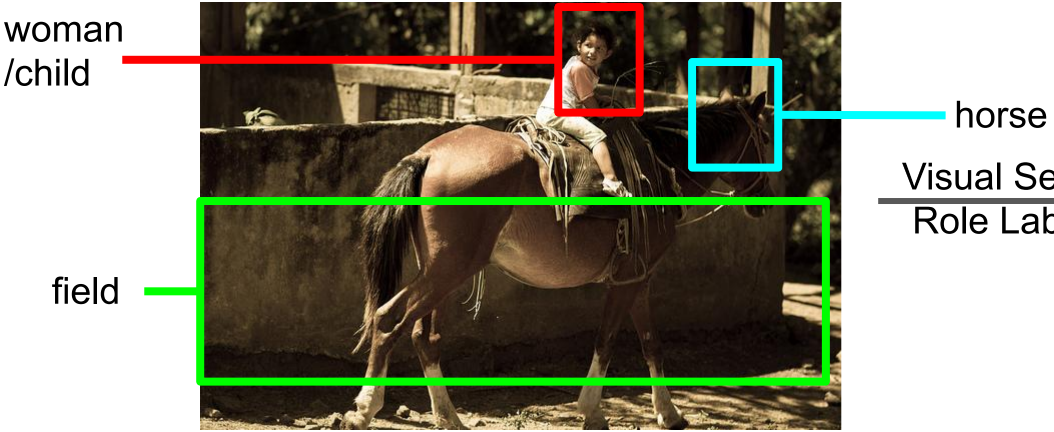
Constraints

caption: **dogs eat food at home** frame: **eat_agent_dog, eat_food_food, eat_place_home**

1. two arguments belonging to the same predicate **should not** exist in a phrase **unless** the predicate also exists in that phrase (~~food at home~~, food at, at home)

2. an argument **cannot** be the head of a phrase that also contains its predicate

Results of Latent Multimodal Tree Learning

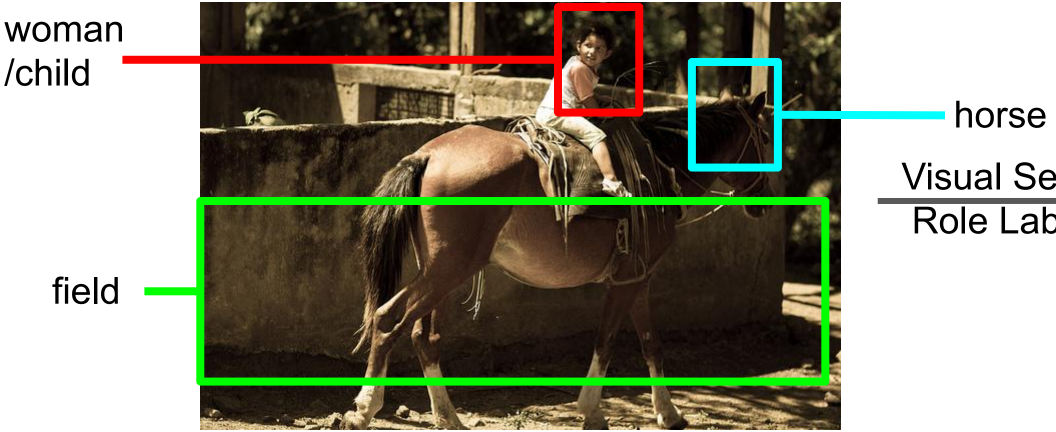


Visual Semantic Role Labelling →

agent	woman
vehicle	horse
place	field

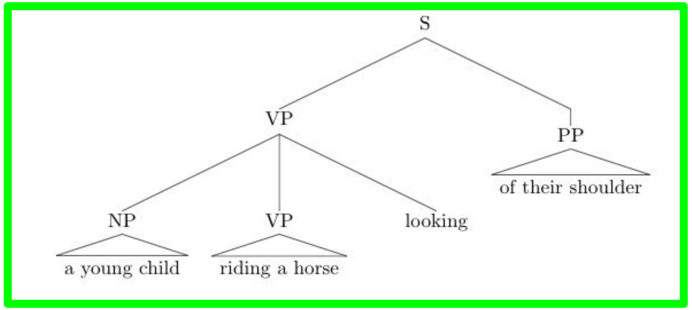
A young child riding a horse looking of their shoulder

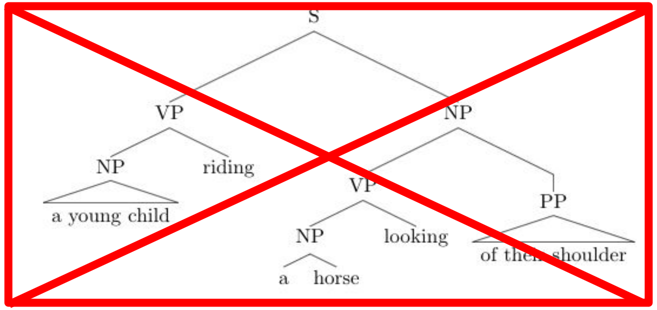
Results of Latent Multimodal Tree Learning



Visual Semantic Role Labelling

agent	woman
vehicle	horse
place	field





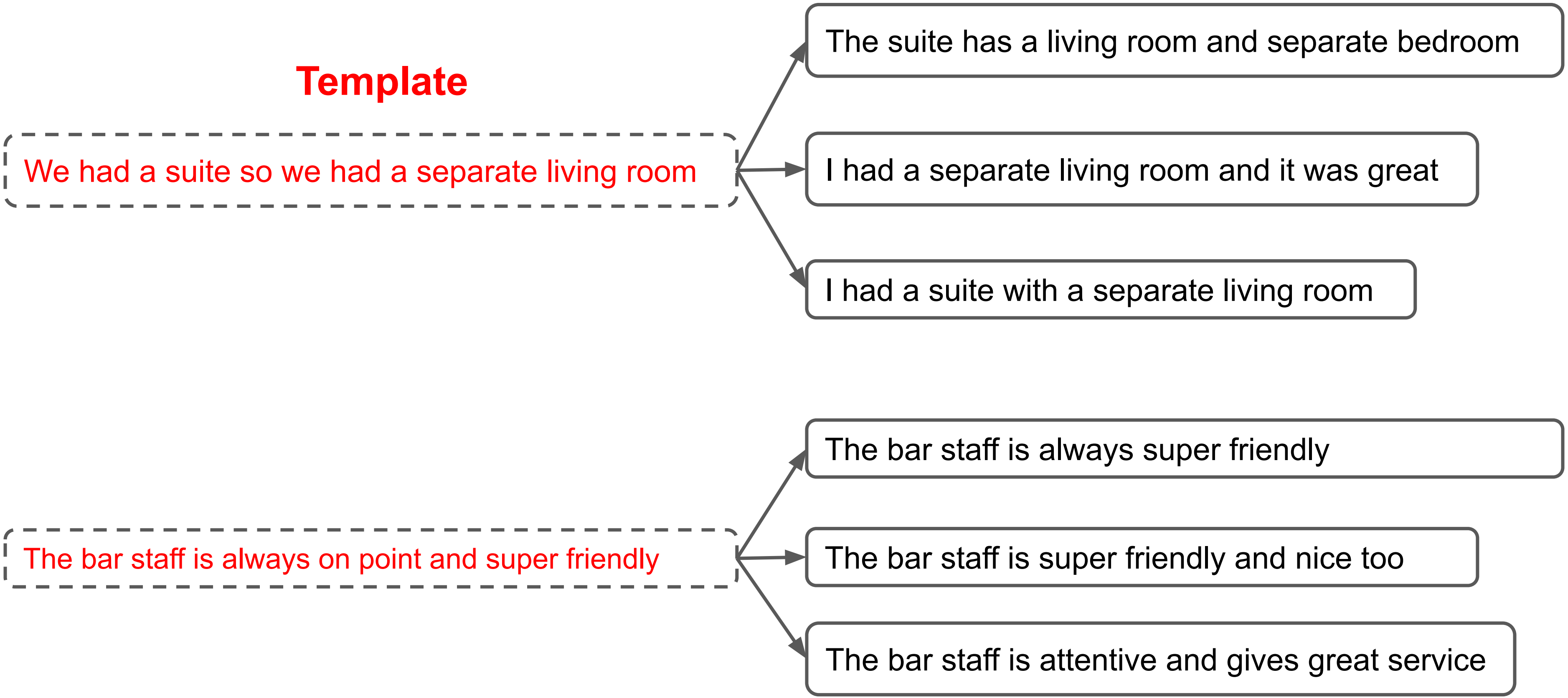
alignment	
caption	frame
child	woman

A young child riding a horse looking of their shoulder

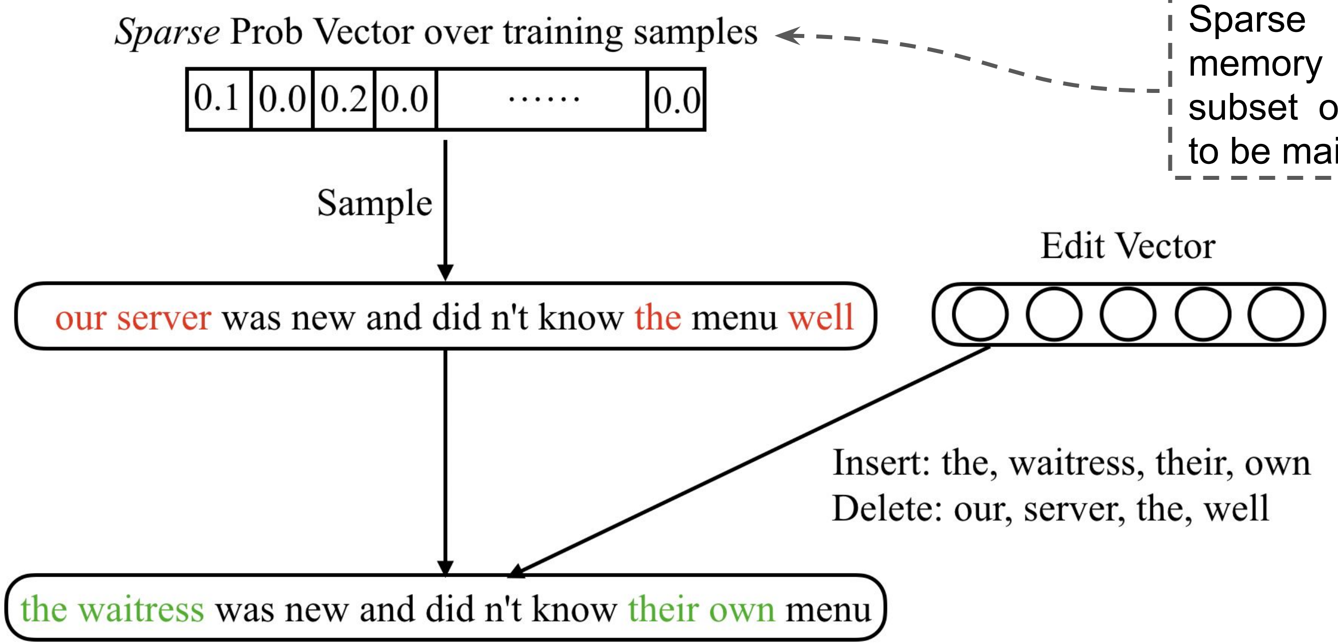
Latent Template Learning: Motivation

- Research on child language development (e.g. usage-based theory of Tomasello 2005) shows **children may learn templates, then generalize**
- Can we create language generation models that learn in a similar way?
- Maybe templates can be associated with semantic frames?

Latent Template Learning: Concept

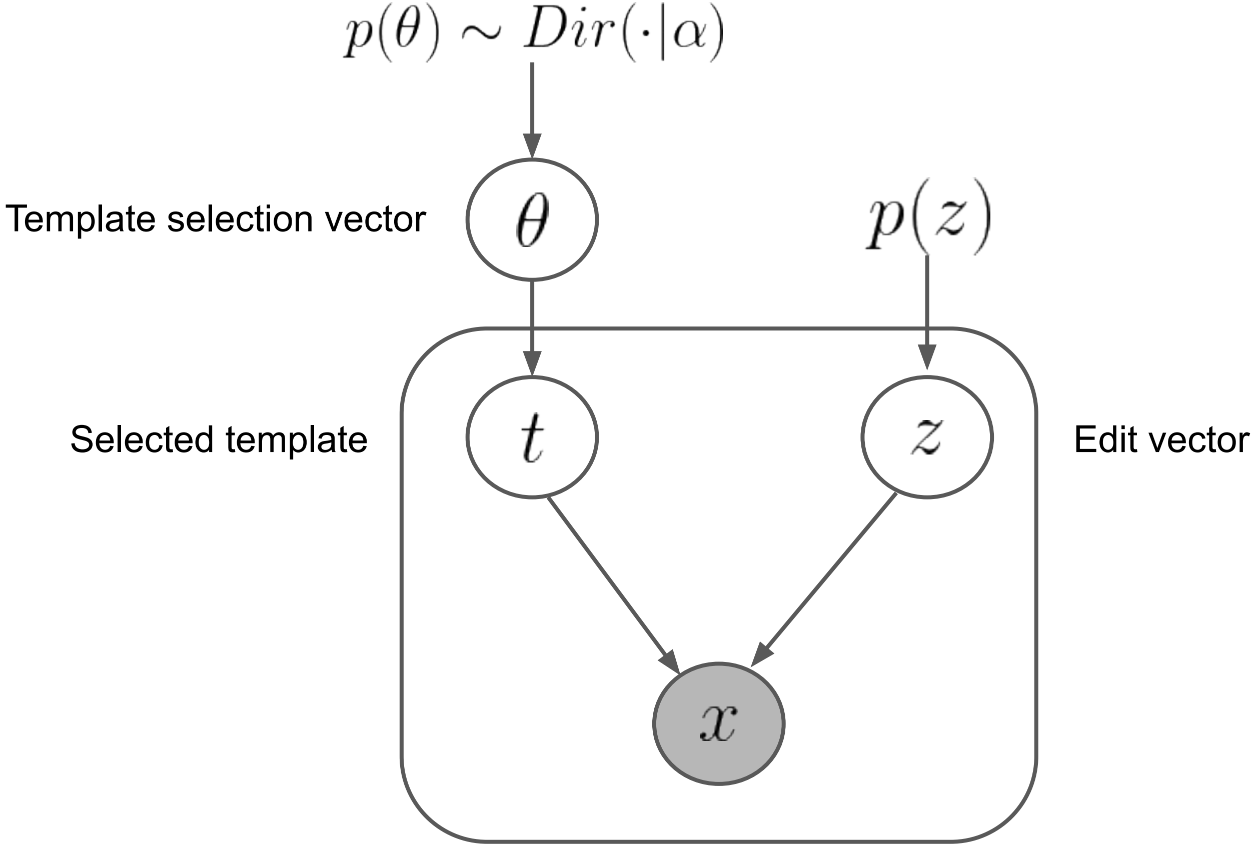


Latent Template Learning: Generative Model

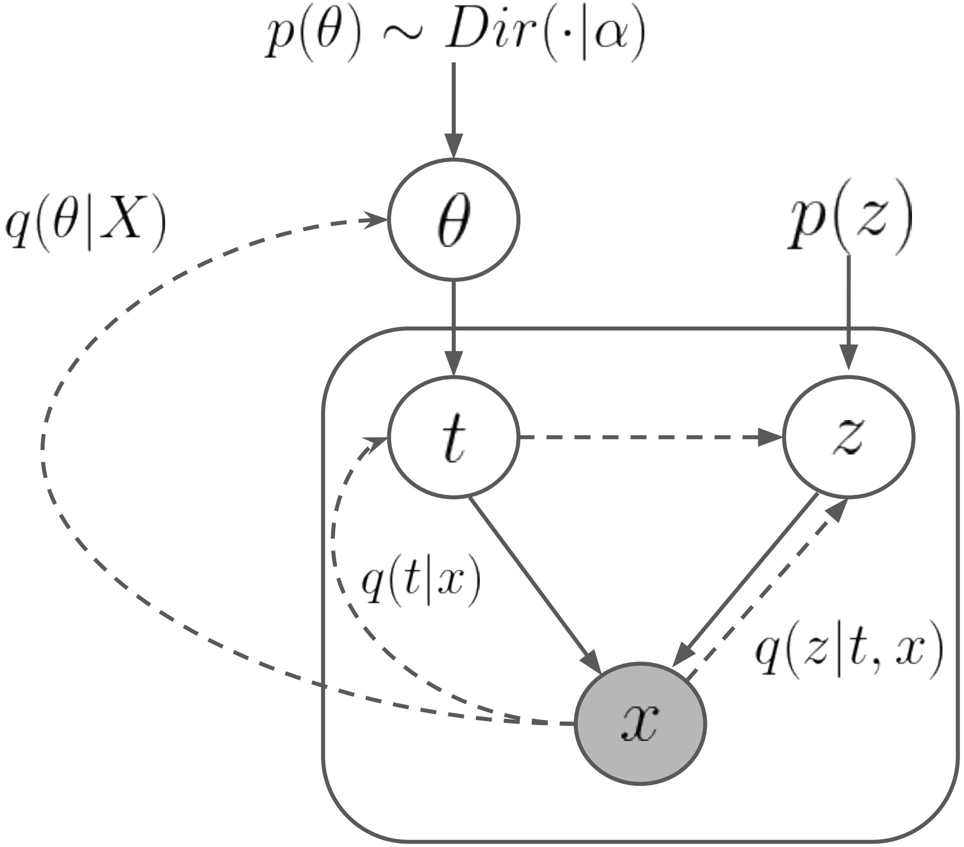


Sparse constraint makes it memory efficient: only a small subset of training samples need to be maintained as templates

Latent Template Learning: Generative Model



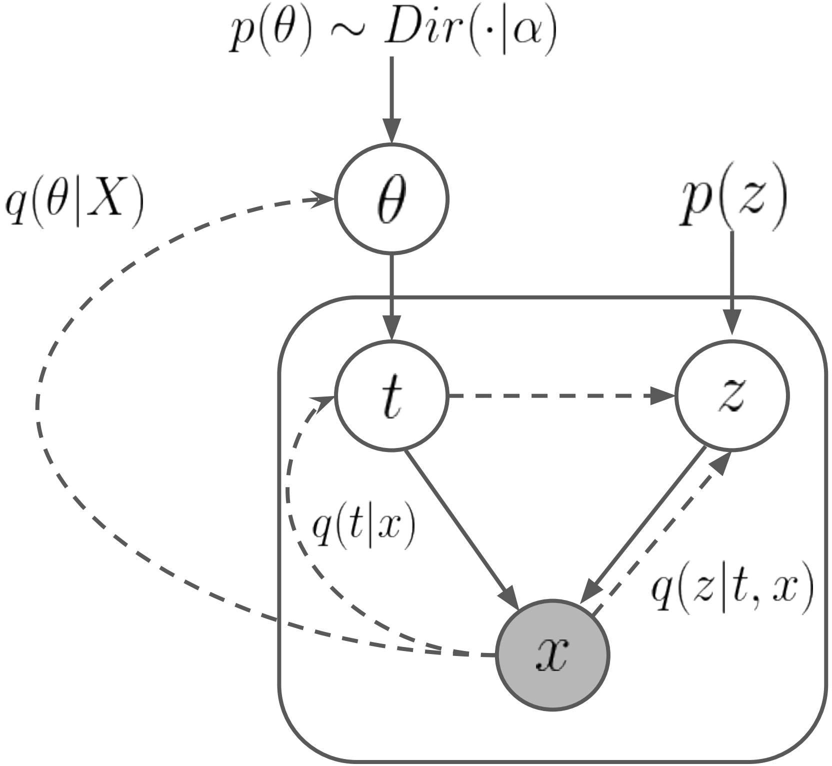
Learning of the Latent Template Model



$q(t|x)$: template retriever

$q(z|t, x)$: inverse editor

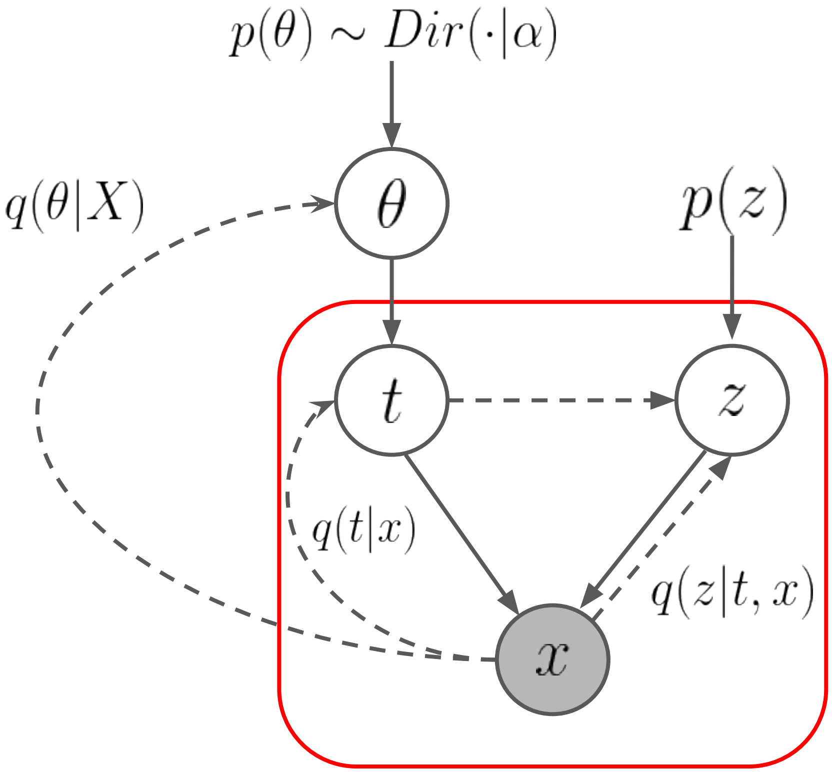
Learning of the Latent Template Model



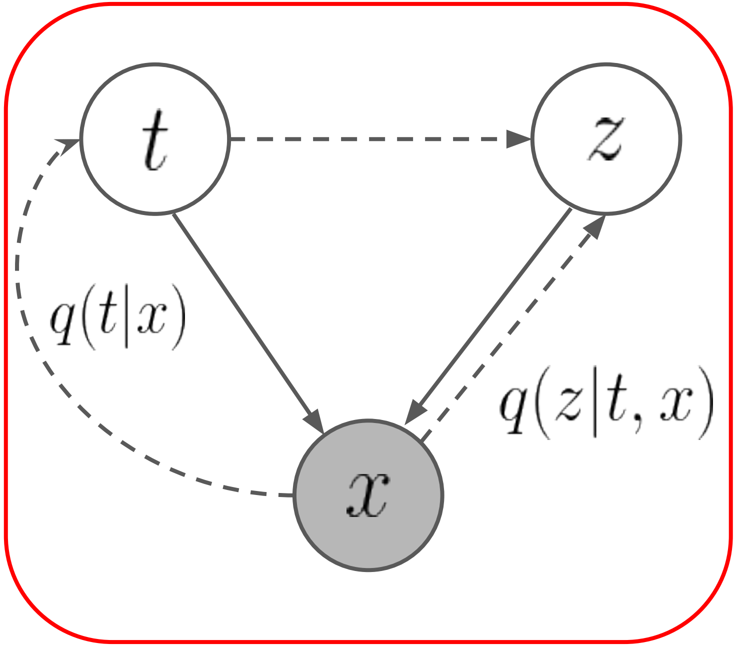
$$q(\theta, t, z|X) = q(\theta|X) \prod_i q(t_i|x_i) q(z_i|x_i, t_i)$$

$$\mathrm{ELBO} = E_q[\log p(\theta)p(t|\theta)p(z)p(x|t, z) - \log q(\theta, t, z|X)]$$

Learning of the Latent Template Model



Learning of the Latent Template Model



$p(x|t,z)$ (editor)

Seq2Seq model

$q(t|x)$ (retriever)

retrieve based on Bert-based embeddings

$q(z|t,x)$ (inverse editor)

= D X = X

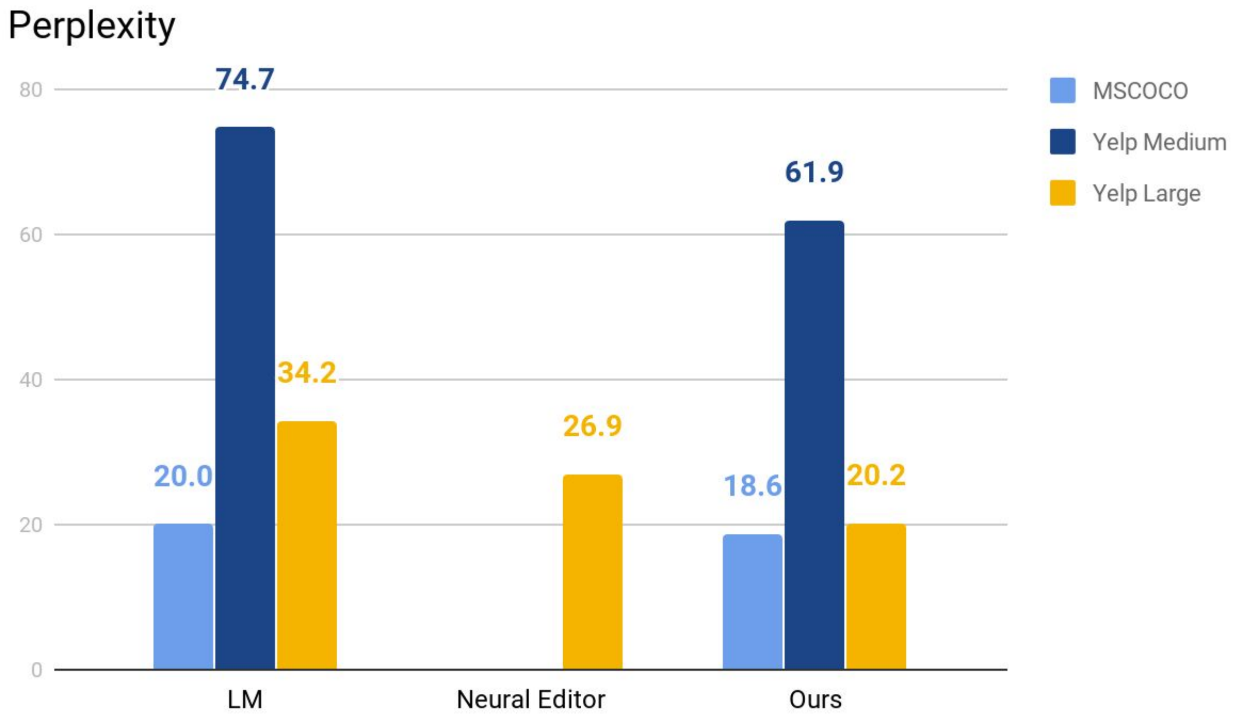
The white dog is barking

The - cat is running

Latent Template Learning: Experimental Setup

- **Data:**
  - MSCOCO sampled set: 40K training examples
  - Yelp Medium/Yelp Large: 1.5M and 17M training examples respectively
- **Baselines:**
  - *LM*: vanilla LSTM language model without latent variables
  - *Neural Editor* (Guu et al. 2018): prototype-based language model but with dense prototype library (e.g. the entire training set) and prefixed prototypes for each training example through heuristics

Results on Language Modeling



Results on Efficiency

	Model	PPL	# Templates	Test speed (sents/s)
Yelp Medium	LM	74.7	-	236
	Ours	61.9	1.5K	107
Yelp Large	LM	34.2	-	272
	Neural Editor	26.9	17M	0.1
	Ours	**20.2**	**2K**	108

We achieve 1000x memory savings and 1000x speed-up at test time over the previous neural editor baseline

Analysis on Sparsity Variation

Sparsity can be controlled through the Dirichlet prior, and templates (prototypes) tend to focus on syntax when they grow sparser

Model	Overall	NOUN	DET	AUX	PRON	ADJ	VERB	CCONJ
Ours (31K prototypes)	91.2K	14.4K	9.6K	9.3K	9.0K	7.2K	6.4K	5.5K
Ours (1.5K prototypes)	74.7K	9.9K	8.5K	8.2K	7.3K	5.6K	4.4K	5.0K
Relative Change	-18.1%	**-31.3%**	-11.5%	-11.8%	**-18.9%**	**-22.2%**	**-31.3%**	-9.1%

Number of matching tokens between examples and their templates under two different sparsity settings. Results are reported in cluster of POS tags.

Analysis on Varying Sparsity

Sparsity can be controlled through the Dirichlet prior, and templates (prototypes) tend to focus on syntax when they grow sparser

Data Examples	Prototypes
the best corned beef hash i ’ve ever had !	(dense) the best real corned beef hash i ’ve had . (sparse) the chicken satay is the best i ’ve ever had .
the grilled chicken was flavorful , but too flavorful .	(dense) the chicken was moist but it lacked flavor . (sparse) my sandwich was good but the chicken was a little plain .
i asked her what time they close and she said <cardinal> o’clock .	(dense) i asked what time they closed <date> , and was told <cardinal> . (sparse) we asked how long the wait was and we were informed it would be <time> .

Qualitative examples of retrieved prototypes under relatively dense and sparse settings

Generation Conditioned on Templates

Prototype: A man walking on the beach carrying a surfboard	Prototype: A group of people are raising an umbrella on a beach
Two people standing next to each other on a beach A person standing on the beach holding a surfboard A man walking along the beach with a surfboard A man walking on the beach with a surfboard A young man walking on the beach with a surfboard	A group of people are walking on the beach with umbrellas A group of people are walking on the beach next to each other A group of people are walking on the beach with umbrellas A group of people are holding umbrellas on the beach A group of people are walking on the beach
Prototype: there is a white truck that is driving on the road	Prototype: A couple of bags of luggage sitting up against a wall
there are many cows that are standing in the dirt there are many cows that are standing in the dirt the truck is driving down the road in the rain this truck is driving down the road in the rain This truck is pulled up to the side of the road	A large pile of luggage sitting on top of a wall A pile of luggage sitting on top of a wall Two bags of luggage sitting on the ground Two bags of luggage sitting in a room A couple of bags of luggage on a wooden floor
Prototype: A man riding a sailboat in the ocean next to a shore	Prototype: A beer bottle sitting on a bathroom sink next to a mirror
A man on a boat in a body of water A man riding a boat on a body of water A man riding a boat in a body of water A man riding a small boat on a body of water A man riding a wave on top of a boat	A white cell phone sitting next to a toilet in a bathroom A white bottle of wine sitting next to a toilet A glass of wine sitting next to a toilet in a bathroom A pair of scissors is placed next to a toilet A pair of scissors sitting next to each other on a toilet

Latent Multimodal Template Learning

On image caption dataset like MSCOCO, we can utilize associated images to help retrieve sentence templates

A picture of chocolate milk and a variety of donuts

Joint Encoder

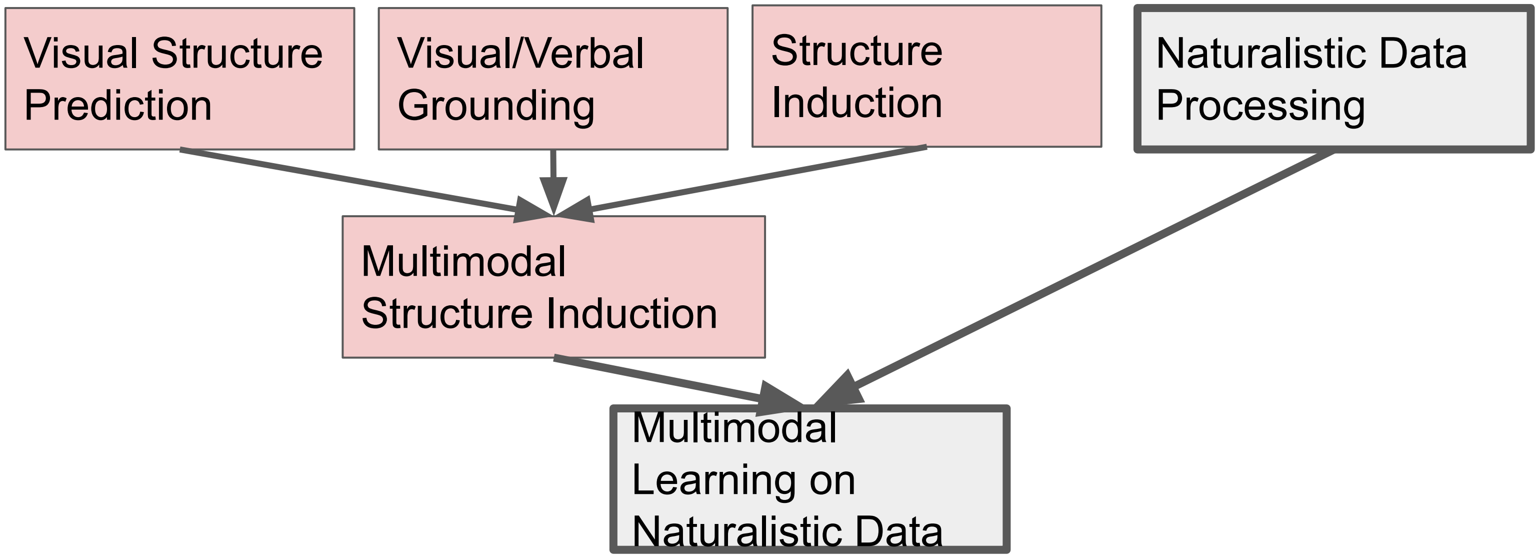
Retriever is based on multimodal embeddings

Preliminary Results on MSCOCO

Model	PPL
LSTM LM	18.85
Our template-based model	19.96

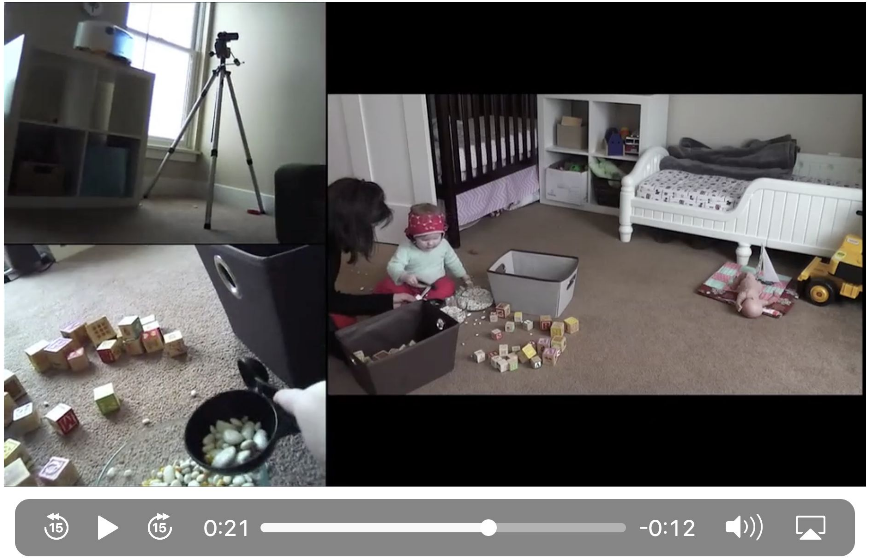
Data samples	Retrieved templates
A kitchen looks very clean with corner cabinets .	A picture of a kitchen that is very clean .
people playing a baseball and many people watching .	A group of people are playing baseball with an audience in the background .
A notebook computer is set on a table .	A laptop and mouse sits on a table .
Three men sitting at a table looking at a cell phone .	A man sitting at a table with a cell phone .

Naturalistic Data Processing



SeedlingS Corpus

- 500 hours of audio and video from 46 children 6-17 months of age
- Objects being referenced in typical speech acts, and visible to child, annotated
- **Manual annotation:** 100+ noun types with 100+ occurrences
- **Not annotated:** full transcripts, all visually present objects
- Available through Databrary: https://nyu.databrary.org/



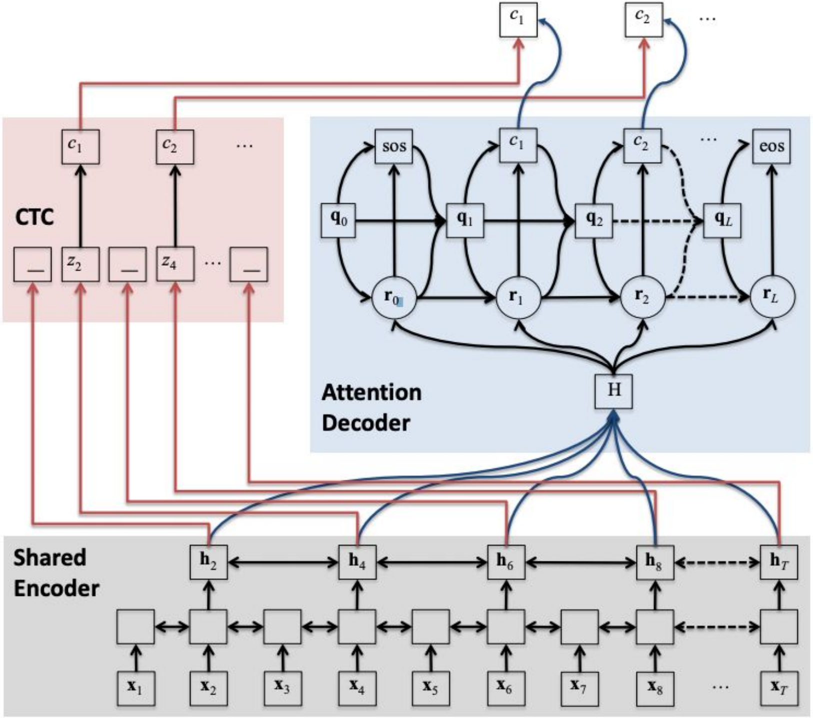
Screenshot: https://bergelsonlab.com/seedlings/

Automatic Speech Recognition

- **Pre-trained models** from more established Speech Recognition corpus, (Libri Speech and SwitchBoard in our case)
- **3 models:** ESP-Net[1], EESEEN-WFST[2], and EESEN-rnnLM decoding, trained with CTC loss
- **Major Challenges:**
  - fully annotated transcription not available for evaluation
  - much more noisy than the pre-trained datasets
  - multiple speakers present

1. EESEN: https://github.com/srvk/eesen

ESP-Net VS EESEN



ESP-Net architecture
Watanabe et al. 2018

- EESEN:

  Requires separate Language Model, conditional independence assumption

- ESP-Net:

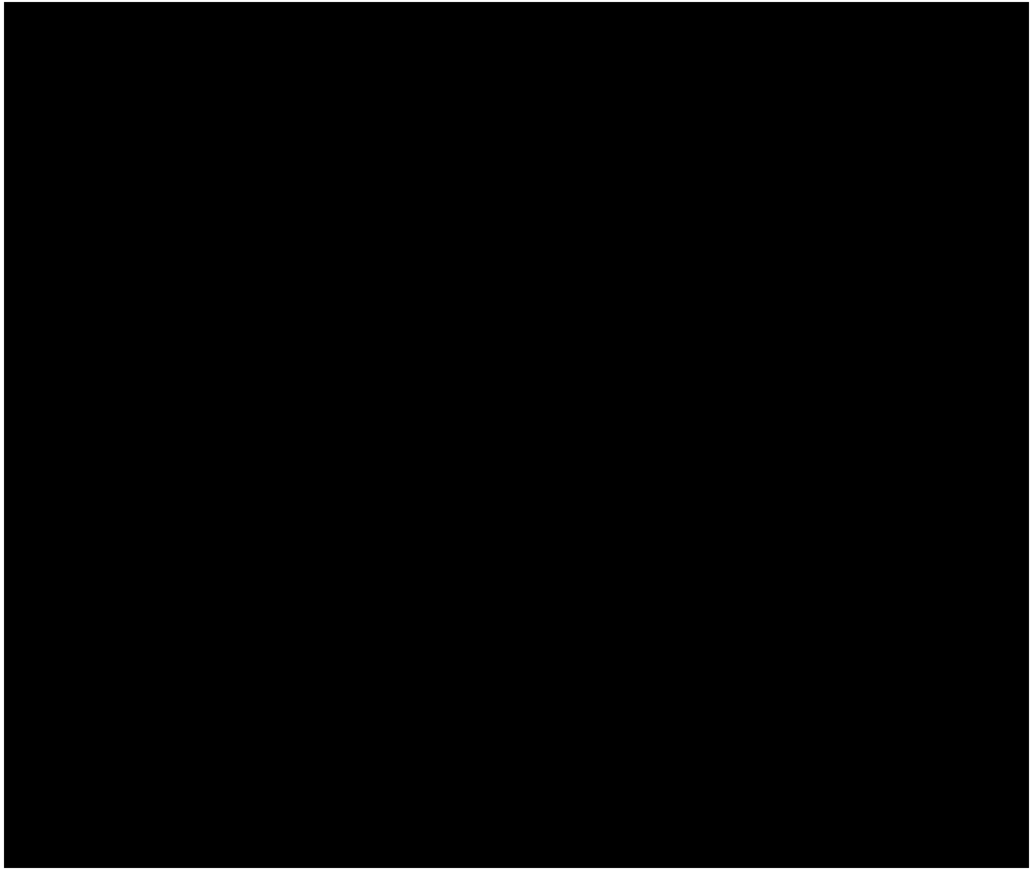
  Utilized hybrid CTC attention Loss that utilizes both benefits

  $L_{mul} = \lambda \log p_{ctc}(C|X) + (1 - \lambda)\log p_{att}(C|X)$

  $C = \{c_l \in U | l = 1, \cdots, L\}$, U is a set of distinct letters, $X = \{x_t \in R^D | t = 1, \cdots, T\}$, $0<\lambda<1$ λ is a tunable parameter

  Faster decoding, no need for LM, irregular alignments, directly estimates the posterior,

ASR: Seedling Dataset Samples(ESPnet vs EESEN)



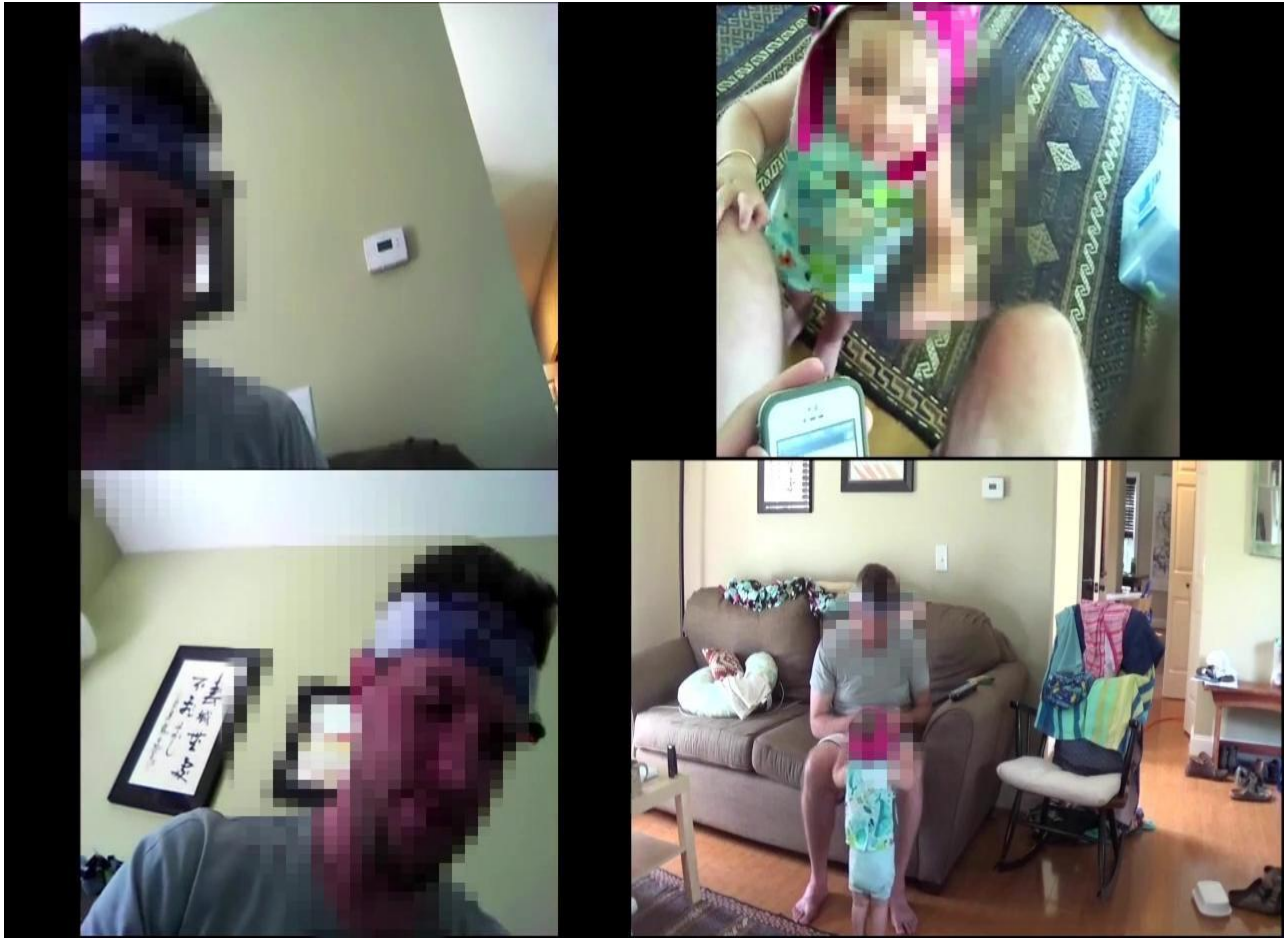
ESPnet: Hey, do you want to play anything or read a book or anything a book? Okay, which book which book you want to read? The watch one little baby who is born far away. And another who is born on the very next day. And both of these babies as everyone knows.

Turn the Page. Had Ten Little Fingers ten fingers and ten little toes. There was only there was one little baby who is born in a town and another who is wrapped in either down. And both of these babies as everyone knows add ten little fingers and ten.

Have you any water recently? Get some water, please. Get some water please some water. Yeah water is delicious. Why don't you have some? Give me some water, please.

There was one little baby who is born in the house and another who Snuffer suffered from sneezes and chills. And both of these babies with everyone knows. at ten little fingers and ten little toes just like

ASR: Less Successful Example



ESP-Net: You get a car going?

Atlantic what sound does a car make?

I'm home. That's right.

Bye-bye. Be back soon.

I'll be back soon. I got to take a picture of you for Mom.

I have to take a picture of you for Mom.

Mama, that's right. Can you smile can you say hi Mom? Hi, Mom.

Yes, indeed. That's wonderful. I want let me send this to Mom and then I'll let you see my phone. Okay?

Liquidators

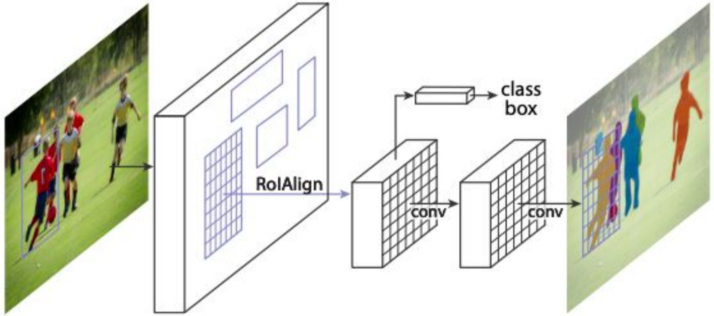
ASR: Quantitative Results

- Measure **how well the ASR results match with the annotated nouns**.
- A word is treated "recognized" if it occurs within a fixed window of the annotated word on either side.
-

	EESEN	ESPNet
Overall Recall	37.87%	41.51%
Father	45%	51%
Grandma	41.6%	48%
Mother	40.5%	42.5%
Aunt	26.4%	35.1%
Brother	12.8%	20.8%

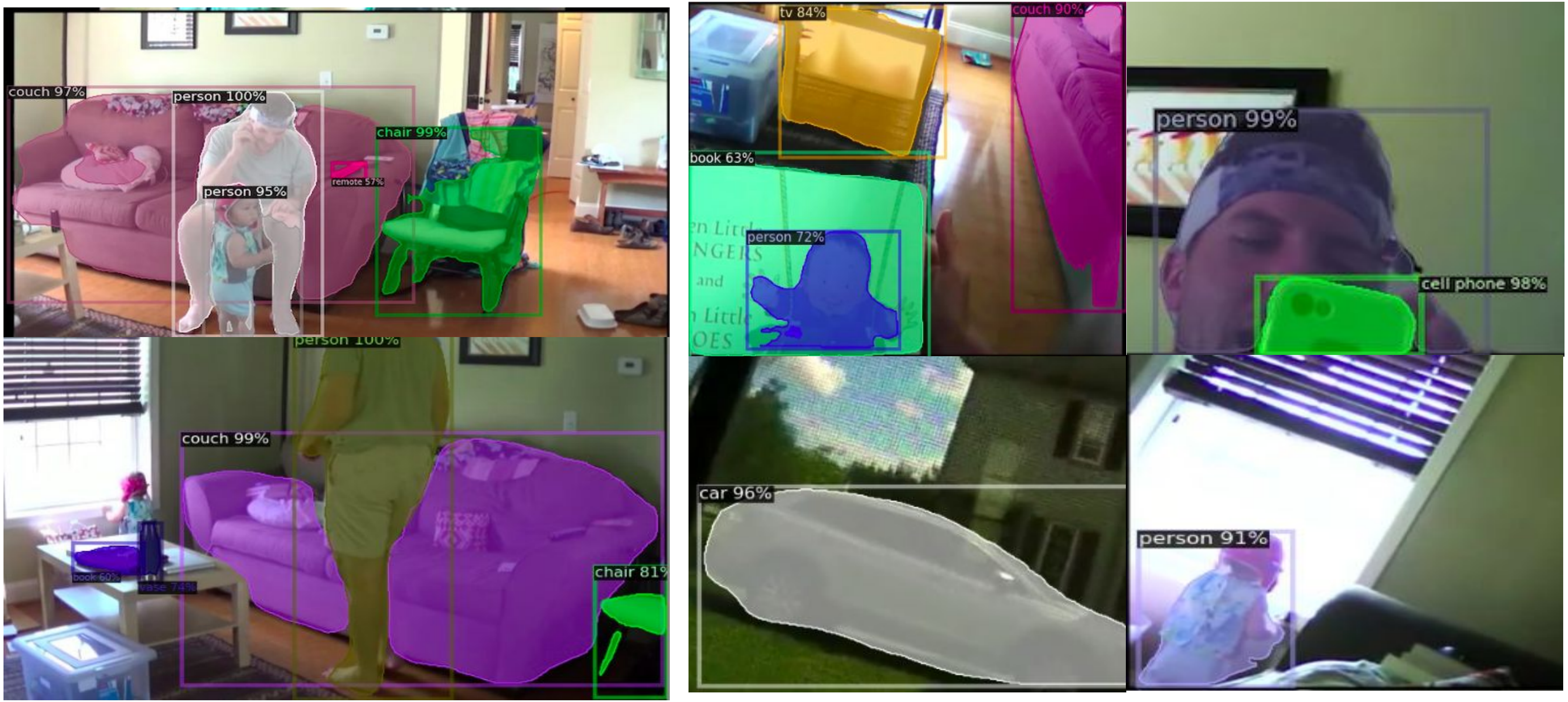
Object Detection: Methodology

- **Model: Mask-RCNN**
  - Detect + segmentation
  - Pretrained on MS-COCO



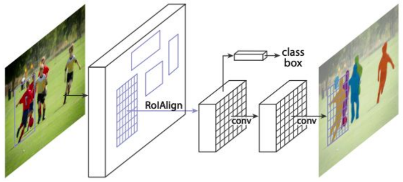
- **Challenges:**
  - Domain gaps between COCO and Seedlings
    - High-quality still images vs low-quality video frames captured by wearable cameras.
    - Small objects in the scene are challenging to detect.
  - Limited object vocabulary (80 classes)

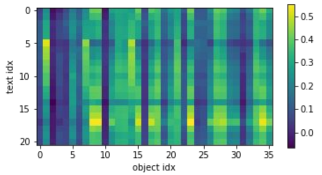
Object Detection: Results (left: 3rd right: 1st person view)

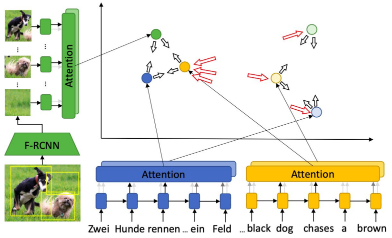


Multimodal association in SeedlingS Corpus

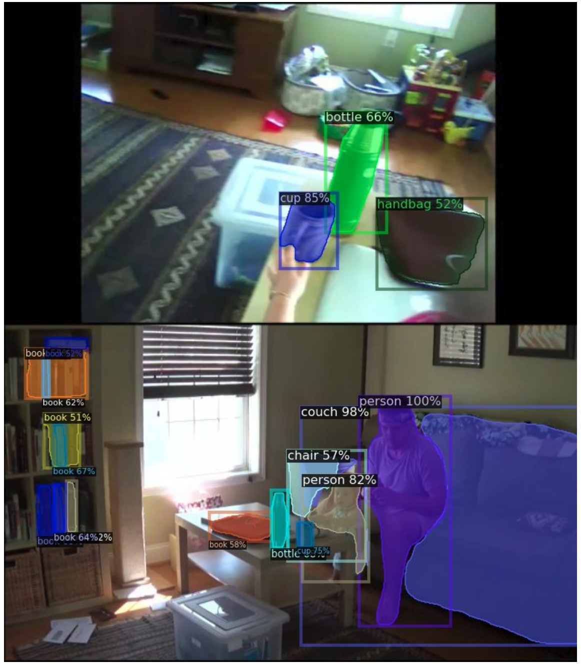
- Goal:
  - Associating parents’ speech with visual object in the video of Seedlings’ Corpus
- Approaches:
  - Alignment with Object detection
    - Top-1 *sim(w2v of object name, text token w2v)*
    - Limitation:
      - small pool of object class names (MS-COCO: 80 classes)
      - Noisy, irrelevant objects
  - Alignment with Multilingual Multimodal Embeddings
    - *sim(visual object, text token)*
- Current Problem/Challenges:
  - Domain gap between written (e.g. caption) and spoken language (e.g. baby’s talk/speech)
  - Lack of reliable annotation.







Example1: Dad’s Coffee Mug from this morning

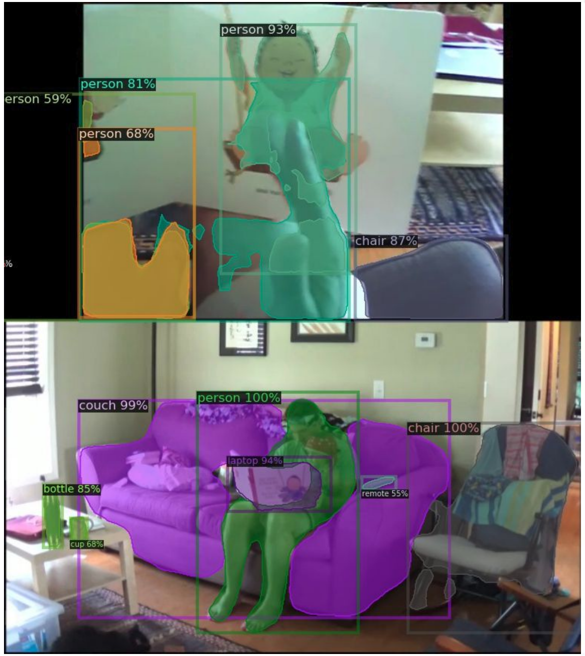


Object detection results

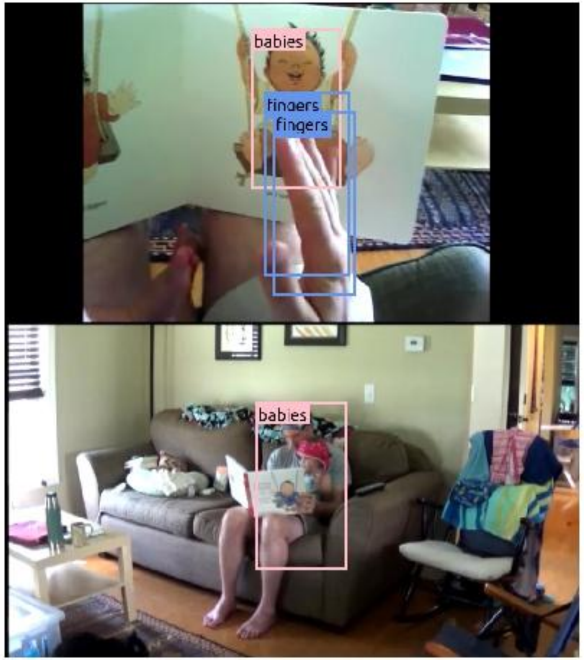
Visual-Speech alignment

Asr transcription: That's Dad's coffee mug from this morning.

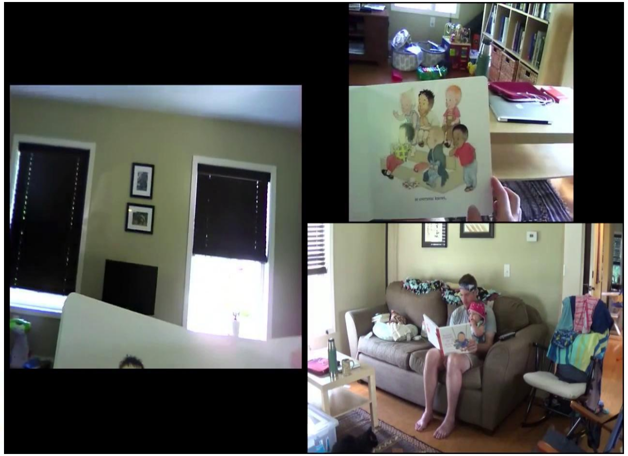
Babies had ten little fingers and ten little toes..



Object detection results



Visual-Speech alignment



And both of these babies as everyone knows.

Yeah.

Had ten little fingers and ten little toes.

Next Steps

Application of multi-modal language learning to human language acquisition data

Refinement of unsupervised visual structure induction, etc.

Thank You!
Questions?