Learning Syntactic Structures from Visually Grounded Text and Speech

Freda Shi

Toyota Technological Institute at Chicago & the University of Waterloo

freda@ttic.edu/fhs@uwaterloo.ca

Oct. 24, 2023

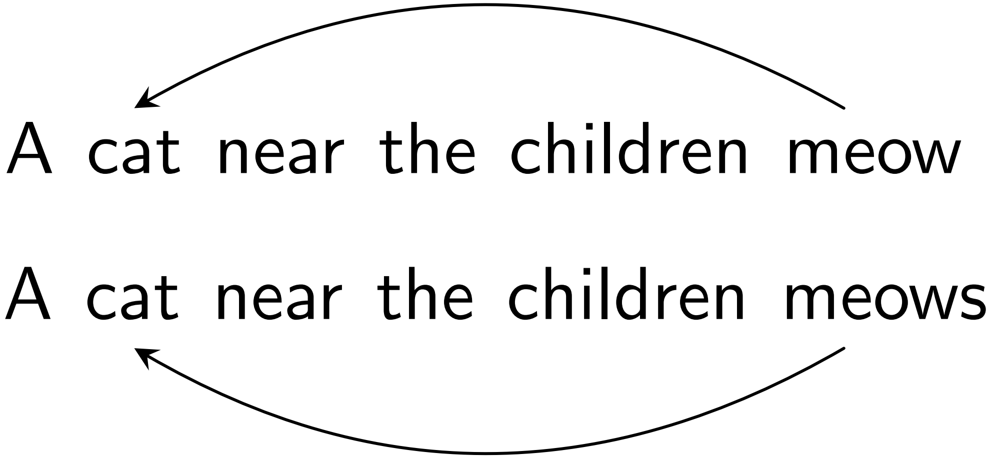
@the University of Michigan

A Minimal Pair

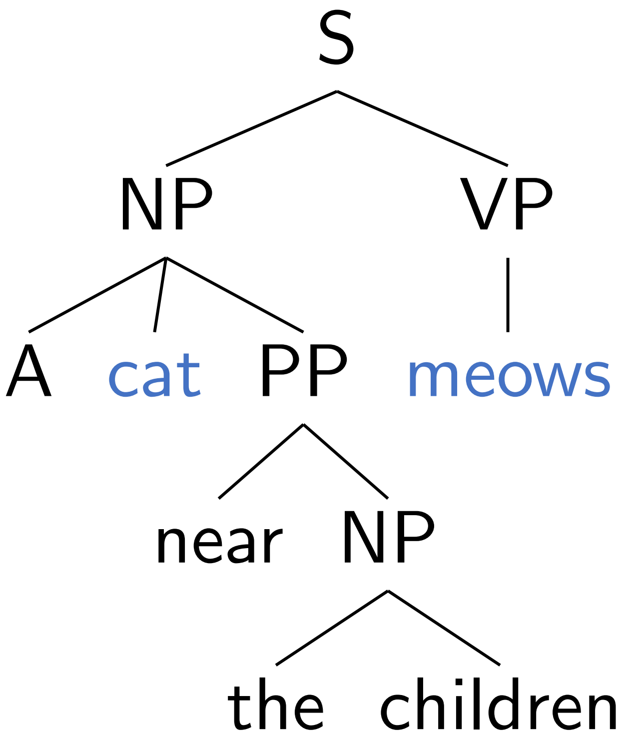
A cat near the children meow (*)

A cat near the children meows

Syntactic Structures



Dependency



Constituency

Syntactic Structures

- Languages are highly structured
- The explicit structures are almost never given (to native speakers)

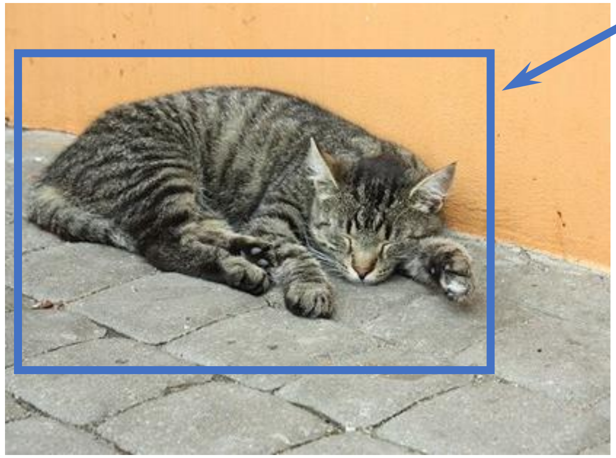
In the real world, we learn and use language in grounded settings

A cat is standing on the lawn.

How did we learn our first language?

A cat is standing on the lawn.

A cat, as a whole,
means something concrete



A cat is sleeping
There is a cat sleeping on the ground

How did we learn our first language?

Our Observation

A cat, as a whole,
means something concrete

Definition of *Constituent*

A cat, as a whole,
functions as a single unit in sentences

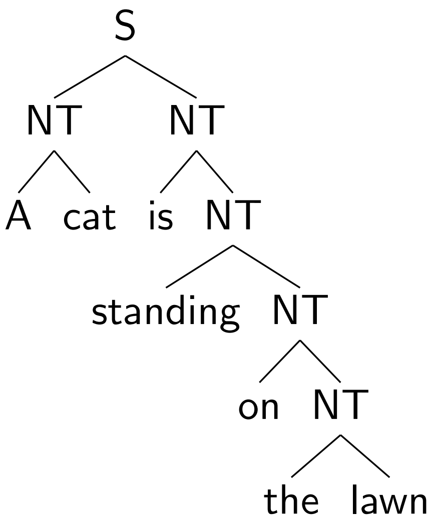
Our Hypothesis

More visually concrete word spans are more likely to be constituents

Visually Grounded Grammar Induction

- Input: captioned images
- Output: linguistically plausible structure for captions

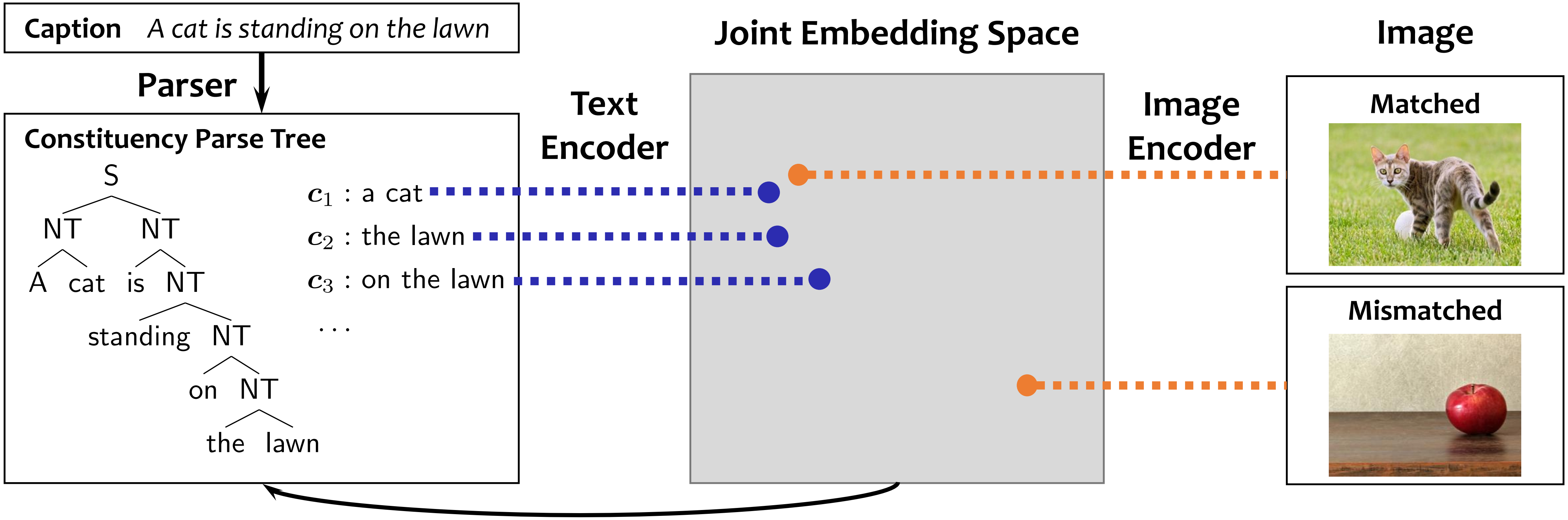
A cat is standing on the lawn



[**Shi***, Mao*, Gimpel, Livescu. Visually grounded neural syntax acquisition. ACL 2019]

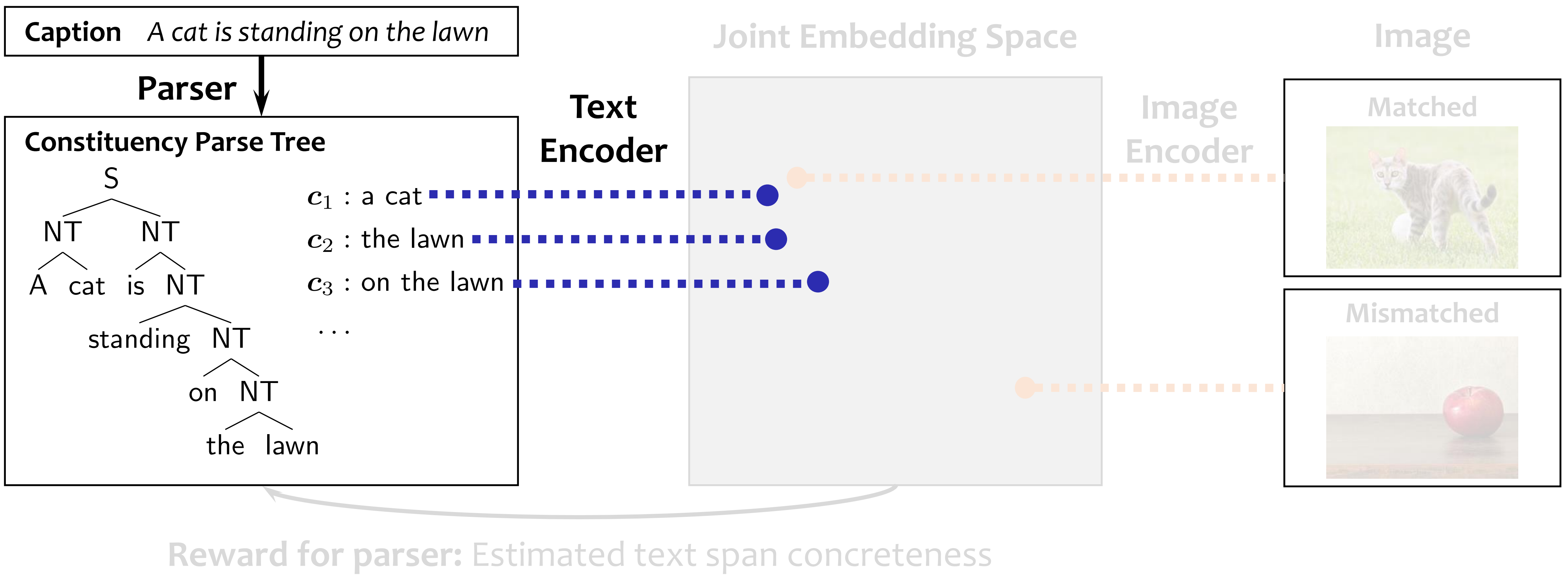
The Visually Grounded Neural Syntax Learner (VG-NSL)

Hypothesis: more visually concrete word spans are more likely to be constituents

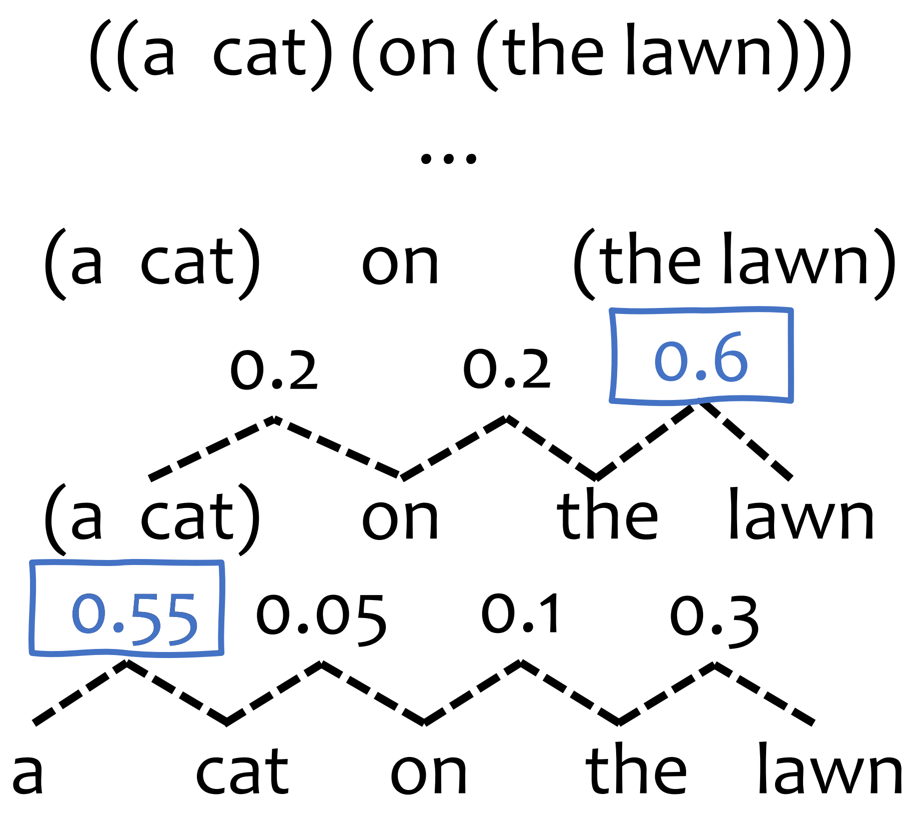


Reward for parser: Estimated text span concreteness

VG-NSL: Text Parser and Encoder



VG-NSL: Text Parser and Encoder



Repeat the score-sample-combine process for $n-1$ times

$\mathbf{\Theta}$: Parameters for structure

$\mathbf{V}$: Parameters for word meanings

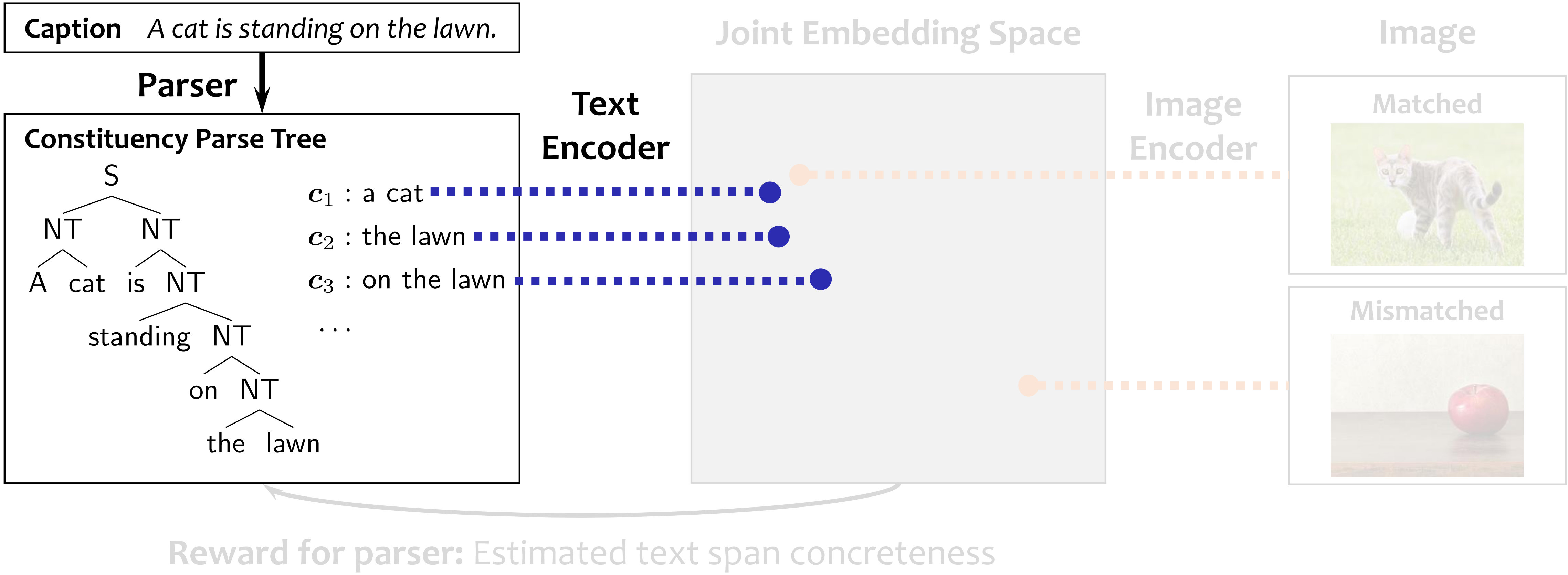
$$\mathbf{v}_{\text{a cat}} = \frac{\mathbf{v}_{\text{a}} + \mathbf{v}_{\text{cat}}}{||\mathbf{v}_{\text{a}} + \mathbf{v}_{\text{cat}}||_2}$$

Compute score

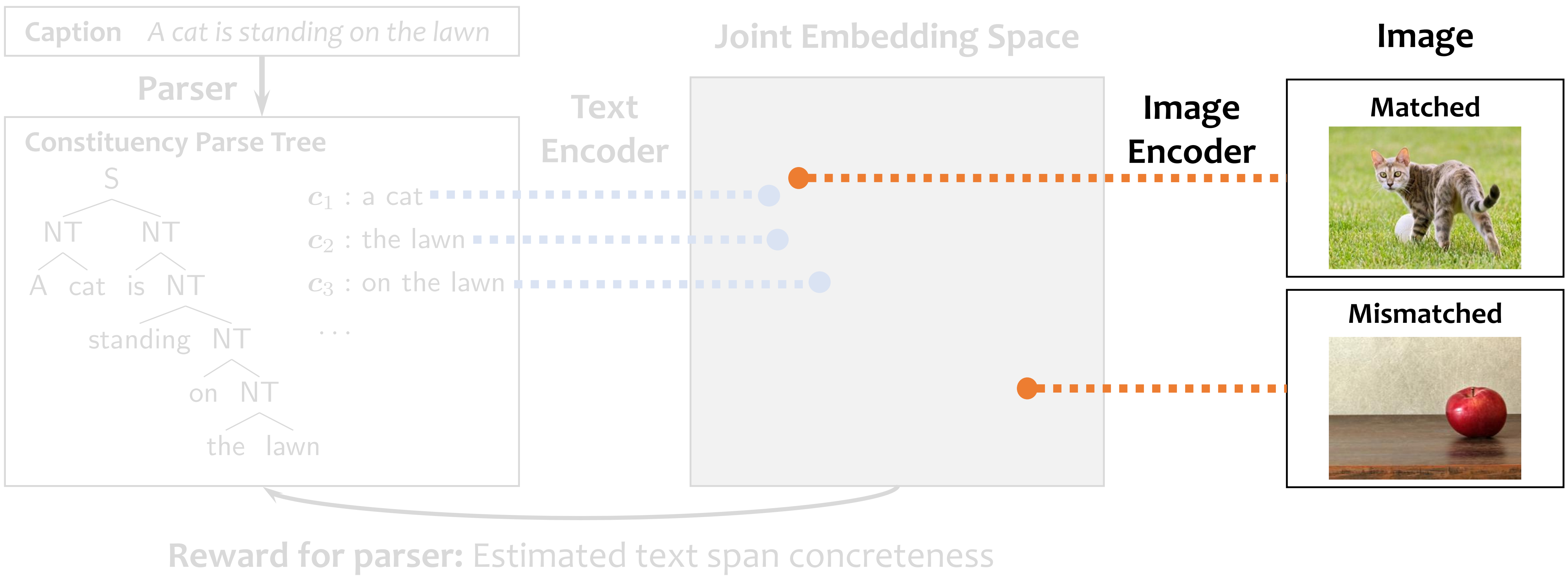
$$\text{FFN}_{\mathbf{\Theta}}\left(\begin{pmatrix} \mathbf{v}_{\text{the}} \\ \mathbf{v}_{\text{lawn}} \end{pmatrix}\right) = 0.3$$

VG-NSL: Text Parser and Encoder

$\boldsymbol{\Theta}$: Parameters for structure $\qquad$ $\mathbf{V}$: Parameters for word meanings



VG-NSL: Image Encoder



VG-NSL: Image Encoder

Frozen
ResNet152
(He et al., 2015)

Linear Projection

Trainable Parameter

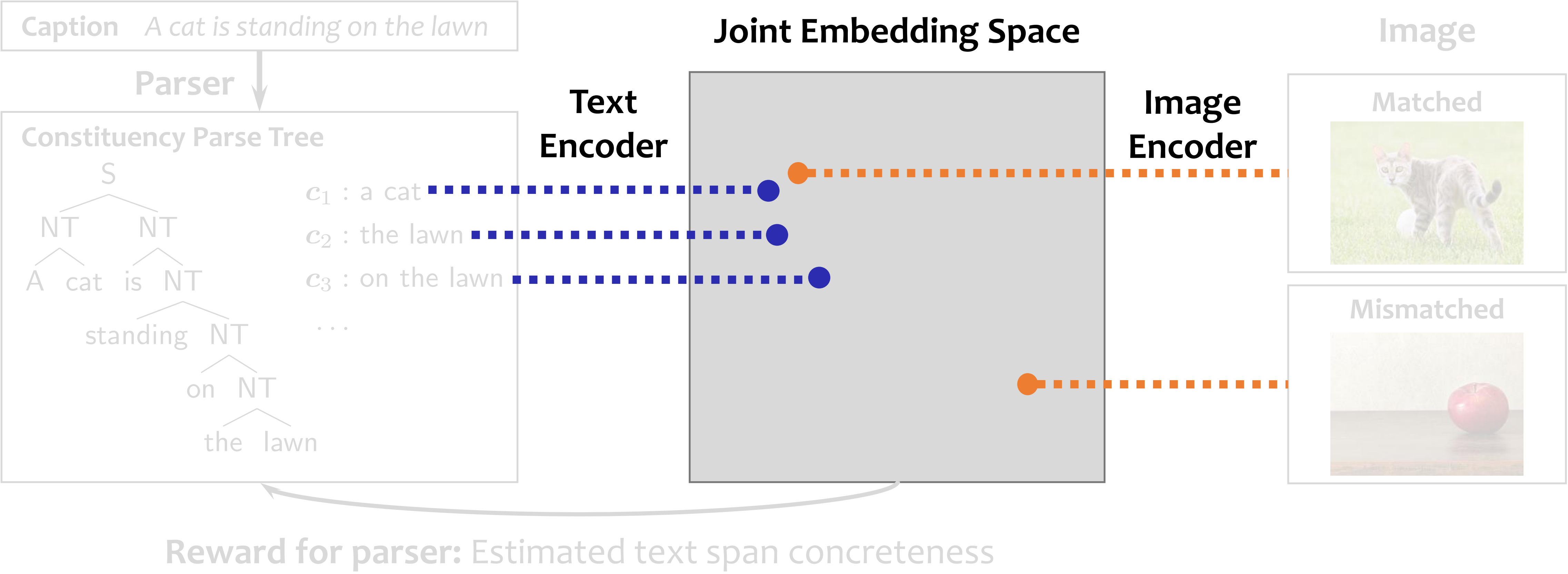
$$\mathbf{u} \rightarrow \mathbf{\Phi u}$$

ResNet Image Representation

Linear Projection

VG-NSL: Joint Embedding Space

Model parameters: $\boldsymbol{\Theta}$--text structure; Φ, $\mathbf{V}$--visual/textual semantics



VG-NSL: Joint Embedding Space

- **Key idea:** high similarity for matched image-constituent pairs,
  low similarity for mismatched pairs
- **Approach:** minimize the hinge-based triplet loss (Kiros et al., 2015)

$$\mathcal{L}\left(i_{\mathbf{\Phi}}, c_{\mathbf{V}}\right) = \sum_{(i',c') \neq (i,c)} \left[\operatorname{sim}\left(i_{\mathbf{\Phi}}, c'_{\mathbf{V}}\right) - \operatorname{sim}\left(i_{\mathbf{\Phi}}, c_{\mathbf{V}}\right) + \delta\right]_+ + \left[\operatorname{sim}\left(i'_{\mathbf{\Phi}}, c_{\mathbf{V}}\right) - \operatorname{sim}\left(i_{\mathbf{\Phi}}, c_{\mathbf{V}}\right) + \delta\right]_+$$


an apple


a cat


a cat


a cat

$\operatorname{sim}(\cdot,\cdot) = \cos(\cdot,\cdot)$ $\quad [\cdot]_+ = \max(0,\cdot)$ $\quad \delta$: margin score

VG-NSL: Quantify Visual Concreteness

- **Joint embedding space:** High similarity $\leftrightarrow$ stronger correspondence

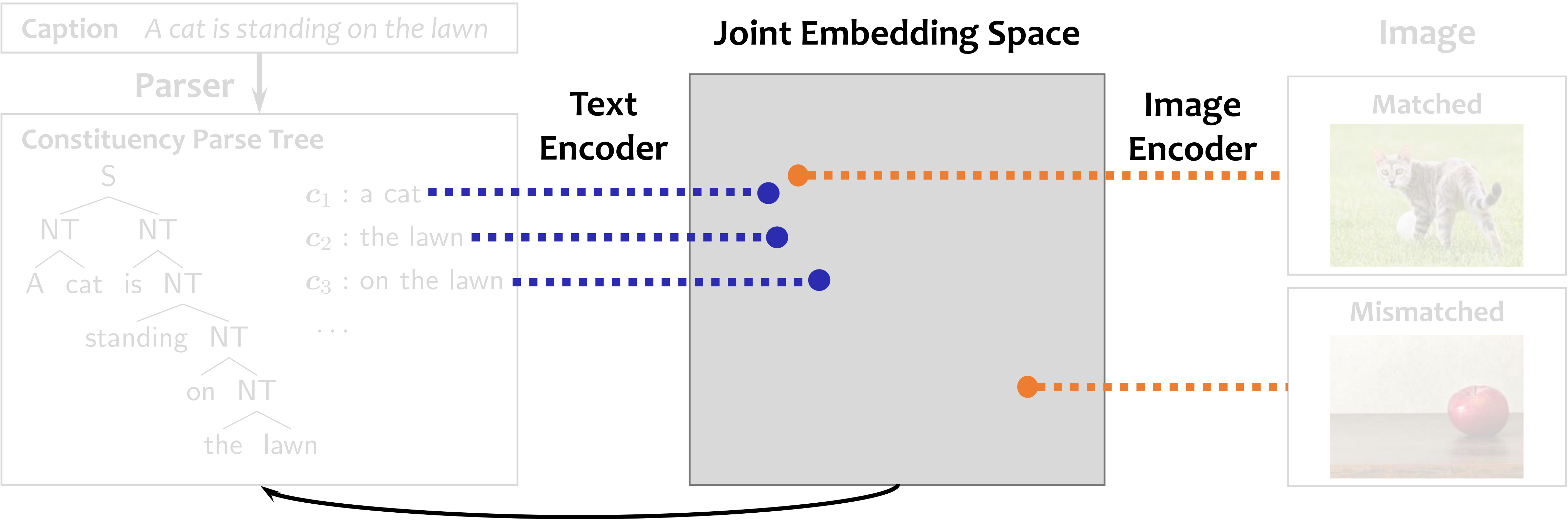
image i	**another image** i'	**candidate constituents**
		*a cat* *on the*

$$\ell(c; i, i') = \mathrm{sim}\left(i'_{\mathbf{\Phi}}, c_{\mathbf{V}}\right) - \mathrm{sim}\left(i_{\mathbf{\Phi}}, c_{\mathbf{V}}\right)$$

-0.8	$\mathrm{sim}\left(i'_{\mathbf{\Phi}}, \textit{a cat}\right) = 0.1$	$\mathrm{sim}\left(i_{\mathbf{\Phi}}, \textit{a cat}\right) = 0.9$
0	$\mathrm{sim}\left(i'_{\mathbf{\Phi}}, \textit{on the}\right) = 0.2$	$\mathrm{sim}\left(i_{\mathbf{\Phi}}, \textit{on the}\right) = 0.2$

- Idea: smaller $\ell(c) \leftrightarrow c$ is more concrete

VG-NSL: Training the Parser



Reward for parser: Estimated text span concreteness

VG-NSL: Training the Parser

- $\ell(c; i, i') = \mathrm{sim}(i'_{\mathbf{\Phi}}, c_{\mathbf{V}}) - \mathrm{sim}(i_{\mathbf{\Phi}}, c_{\mathbf{V}})$ quantifies visual abstractness of word spans, and we can define concreteness similarly

$$\mathrm{concreteness}(c; i, i') = \left[\mathrm{sim}(i_{\mathbf{\Phi}}, c_{\mathbf{V}}) - \mathrm{sim}(i'_{\mathbf{\Phi}}, c_{\mathbf{V}}) + \delta\right]_+$$

- REINFORCE (Williams, 1992)

$$\underbrace{\boxed{\mathbf{\Theta}}}_{\text{parser parameters}} \leftarrow \mathbf{\Theta} + \underbrace{\boxed{\eta}}_{\text{learning rate}} \nabla_{\mathbf{\Theta}} \sum_{i, i', c} p_{\mathbf{\Theta}}(c) \underbrace{\boxed{\mathrm{concreteness}(c; i, i')}}_{\text{reward}}$$

- After training, the parser can parse sentences without images

VG-NSL: Head-Initiality as Abstract-Initiality

((A cat) on) (the lawn)   (A cat) (on (the lawn))

Fact #1: *On* is the head of *on the lawn*

Fact #2: English is strongly head-initial
Many other Indo-European languages are head-initial as well

Fact #3: In visually grounded settings, most abstract words are function words (e.g., prepositions, determiners, complementizers)

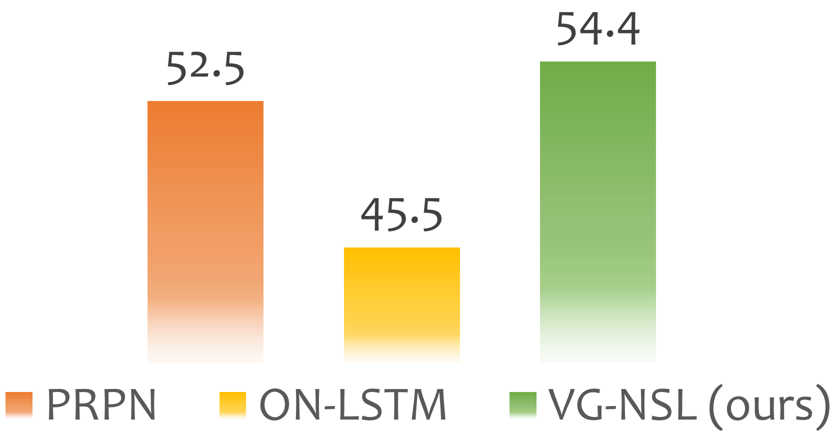
Empirical Solution (mimic the head-initial property with abstractness):
  Discourage abstract words from combining to the front

$$\mathrm{reward}(c) = \frac{\mathrm{concreteness}(c)}{\lambda \cdot \mathrm{abstractness}(c_{right}) + 1} \quad (\lambda > 0)$$

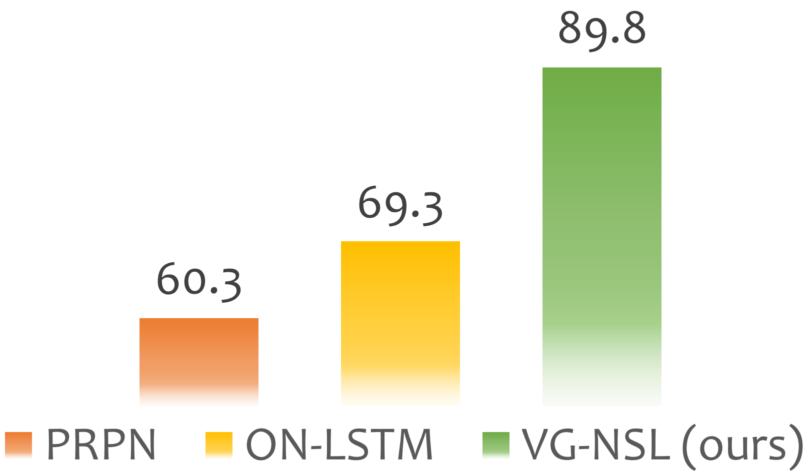
VG-NSL: English Results

- Text-only models: PRPN (Shen et al., 2018), ON-LSTM (Shen et al., 2019)
- Evaluated on MSCOCO (Lin et al., 2014)

F_1 score (↑)

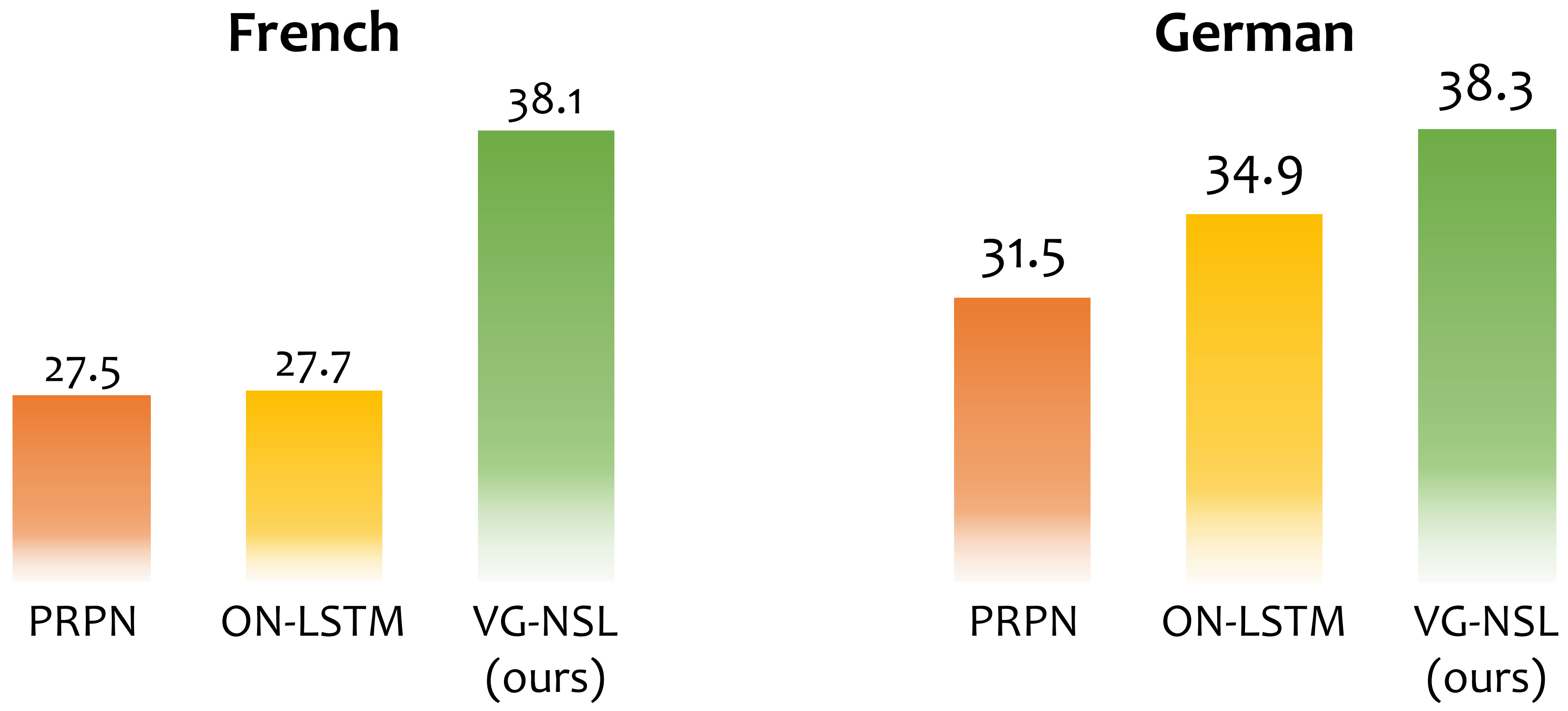


Self F_1 score across 5 random seeds (↑)



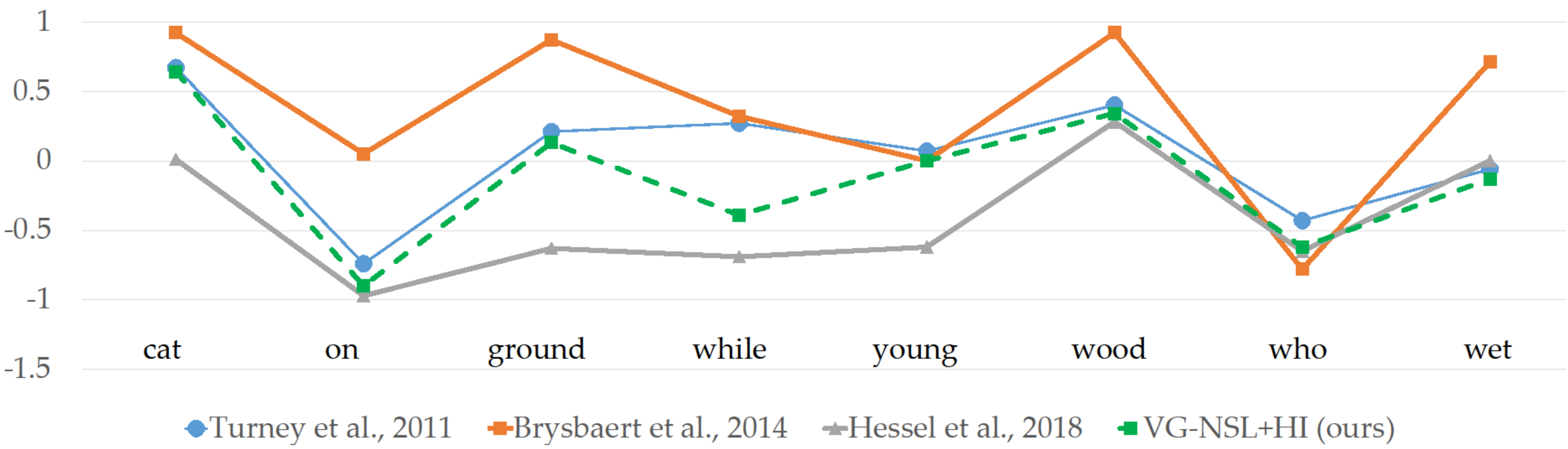
VG-NSL: Multilingual Results

- Extension to multiple languages, evaluated on Multi30K (Elliott et al., 2017)



VG-NSL: Estimated Concreteness

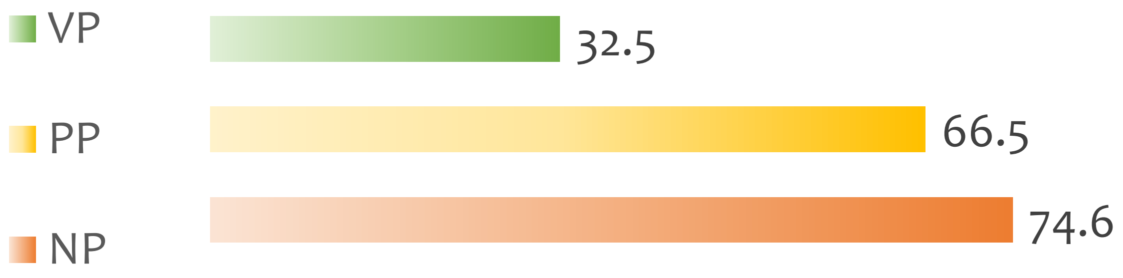
- Normalized concreteness ∈ [0, 1]



VG-NSL: Discussion

- VG-NSL's concreteness-based bottom-up parser is good at capturing NPs and PPs, but less good at capturing VPs

Recall per Category (↑)



- Follow-up work: more sophisticated inductive biases (e.g., PCFG) and other modalities (e.g., video)

Other Work on Grounded Grammar Induction

What is Learned in Visually Grounded Neural Syntax Acquisition

Noriyuki Kojima, Hadar Averbuch-Elor,
Alexander Rush and **Yoav Artzi**
Department of Computer Science and Cornell Tech, Cornell University
{nk654,he93,arush}@cornell.edu
{yoav}@cs.cornell.edu

Visually Grounded Compound PCFGs

Yanpeng Zhao[ε]
[ε]ILCC, University of Edinburgh
yanp.zhao@ed.ac.uk

Ivan Titov[εæ]
[æ]ILLC, University of Amsterdam
ititov@inf.ed.ac.uk

Video-aided Unsupervised Grammar Induction

Songyang Zhang[1]*, **Linfeng Song**[2], **Lifeng Jin**[2], **Kun Xu**[2], **Dong Yu**[2] and **Jiebo Luo**[1]
[1]University of Rochester, Rochester, NY, USA
szhang83@ur.rochester.edu, jluo@cs.rochester.edu
[2]Tencent AI Lab, Bellevue, WA, USA
{lfsong,lifengjin,kxkunxu,dyu}@tencent.com

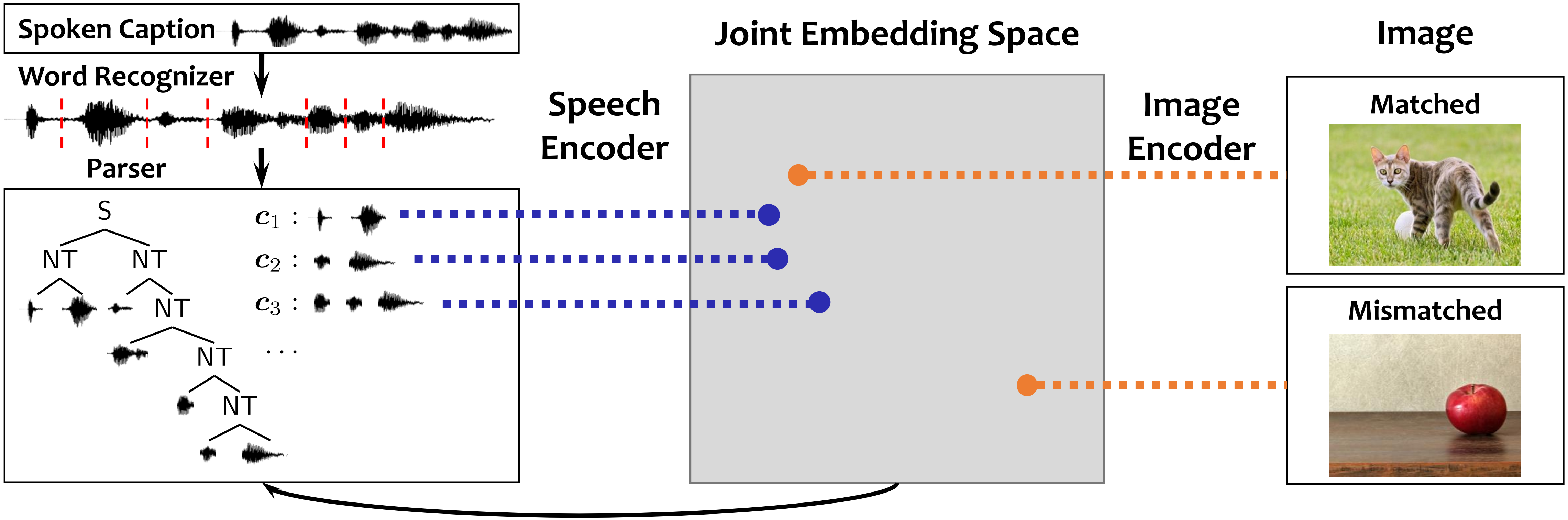
UNSUPERVISED VISION-LANGUAGE GRAMMAR INDUCTION WITH SHARED STRUCTURE MODELING

Bo Wan[1], **Wenjuan Han**[2]*, **Zilong Zheng**[2], **Tinne Tuytelaars**[1]
1. Department of Electrical Engineering, KU Leuven;
2. Beijing Institute for General Artificial Intelligence, Beijing, China
{bwan;Tinne.Tuytelaars}@esat.kuleuven.be;
{hanwenjuan;zlzheng}@bigai.ai

VG-NSL: Discussion

- Motivation of grammar induction/unsupervised parsing
  - Understanding quantitatively how much syntax is encoded in data
  - Arguing for or against the poverty of the stimulus (Chomsky, 1980)
  - Byproduct: methods derived could benefit other tasks
- Modeling human language acquisition
  - Pretrained **text** models are less desirable due to corpus-size mismatch
  - Pretrained speech models are okay in terms of developmental plausibility
    - HuBERT-960hr gives reasonable performance
    - Even the 60K-hour Libri-light data is acceptable: 60,000/24/365 = 6yrs
  - Humans learn languages in grounded settings
    - Much of humans' early exposure to language is in speech

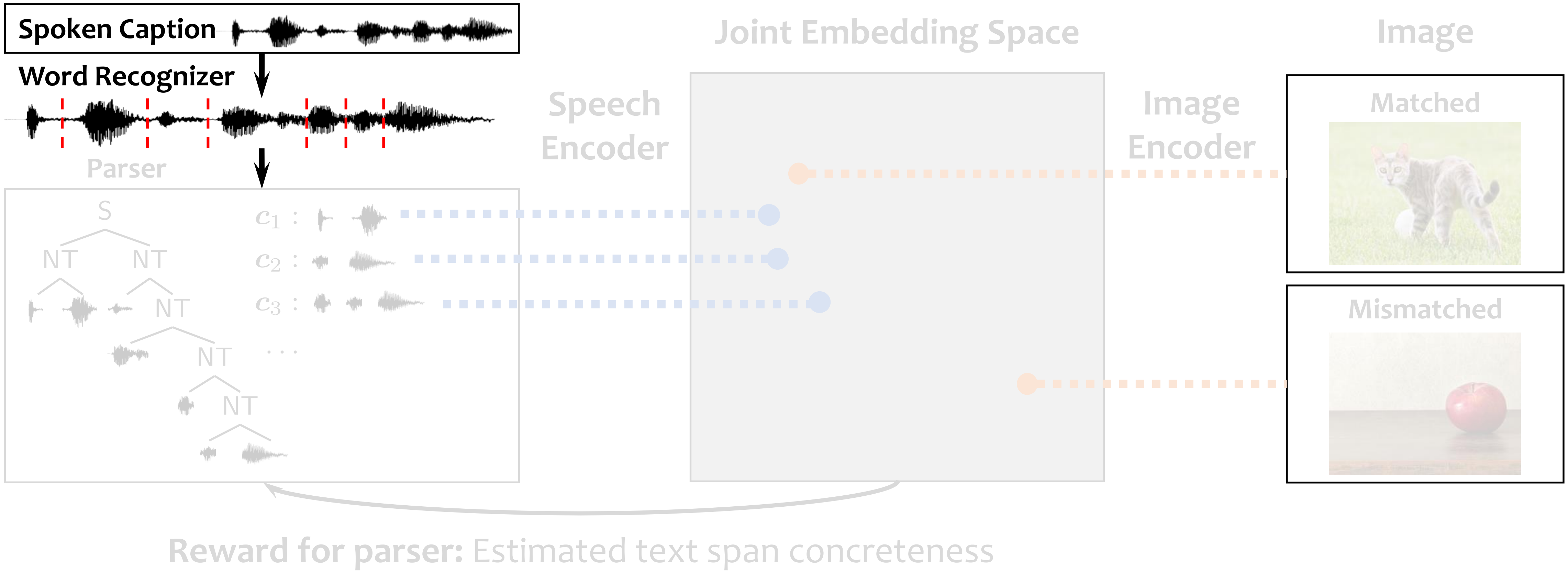
The Audio-Visual Syntax Learner (AV-NSL)



Reward for parser: Estimated text span concreteness

[Lai*, **Shi***, Peng*, et al. Audio-Visual Neural Syntax Acquisition. ASRU 2019]

AV-NSL: Word Recognition/Segmentation

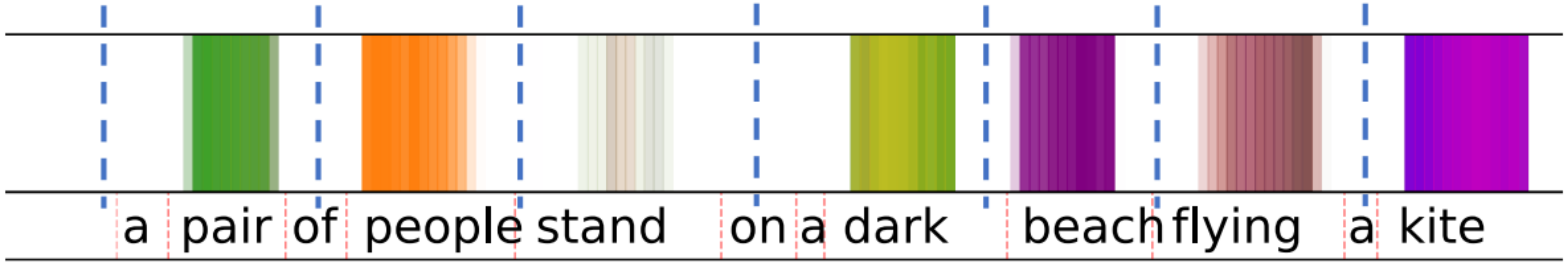


AV-NSL: Word Recognition/Segmentation

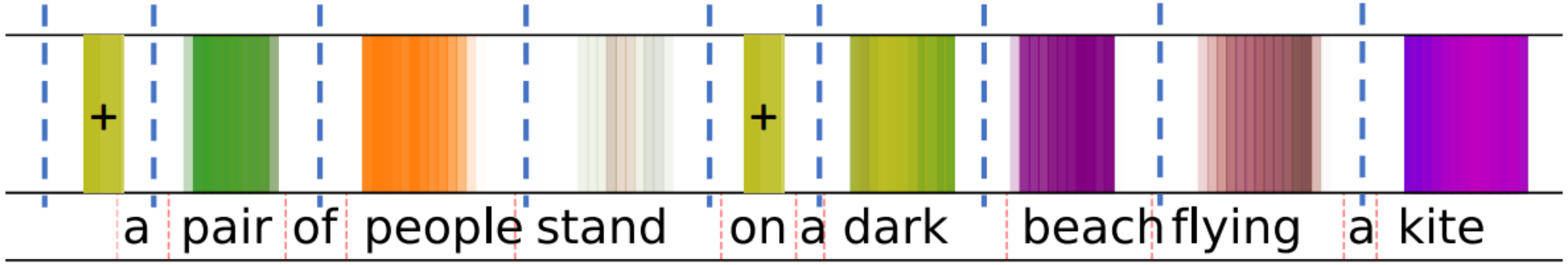
- How should we obtain word segments from a spoken utterance?
- Segmentation with forced alignment: Template-based matching between text and speech (e.g., MFA; McAuliffe et al., 2007)
- Humans learn to listen and speak before learning to read and write
  - Unsupervised word recognition/segmentation is desirable

AV-NSL: Word Recognition/Segmentation

- Word segmentation emerges from VG-HuBERT [CLS] token's attention weights (Peng and Harwath, 2022)



- Insert tokens in long gaps (threshold tuned w/o supervision)



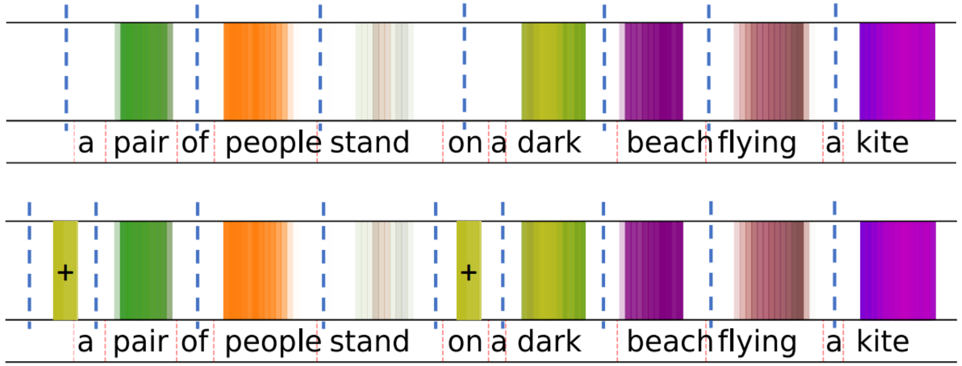
AV-NSL: Word Recognition/Segmentation

- Word segmentation with minimum Bayes' risk (MBR) decoding
- Collect multiple word segmentation proposals with different hyperparameters (e.g., threshold for inserting new segment)

$$\hat{\mathcal{S}} = \arg\max_{\mathcal{S}} \boxed{\sum_{\mathcal{S}'} \boxed{F_1(\mathcal{S}, \mathcal{S}')}}$$

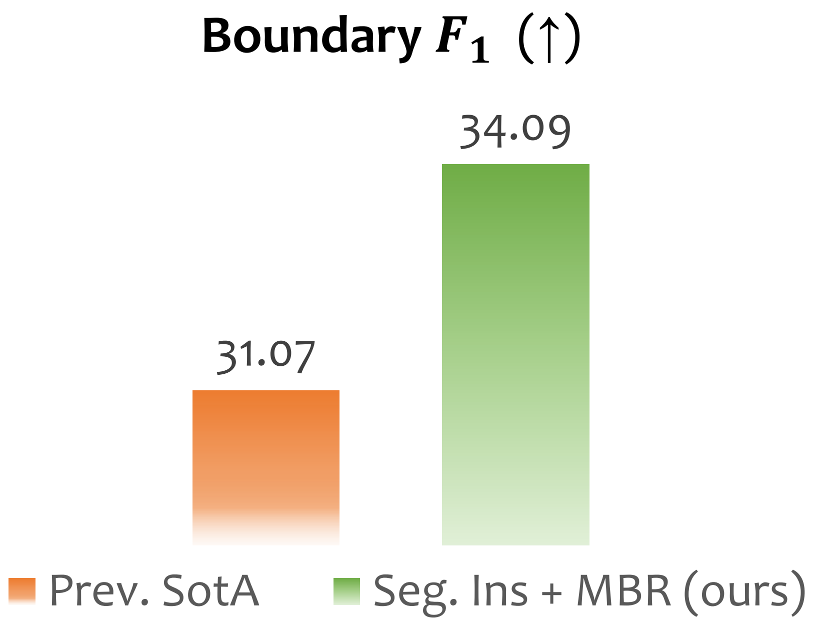
negative Bayes' risk

boundary precision/recall within $\pm 20ms$



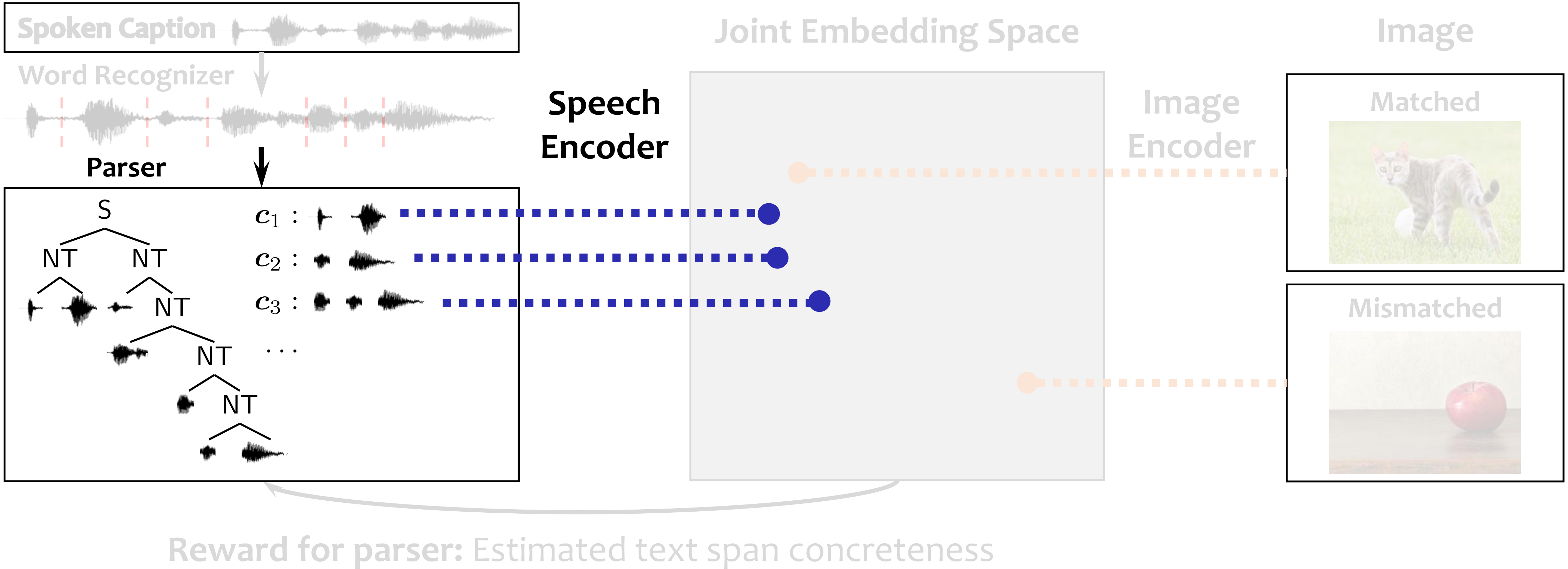
AV-NSL: Word Recognition/Segmentation

- Hyperparameters searched:
  - Threshold to be considered as a long-enough gap (for segment insertion)
  - Threshold to filter out frames that receive less attention
  - VG-HuBERT layer index



AV-NSL: Speech Span Encoders

VG-HuBERT (Peng and Harwath, 2022) as the speech span encoder



AV-NSL: Evaluation

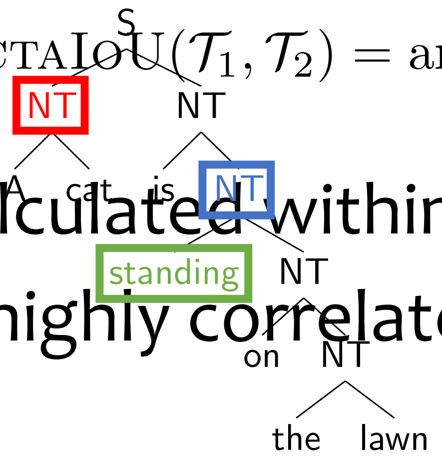
- Text-based segmentation: F_1 score (same as text parsing)
- What if the word segmentation doesn't align with the text?
- Prior work (Roark et al., 2006): project speech to the text domain

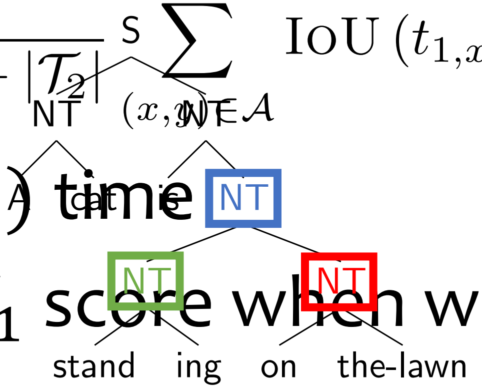
- Our proposal: use a structured alignment—based intersection-over-union ratio to measure the similarity between speech constituency parse trees
- IoU between two spans: $\mathrm{IoU}(I_1, I_2) = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}$

AV-NSL: Evaluation with Structured Average IoU

- Align two constituency parse trees over the same spoken utterance
  - Each node aligns with at most one node in the other tree
  - If node a (in tree 1) and b (in tree 2) are aligned
    - Any descendant of a may align with a descendant of b or remain unaligned, and vice versa
    - Any ancestor of a may align with an ancestor of b or remain unaligned, and vice versa

$$\textsc{StructaIoU}(\mathcal{T}_1, \mathcal{T}_2) = \arg\max_{\mathcal{A}} \frac{1}{|\mathcal{T}_1| + |\mathcal{T}_2|} \sum_{(x,y)\in\mathcal{A}} \textsc{IoU}\left(t_{1,x}, t_{2,y}\right)$$





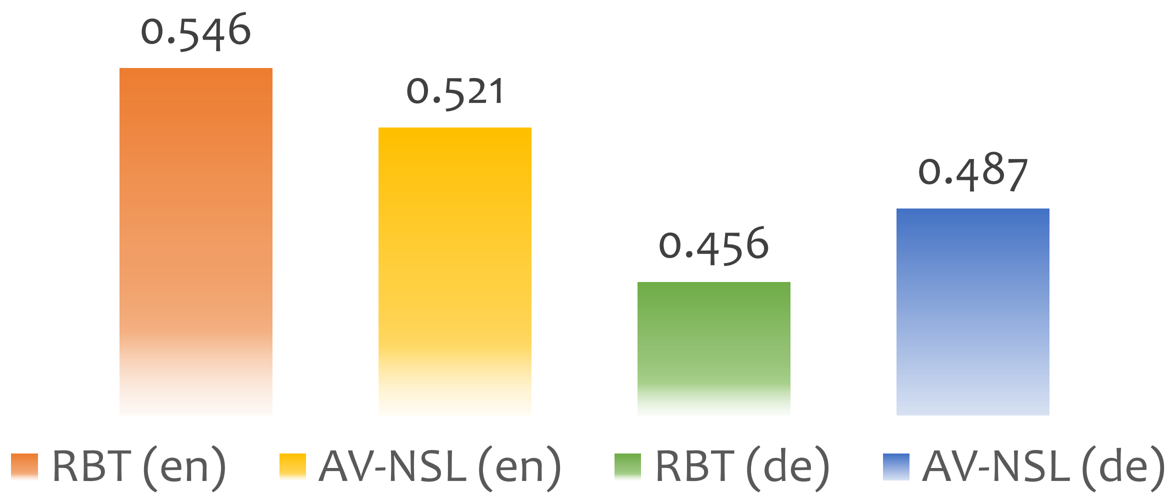
- This can be calculated within $\mathcal{O}(n^2m^2)$ time
- StructaIoU is highly correlated with F_1 score when word segmentation is present

[**Shi**, Gimpel, Livescu. Structured Tree Alignment for Evaluation of Constituency Parsing. Work in Progress]

AV-NSL: Results

- Right-branching trees serve as a strong baseline for European languages
- There is still a gap between the current state and a decent grammar induction model from visually grounded speech

StructaIoU score (w/o gold word segmentation ↑)

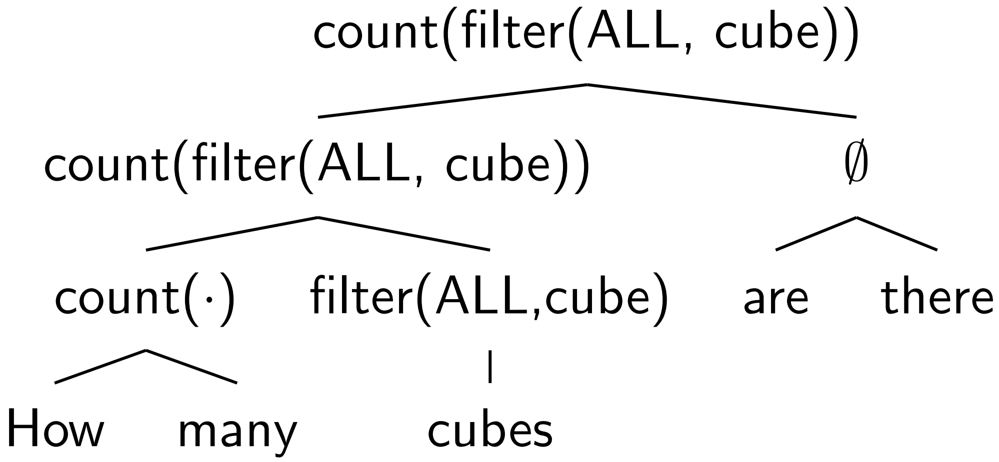


Joint Syntax and Semantics Induction

- Combinatory categorial grammar induction in visually grounded settings

CLEVR (Johnson et al., 2017)

Question: How many cubes are there?
Answer: 4



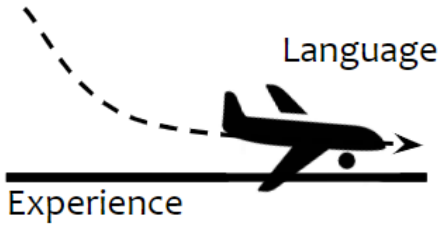
Question answering accuracy (↑) on program-depth generalization:

81.6 (prior SotA) → **98.5**

[Mao, **Shi**, Wu, Levy, Tenenbaum. Grammar-Based Grounded Lexicon Learning. NeurIPS 2021]

Looking ahead…

- Language is never text in isolation
  - Computational linguistics research should benefit more from state-of-the-art machine learning techniques, including (and especially) computer vision, speech, and robotics
- Grounding in NLP does not necessarily mean vision-text models---other grounding forms include but are not limited to
  - Execution results of programs, semantic parses of natural language
  - Sentences with shared semantics but in different languages
  - Knowledge bases
  - A metaphor for *grounding* ☺

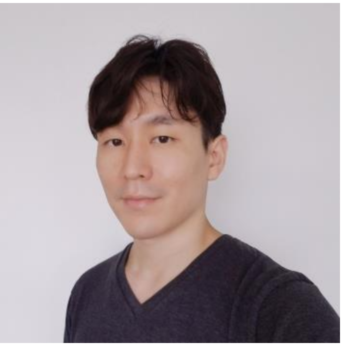


Thanks!

Kevin Gimpel

James Glass

David Harwath



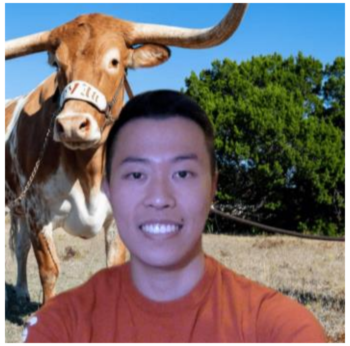
Yoon Kim

Cheng-I Jeff Lai

Roger Levy

Karen Livescu

Jiayuan Mao



Puyuan Peng

Josh Tenenbaum

Jiajun Wu

& other collaborators