Language Grounding

CPSC 677 Presentation

Borui Wang

October 13, 2020

1

Introduction

What is Language Grounding?

Language grounding is to connect and embed linguistic symbols to **perceptual experiences** and **actions** in the real world, such that an AI system can learn to better understand and use natural languages of human.

"flower"
"花"
"fiore"
"꽃"
"flor"

grounds into

Why Language Grounding?

- To model the way language conveys meaning, traditional approaches in natural language processing consider language as a purely symbolic system based on words and syntactic rules (Chomsky, 1980; Burgess and Lund, 1997). The meaning of words and sentences is usually represented only in terms of other words or textual symbols.

- However, meaning does not arise from the statistical distribution of words, but from their use by people to communicate. Many of the assumptions and understandings on which communication relies lie outside of text. We must consider what is missing from models trained solely on text corpora, even when those corpora are meticulously annotated or Internet-scale.

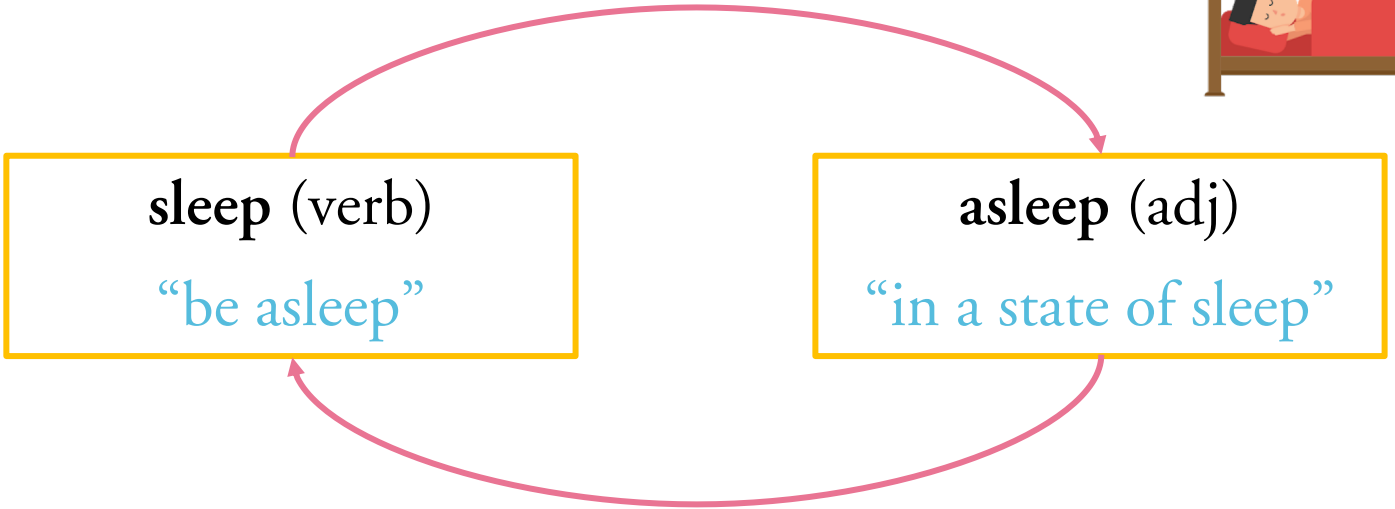
Why Language Grounding?

- Language understanding research is held back by a failure to relate language to the physical world it describes and to the social interactions it facilitates. Despite the incredible effectiveness of language processing models to tackle tasks after being trained on text alone, successful linguistic communication relies on a shared experience of the world. It is this shared experience that makes utterances meaningful (Bisk et al., 2020). Therefore, truly understanding the meaning of language requires grounding semantics in perception and action in the world (Mooney 2013).

- The present success of representation learning approaches trained on large, text-only corpora requires the parallel tradition of research on the broader physical and social context of language to address the deeper questions of communication. (Bisk et al., 2020)

The Problem of Circular Definition

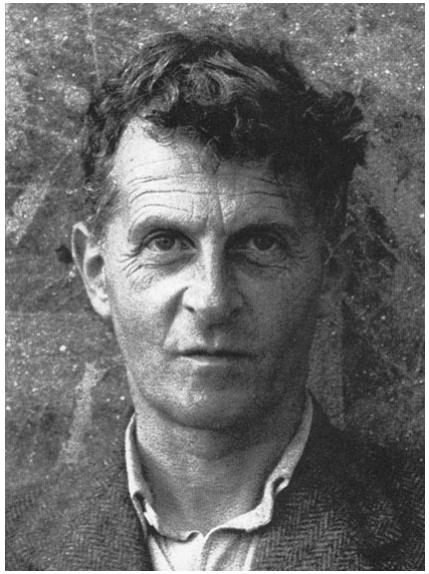
- One of the key problems of the text-only approach to NLP is the problem of circular definition, which is illustrated in the following example (Mooney 2013) from WordNet:



The Problem of Bias between Text and Reality

- Another key problem of the text-only approach to NLP is that the frequency at which objects, relations, or events occur in natural language is significantly different from their real-world frequency (e.g., in texts, people are murdered four times more than they breathe). Thus, leveraging visual resources, in addition to textual resources, is a promising way to acquire commonsense knowledge (Lin and Parikh, 2015; Yatskar et al., 2016) and to cope with this bias between text and reality. (Border et al., 2019)

Historical Roots of Language Grounding



‘Meaning as Use’ & ‘Language Games’

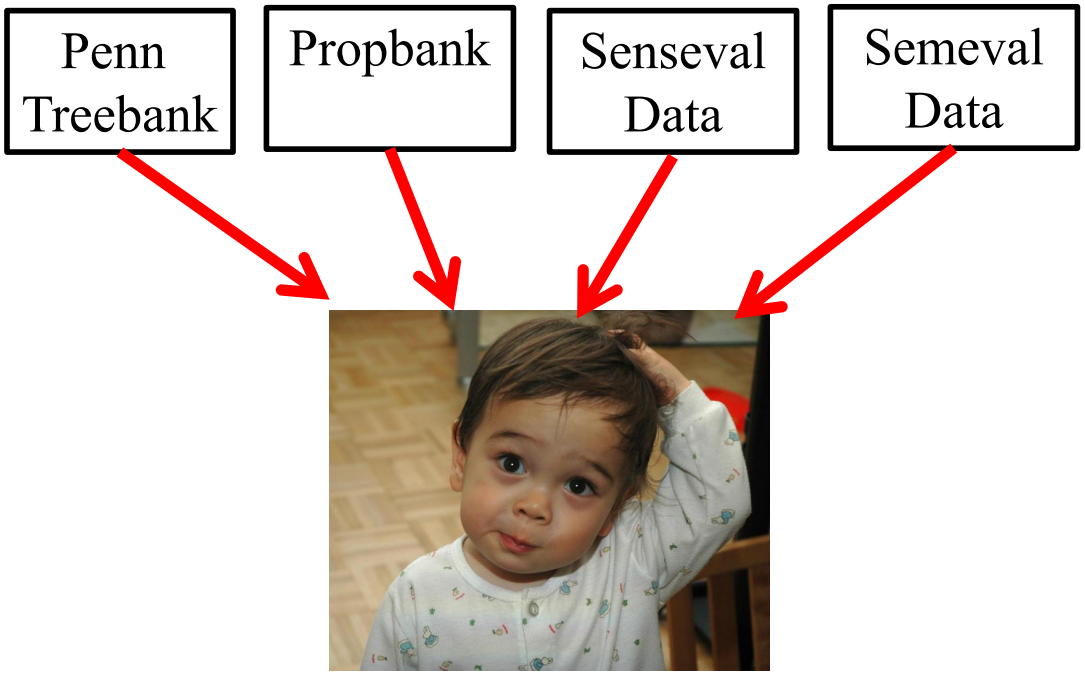
Ludwig Wittgenstein (1953)

‘Symbol Grounding’

Stevan Harnad (1990)

Evidence from Language Learning of Human Children

- Children do not learn language from supervised data.



(Mooney, 2013)

Evidence from Language Learning of Human Children

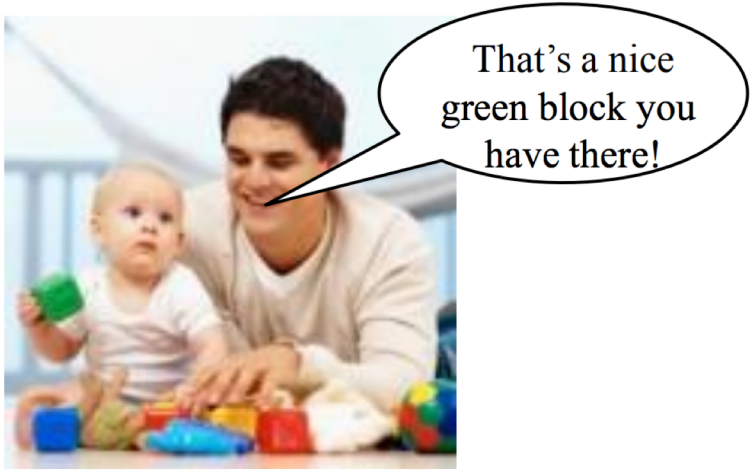
- Children do not learn language from unsupervised raw text.

Unsupervised language learning is difficult and not an adequate solution since much of the requisite semantic information is not in the linguistic signal.

(Mooney, 2013)

Evidence from Language Learning of Human Children

- Children learn language from **perceptual context**.



The natural way to learn language is to perceive language in the context of its use in the physical and social world. This requires inferring the meaning of utterances from their perceptual context.

(Mooney, 2013)

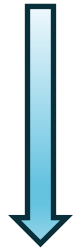
Levels of Language Grounding – World Scope

- In the field of NLP, we can classify research works into 5 levels of **World Scope**, according to their levels of language grounding (Bisk et al., 2020).

- These 5 levels of World Scope is defined as:

  - WS1. Corpus Level (our past)
  - WS2. Internet Level (most of current NLP research works)
  - WS3. **Perception Level** (multimodal NLP)
  - WS4. Embodiment Level
  - WS5. Social Level

Language Grounding



World Scope 3 – The World of Sights and Sounds

- Language learning needs perception, because perception forms the basis for many of our semantic axioms. Learned, physical heuristics, such as the fact that a falling cat will land quietly, are generalized and abstracted into language metaphors like 'as nimble as a cat' (Lakoff, 1980).

- World knowledge forms the basis for how people make **entailment** and **reasoning decisions**, commonly driven by **mental simulation** and **analogy** (Hofstadter and Sander, 2013). Perception is the foremost source of reporting bias. The assumption that we all see and hear the same things informs not just what we name, but what we choose to assume and leave unwritten. Further, there exists strong evidence that children require grounded sensory perception, not just speech, to learn language (Sachs et al., 1981; O'Grady, 2005; Vigliocco et al., 2014).

(Bisk, 2020)

World Scope 3 – The World of Sights and Sounds

- Advances in **computer vision** have enabled building **semantic representations** rich enough to interact with natural language. In the last decade of work descendant from image captioning (Farhadi et al., 2010; Mitchell et al., 2012), a myriad of tasks on visual question answering (Antol et al., 2015; Das et al., 2018; Yagcioglu et al., 2018), natural language and visual reasoning (Suhr et al., 2019b), visual commonsense (Zellers et al., 2019a), and multilingual captioning/translation via video (Wang et al., 2019b) have emerged. These combined text and vision benchmarks are rich enough to train large-scale, multimodal transformers (Li et al., 2019a; Lu et al., 2019; Zhou et al., 2019) without language pretraining (e.g. via conceptual captions (Sharma et al., 2018)) or further broadened to include audio (Tsai et al., 2019). Vision can also help ground speech signals (Srinivasan et al., 2020; Harwath et al., 2019) to facilitate discovery of linguistic concepts (Harwath et al., 2020).

(Bisk, 2020)

World Scope 4 – Embodiment and Action

- In human development, **interactive multimodal sensory experience** forms the basis of action-oriented categories (Thelen and Smith, 1996) as children learn how to manipulate their perception by manipulating their environment. Language grounding enables an agent to connect words to these action-oriented categories for communication (Smith and Gasser, 2005), but requires action to fully discover such connections. Embodiment - situated action taking - is therefore a natural next broader context.
- **Robotics** and embodiment are not available in the same off-the-shelf manner as computer vision models. However, there is rapid progress in simulators and commercial robotics, and as language researchers we should match these advances at every step. As action spaces grow, we can study complex language instructions in simulated homes (Shridhar et al., 2020) or map language to physical robot control (Blukis et al., 2019; Chai et al., 2018).

(Bisk, 2020)

World Scope 5 – The Social World

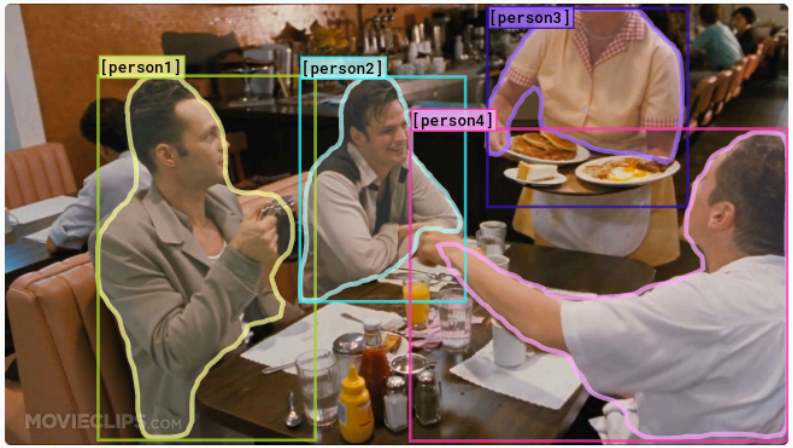
- Interpersonal communication is the foundational use case of natural language (Dunbar, 1993). The physical world gives meaning to metaphors and instructions, but utterances come from a source with a purpose.

- Work in the philosophy of language has long suggested that function is the source of meaning, as famously illustrated through Wittgenstein's "language games" (Wittgenstein, 1953, 1958). In linguistics, the usage-based theory of language acquisition suggests that constructions that are useful are the building blocks for everything else (Langacker, 1987, 1991). The economy of this notion of use has been the subject of much inquiry and debate (Grice, 1975). In recent years, these threads have begun to shed light on what use-cases language presents in both acquisition and its initial origins in our species (Tomasello, 2009; Barsalou, 2008), indicating the fundamental role of the social world.

(Bisk, 2020)

Popular Datasets for Language Grounding

GQA, from Stanford, 2019



VCR, from UW/AI2, 2018

Popular Datasets for Language Grounding

an old locomotive is on a gravelly track.
a very colorful old style train engine on the tracks.
a colorful train engine sits on the tracks.
a train going down a train track by a person.
an engine of a train sitting on a track near a man.

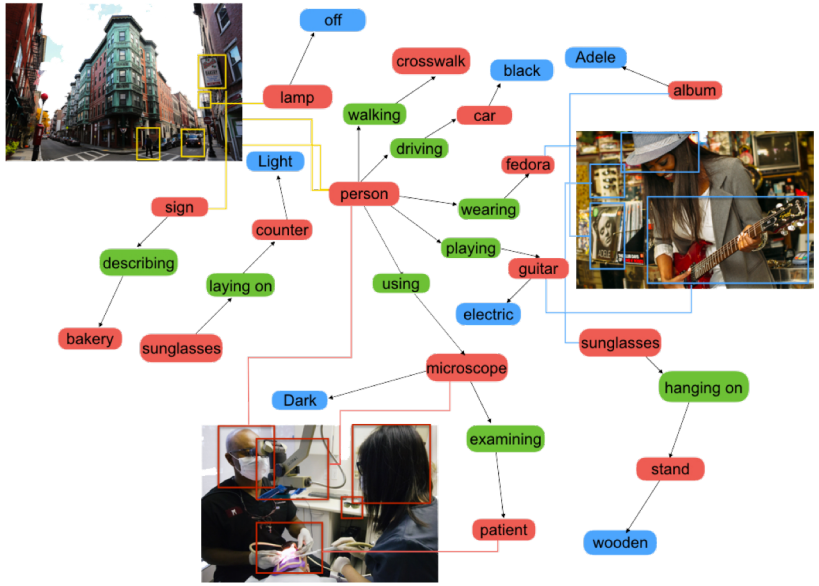
MS COCO, 2019

The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

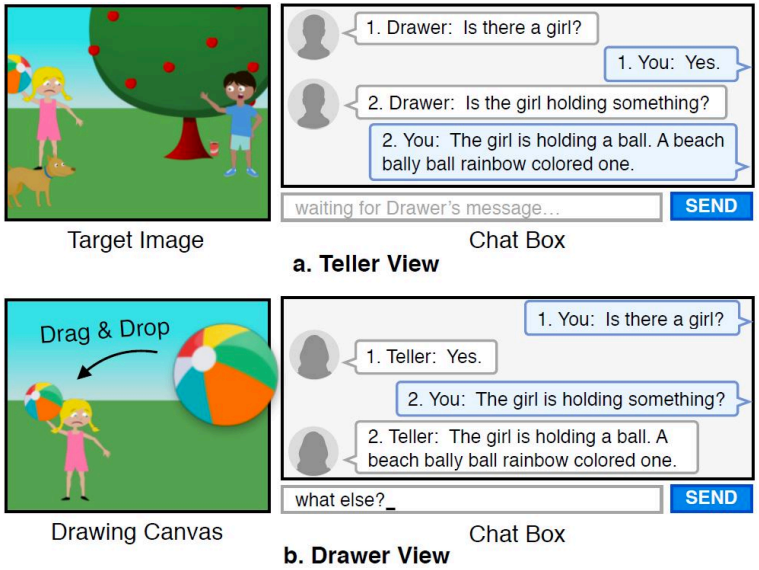
One image shows exactly two brown acorns in back-to-back caps on green foliage.

NLVR, from Cornell, 2018

Popular Datasets for Language Grounding



Visual Genome, from Stanford, 2017



CoDraw, from FAIR, 2017

More language grounding datasets available at:
http://www.denizyuret.com/2018/12/grounded-language-learning-datasets.html

2

Recent Papers

Paper One

Learning Visually Grounded Sentence Representations

Douwe Kiela, Alexis Conneau, Allan Jabri, Maximilian Nickel

ACL 2018

Motivation

- One of the biggest challenges in NLP's is to build **universal sentence representations**: generic representations of sentence meaning that can be "plugged into" any kind of system or pipeline.

- But purely text-based semantic models suffer from the **grounding problem**, which is characterized by the lack of an association between symbols and external information.

Proposed Solution

- Address this problem by aligning text with paired visual data and hypothesize that sentence representations can be enriched with external information — i.e., grounded — by forcing them to capture visual semantics.

- Instead of predicting actual images, we train a deep recurrent neural network to predict the **latent feature representation** of images. We are specifically interested in the semantic content of visual representations and how useful that information is for learning sentence representations.

- One can think of this as trying to imagine, or form a "mental picture", of a sentence's meaning.

Approach

- Use a dataset of image captions, e.g. MS COCO (Lin et al., 2014):

$$\mathcal{D} = \{(I_k, C_k)\}_{k=1}^{N}$$

where, each I_k is associated with one or more captions $C_k = \{C_1, \ldots, C_{|C|_k}\}$

- The Objective is to encode a given sentence, i.e., a caption C, and learn to ground it in the corresponding image I.

- Train a bidirectional LSTM on the caption, and then take the elementwise maximum to obtain a sentence encoding.

Model Architecture

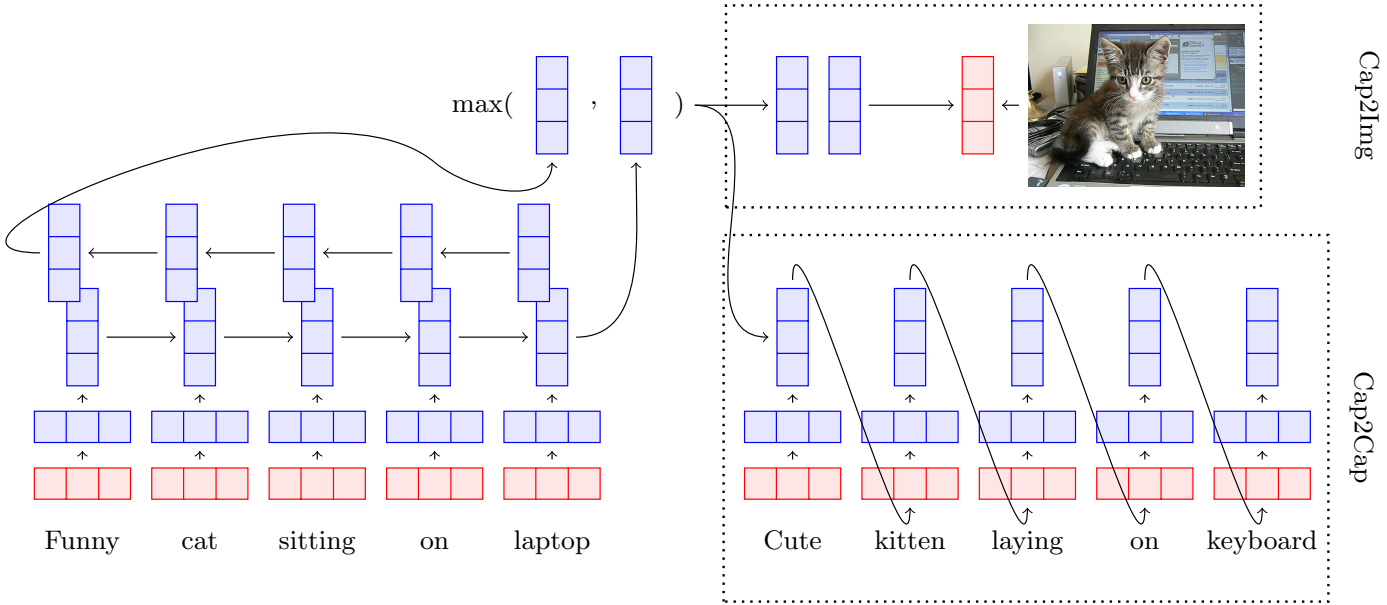


Figure 1: Model architecture: predicting either an image (Cap2Img), an alternative caption (Cap2Cap), or both at the same time (Cap2Both).

Three Methods for Grounding the Sentence Embedding

a) **Cap2Img** - try to predict the image features ⇒ strong perceptual grounding

b) **Cap2Cap** - learn to predict which other captions are valid descriptions of the same image ⇒ weaker implicit grounding

c) **Cap2Both** – optimize the objectives in both (a) and (b) jointly, predict both images and alternative captions for the same image ⇒ incorporates both strong perceptual and weak implicit grounding together

Bidirectional LSTM and Transformation of Word Embeddings

- To exploit contextual information in both input directions, we process input sentences using a bidirectional LSTM:

$$\mathbf{h}_{t+1}^{f} = \text{LSTM}(\mathbf{x}_t, \mathbf{h}_t^{f}, \mathbf{c}_t^{f} \mid \Theta^{f})$$
$$\mathbf{h}_{t+1}^{b} = \text{LSTM}(\mathbf{x}_{T-t}, \mathbf{h}_t^{b}, \mathbf{c}_t^{b} \mid \Theta^{b})$$

- And we use elementwise maximum to combine them into one single embedding:

$$\mathbf{h}_T = \max(\mathbf{h}_t^{f}, \mathbf{h}_t^{b})$$

- Use GloVe vectors for word embeddings, and learn a global transformation of GloVe space to grounded word space. More specifically, we learn a linear map $U \in \mathbb{R}^{n \times n}$ such that $\mathbf{x} = U\overline{\mathbf{x}}$ and use $\mathbf{x}$ as input to the BiLSTM. The linear map U and the BiLSTM are trained jointly.

Cap2Img

- Aims to predict the latent features of an image from its caption.
- The mapping of caption to image space is performed via a series of projections with non-linearity:

$$\mathbf{p}_0 = \mathbf{h}_T$$
$$\mathbf{p}_{\ell+1} = \psi(P_\ell \mathbf{p}_\ell)$$

- By jointly training the BiLSTM with these latent projections, we can ground the language model in its visual counterpart.

Cap2Img

- Minimize the ranking loss:

$$\mathcal{L}_{\text{C2I}}(\Theta) = \sum_{(I,C)\in\mathcal{D}} f_{\text{rank}}(I, C) + f_{\text{rank}}(C, I) \qquad (1)$$

where

$$f_{\text{rank}}(a, b) = \sum_{b'\in\mathcal{N}_a} [\gamma - \text{sim}(a, b) + \text{sim}(a, b')]_+$$

$$[x]_+ = \max(0, x) \qquad \text{Threshold Function}$$

$$\text{sim}(a, b) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|}. \qquad \text{Cosine Similarity}$$

Cap2Cap

- For a caption pair x = (x_1, …, x_T) and y = (y_1, …, y_S) that describes the same image, employ a standard sequence-to-sequence model to learn weakly grounded representations by predicting y from x:
- Joint probability of y given x: $p(y \mid x) = \prod_{s=1}^{S} p(y_s \mid \mathbf{h}_T, y_1, \ldots, y_{s-1}, \Theta)$
- Then use multiclass classification over the vocabulary of the corpus to model the conditional probability of y_S: $p(y_s = k \mid \mathbf{h}_T, y_1, \ldots, y_{s-1}, \Theta) = \frac{e^{\langle \mathbf{v}_k, \mathbf{y}_s \rangle}}{\sum_{j=1}^{|\mathcal{V}|} e^{\langle \mathbf{v}_j, \mathbf{y}_s \rangle}}$
- We minimize the negative log-likelihood over all caption pairs to train the model:

$$\mathcal{L}_{\text{C2C}}(\theta) = -\sum_{x,y \in \mathcal{D}} \sum_{s=1}^{|y|} \log p(y_s \mid \mathbf{h}_T, y_1, \ldots, y_{s-1}, \Theta)$$

Cap2Both

- Integrate both Cap2Img and Cap2Cap into a joint model, and optimize the joint loss function:

$$\mathcal{L}_{\mathrm{C2B}}(\Theta) = \mathcal{L}_{C2I}(\Theta) + \mathcal{L}_{C2C}(\Theta)$$

Grounded Universal Representations

- Grounded representations are potentially less universal than text-based representations, which also cover **abstract concepts**.

- Evidence suggests that meaning is dually coded in the human brain: while abstract concepts are processed in linguistic areas, concrete concepts are processed in both linguistic and visual areas.

- Therefore we optionally complement our systems' representations with more abstract universal sentence representations trained on language-only data. Here we combine grounded and language-only representations using simple concatenation: $r_{gs} = r_{grounded} \,||\, r_{ling-only}$, which has been proven to be a strong and straightforward mid-level multi-modal fusion method.

GroundSent (GS) ⇒ GroundSent-Img / GroundSen-Cap / GroundSent-Both

Evaluation Tasks

- Want to see how well grounded universal sentence representations transfer to different tasks.
- Evaluate all systems with the same evaluation pipeline – SentEval.
- To evaluate: take universal sentence representations and learn a simple classifier on top for each of the transfer tasks.

- **Semantic Classification**: movie review sentiment (MR) (Pang and Lee, 2005), product reviews (CR) (Hu and Liu, 2004), subjectivity classification (SUBJ) (Pang and Lee, 2004), opinion polarity (MPQA) (Wiebe et al., 2005), paraphrase identification (MSRP) (Dolan et al., 2004) and sentiment classification (SST, binary version) (Socher et al., 2013).

- **Entailment**: the large-scale SNLI dataset (Bowman et al., 2015) and the SICK dataset (Marelli et al., 2014).

Results on the COCO5K Caption and Image Retrieval Tasks

	COCO5K									
	Caption Retrieval					Image Retrieval				
Model	**R@1**	**R@5**	**R@10**	**MEDR**	**MR**	**R@1**	**R@5**	**R@10**	**MEDR**	**MR**
DVSA	11.8	32.5	45.4	12.2	NA	8.9	24.9	36.3	19.5	NA
FV	17.3	39.0	50.2	10.0	46.4	10.8	28.3	40.1	17.0	49.3
OE	23.3	NA	65.0	5.0	24.4	18.0	NA	57.6	7.0	35.9
Cap2Both	19.4	45.0	59.4	7.0	26.5	11.7	32.6	46.4	12.0	41.7
Cap2Img	27.1	55.6	70.0	4.0	19.2	17.1	43.0	57.3	8.0	36.6

Table 1: Retrieval (higher is better) results on COCO, plus median rank (MEDR) and mean rank (MR) (lower is better). Note that while this work underwent review, better methods have been published, most notably VSE++ (Faghri et al., 2017).

Results on Sentence Classification and Entailment Tasks

Model	MR	CR	SUBJ	MPQA	MRPC	SST	SNLI	SICK
ST-LN	78.1	80.1	92.7	88.0	69.6/81.2	82.9	73.8	78.5
GroundSent-Cap	79.9	81.4	93.1	88.9	72.9/82.2	85.0	75.5	79.7
GroundSent-Img	79.1	80.8	93.1	89.0	71.9/81.4	86.1	76.1	82.2
GroundSent-Both	79.6	81.7	93.4	89.4	72.7/82.5	84.8	76.1	81.6

Table 2: Accuracy results on sentence classification and entailment tasks.

The Contribution of Grounding

- An important open question is whether the increase in performance in multi-modal semantic models is due to qualitatively different information from grounding, or simply due to the fact that we have more parameters or data from a different distribution.

- In order to examine this, we implement a SkipThought-like model that also uses a bidirectional LSTM with element-wise max on the final hidden layer (STb) as a comparison.

The Contribution of Grounding

Model	MR	CR	SUBJ	MPQA	MRPC	SST	SNLI	SICK
STb-1024	70.3	68.0	87.5	85.5	69.7/80.6	78.3	67.3	76.6
STb-2048	73.1	**75.7**	88.3	86.5	71.6/**81.7**	79.0	71.0	78.8
2×STb-1024	71.4	74.7	88.2	86.6	71.3/80.7	75.8	69.4	78.3
Cap2Cap	71.4	74.7	86.7	**86.7**	70.3/79.8	76.1	68.5	78.2
Cap2Img	72.1	75.5	86.9	86.0	**72.3**/81.1	77.7	71.4	81.2
Cap2Both	71.6	74.4	86.5	85.5	71.4/79.5	78.5	71.3	**81.7**
GroundSent-Cap	73.1	73.0	**88.6**	86.6	70.8/81.2	79.4	70.7	79.1
GroundSent-Img	72.5	74.9	88.4	85.7	71.3/81.2	79.4	70.5	79.7
GroundSent-Both	**73.3**	75.2	87.5	86.6	69.9/79.9	**80.3**	**72.0**	78.1

Table 3: Thorough investigation of the contribution of grounding, ensuring equal number of components and identical architectures, on the variety of sentence-level semantic benchmark tasks. STb=SkipThought-like model with bidirectional LSTM+max. 2×STb-1024=ensemble of 2 different STb models with different initializations. GroundSent is STb-1024+Cap2Cap/Img/Both. We find that performance improvements are sometimes due to having more parameters, but in most cases due to grounding.

Concreteness

- A dataset's concreteness plays an important role in the relative merit of applying grounding: a dataset consisting mostly of abstract words is less likely to benefit from grounding than one that uses mostly concrete words.

Dataset	Concreteness
MR	2.3737 ± 0.965
CR	2.4714 ± 1.025
SUBJ	2.4510 ± 1.007
MPQA	2.3158 ± 0.834
MRPC	2.5086 ± 0.987
SST	2.7471 ± 1.142
SNLI	3.1867 ± 1.309
SICK	3.1282 ± 1.372

Table 4: Mean and variance of dataset concreteness, over all words in the datasets.

Grounded Word Embeddings

- Our models contain a projection layer that maps the GloVe word into a different embedding space.

- The grounded word projections that our network learns yield higher-quality word embeddings on four standard lexical semantic similarity benchmarks: MEN (Bruni et al., 2014), SimLex-999 (Hill et al., 2016b), Rare Words (Luong et al., 2013) and WordSim-353 (Finkelstein et al., 2001).

Model	MEN	SimLex	RW	W353
GloVe	0.805	0.408	0.451	0.738
Cap2Both	0.819	0.467	0.487	0.712
Cap2Img	0.845	0.515	0.523	0.753

Table 5: Spearman ρ_s correlation on four standard semantic similarity evaluation benchmarks.

Paper Two

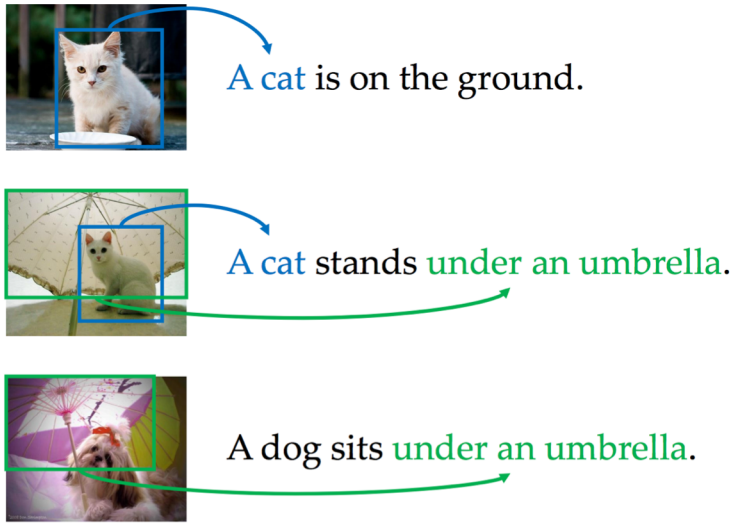
Visually Grounded Neural Syntax Acquisition

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, Karen Livescu

ACL 2019

Problem Formulation

- This paper studies the problem of **visually grounded syntax acquisition**.



- Intuition: similar spans should be matched to similar visual objects and these concrete spans form constituents

More Intuition

- Given no prior knowledge of English, and just based sufficient pairs of images and corresponding descriptive texts (captions), one can infer the correspondence between certain words and visual attributes, (e.g., recognizing that “a cat” refers to the objects in the blue boxes). One can further extract constituents, by assuming that concrete spans of words should be processed as a whole, and thus form the constituents.

- This suggests that we can use image-text pairs to facilitate automated language learning, including both **syntax** and semantics.

- In this paper we focus on the acquisition of **syntax**.

Proposed Approach

- This paper proposes the **Visually Grounded Neural Syntax Learner** (VG-NSL).
- VG-NSL acquires syntax, in the form of constituency parsing, by looking at images and reading captions.
- At a high level, VG-NSL builds latent constituency trees of word sequences and recursively composes representations for constituents. Next, it matches the visual and textual representations. The training procedure is built on the hypothesis that a better syntactic structure contributes to a better representation of constituents, which then leads to better alignment between vision and language.
- VG-NSL uses no human-labeled constituency trees or other syntactic labeling (such as part-of-speech tags). Instead, it defines a **concreteness** score of constituents based on their matching with images, and uses it to guide the parsing of sentences. At test time, no images paired with the text are needed.

Motivation

Compared to existing works on learning linguistic structures from text, we would want to build a syntax acquisition model that:

a) is robust to random initializations and produces consistent and linguistically sound structures;

b) doesn't produce meaningless parses for lower-level constituents;

c) doesn't depend on linguistic annotations, such as part of speech tags.

➢ In order to achieve these, we use parallel data from another modality (i.e. paired images and captions) to induce constituency parse trees with visual grounding.

Visually Grounded Neural Syntax Learner (VG-NSL)

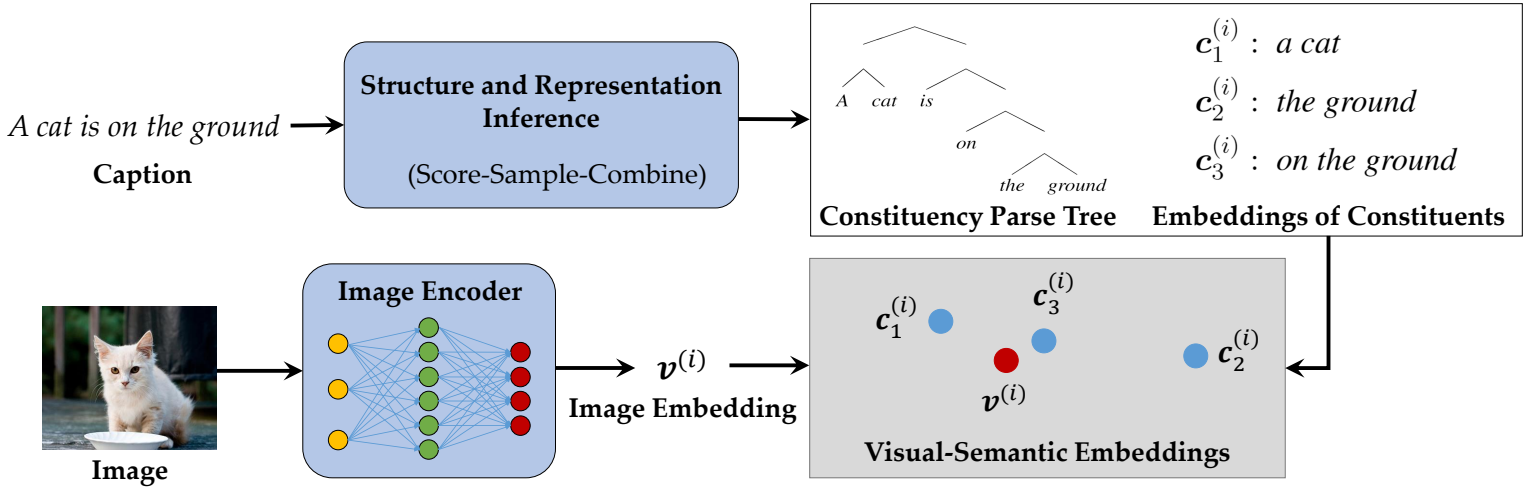
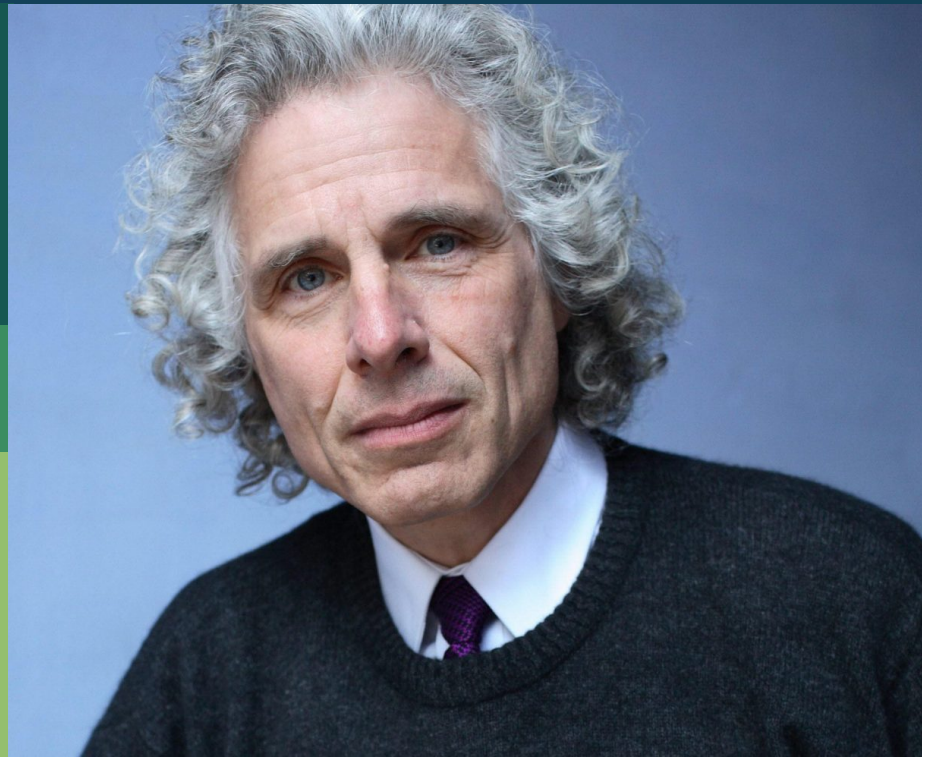
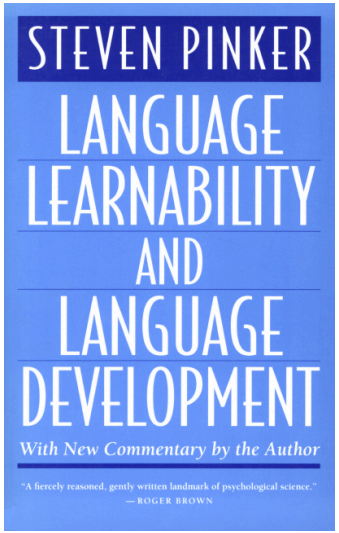


Figure 2: VG-NSL consists of two modules: a textual module for inferring structures and representations for captions, and a visual-semantic module for matching constituents with images. VG-NSL induces constituency parse trees of captions by looking at images and reading paired captions.

Semantic Bootstrapping



VG-NSL is inspired by the idea of **semantic bootstrapping** (Steven Pinker, 1984), which suggests that "children acquire syntax by first understanding the meaning of words and phrases, and linking them with the syntax of words."



Model Architecture of VG-NSL

VG-SNL consists of 2 modules:

- Textual Module: given an input caption (i.e., a sentence or a smaller constituent), as a sequence of tokens, building a latent constituency parse tree, and recursively composes representations for every constituent.

- Visual-Semantic Module: matching textual representations with visual inputs, such as the paired image with the constituents.

Both modules are jointly optimized from natural supervision: the model acquires constituency structures, composes textual representations, and links them with visual scenes, by looking at images and reading paired captions.

Module 1: Textual Representations and Structures

VG-NSL starts by composing a binary constituency structure of text, using an easy-first bottom-up parser (Goldberg and Elhadad, 2010). The composition of the tree from a caption of length n consists of $n-1$ steps. Let $\mathbf{X}^{(t)} = (\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}, \cdots, \mathbf{x}_k^{(t)})$ denote the textual representations of a sequence of constituents after step t, where $k = n - t$. For simplicity, we use $\mathbf{X}^{(0)}$ to denote the *word embeddings* for all tokens (the initial representations).

At step t, a score function $score(\cdot;\Theta)$, parameterized by Θ, is evaluated on all pairs of consecutive constituents, resulting in a vector $\boldsymbol{score}(\mathbf{X}^{(t-1)};\Theta)$ of length $n-t$:

$$score(\mathbf{X}^{(t-1)};\Theta)_j \triangleq score\left(\left[\mathbf{x}_j^{(t-1)}, \mathbf{x}_{j+1}^{(t-1)}\right];\Theta\right).$$

We implement $score(\cdot;\Theta)$ as a two-layer feed-forward network.

A pair of constituents $\left(\mathbf{x}_{j^*}^{(t-1)}, \mathbf{x}_{j^*+1}^{(t-1)}\right)$ is sampled from all pairs of consecutive constituents, with respect to the distribution produced by a `softmax`:[2]

$$\Pr[j^*] = \frac{\exp\left(score\left(\mathbf{X}^{(t-1)};\Theta\right)_{j^*}\right)}{\sum_j \exp\left(score\left(\mathbf{X}^{(t-1)};\Theta\right)_j\right)}.$$

The selected pair is combined to form a single new constituent. Thus, after step t, the number of constituents is decreased by 1. The textual representation for the new constituent is defined as the L2-normed sum of the two component constituents:

$$combine\left(\mathbf{x}_{j^*}^{(t-1)}, \mathbf{x}_{j^*+1}^{(t-1)}\right) \triangleq \frac{\mathbf{x}_{j^*}^{(t-1)} + \mathbf{x}_{j^*+1}^{(t-1)}}{\left\|\mathbf{x}_{j^*}^{(t-1)} + \mathbf{x}_{j^*+1}^{(t-1)}\right\|_2}.$$

[2] At test time, we take the `argmax`.

Module 1: Textual Representations and Structures

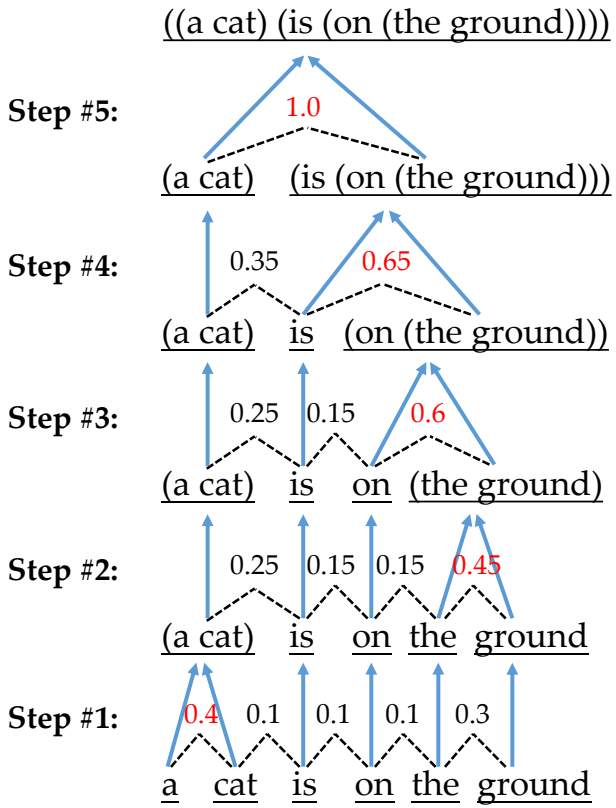


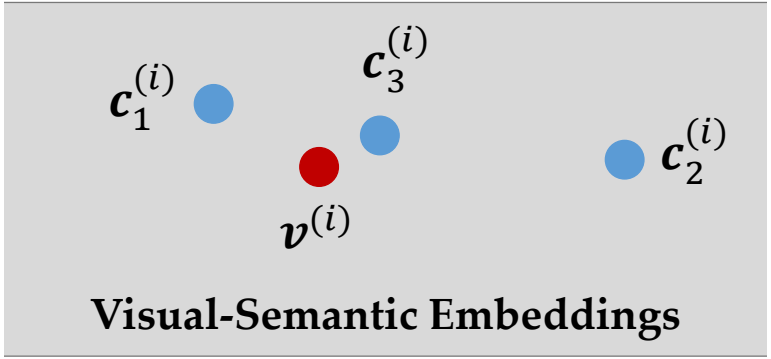
Figure 3: An illustration of how VG-NSL composes a constituency parse tree. At each step, the score function *score* is evaluated on all pairs of consecutive constituents (dashed lines). Next, a pair of constituents is sampled from all pairs w.r.t. a distribution computed by the `softmax` of all predicted scores. The selected pair of constituents is combined into a larger one, while the other constituents remain unchanged (solid lines).

Module 2: Visual-Semantic Embedding

We follow an approach similar to that of Kiros et al. (2014) to define the visual-semantic embedding (VSE) space for paired images and text constituents. Let $\mathbf{v}^{(i)}$ denote the vector representation of an image i, and $\mathbf{c}_j^{(i)}$ denote the vector representation of the j-th constituent of its corresponding text caption. During the matching with images, we ignore the tree structure and index them as a list of constituents. A function $m(\cdot, \cdot; \Phi)$ is defined as the matching score between images and texts:

$$m(\mathbf{v}^{(i)}, \mathbf{c}_j^{(i)}; \Phi) \triangleq \cos(\Phi \mathbf{v}, \mathbf{c})$$

where the parameter vector Φ aligns the visual and textual representations into a joint space.



Training

We optimize the visual-semantic representations (Φ) and constituency structures (Θ) in an alternating approach. At each iteration, given constituency parsing results of caption, Φ is optimized for matching the visual and the textual representations. Next, given the visual grounding of constituents, Θ is optimized for producing constituents that can be better matched with images. Specifically, we optimize textual representations and the visual-semantic embedding space using a hinge-based triplet ranking loss:

$$\mathcal{L}(\Phi; \mathcal{V}, \mathcal{C}) = \sum_{i,k \neq i,j,\ell} \left[m(\mathbf{c}_\ell^{(k)}, \mathbf{v}^{(i)}) - m(\mathbf{c}_j^{(i)}, \mathbf{v}^{(i)}) + \delta \right]_+ + \sum_{i,k \neq i,j} \left[m(\mathbf{c}_j^{(i)}, \mathbf{v}^{(k)}) - m(\mathbf{c}_j^{(i)}, \mathbf{v}^{(i)}) + \delta \right]_+ ,$$

where i and k index over all image-caption pairs in the data set, while j and ℓ enumerate all constituents of a specific caption ($c^{(i)}$ and $c^{(k)}$, respectively), $\mathcal{V} = \{\mathbf{v}^{(i)}\}$ is the set of image representations, $\mathcal{C} = \{\mathbf{c}_j^{(i)}\}$ is the set of textual representations of all constituents, and δ is a constant margin, $[\cdot]_+$ denotes $\max(0, \cdot)$. The loss $\mathcal{L}$ extends the loss for image-caption retrieval of Kiros et al. (2014), by introducing the alignments between images and sub-sentence constituents.

We optimize textual structures via distant supervision: they are optimized for a better alignment between the derived constituents and the images. Intuitively, the following objective encourages adjectives to be associated (combined) with the corresponding nouns, and verbs/prepositions to be associated (combined) with the corresponding subjects and objects. Specifically, we use REINFORCE (Williams, 1992) as the gradient estimator for Θ. Consider the parsing process of a specific caption $c^{(i)}$, and denote the corresponding image embedding $\mathbf{v}^{(i)}$. For a constituent $\mathbf{z}$ of $c^{(i)}$, we define its (visual) concreteness $\textit{concrete}(\mathbf{z}, \mathbf{v}^{(i)})$ as:

$$\textit{concrete}(\mathbf{z}, \mathbf{v}^{(i)}) = \sum_{k \neq i,p} \left[m(\mathbf{z}, \mathbf{v}^{(i)}) - m(\mathbf{c}_p^{(k)}, \mathbf{v}^{(i)}) - \delta' \right]_+ + \sum_{k \neq i} \left[m(\mathbf{z}, \mathbf{v}^{(i)}) - m(\mathbf{z}, \mathbf{v}^{(k)}) - \delta' \right]_+ , \quad (1)$$

where δ' is a fixed margin. At step t, we define the reward function for a combination of a pair of constituents $(\mathbf{x}_j^{(t-1)}, \mathbf{x}_{j+1}^{(t-1)})$ as:

$$r(\mathbf{x}_j^{(t-1)}, \mathbf{x}_{j+1}^{(t-1)}) = \textit{concrete}(\mathbf{z}, \mathbf{v}^{(i)}) \quad (2)$$

where $\mathbf{z} \triangleq \textit{combine}(\mathbf{x}_j^{(t-1)}, \mathbf{x}_{j+1}^{(t-1)})$. In plain words, at each step, we encourage the model to compose a constituent that maximizes the alignment between the new constituent and the corresponding image. During training, we sample constituency parse trees of captions, and reinforce each composition step using Equation 2. During test, no paired images of text are needed.

Experiments and Evaluations

A horse carrying a large load of hay and two people sitting on it.

We evaluate VG-NSL for unsupervised parsing in 4 dimensions:

1) F_1 score with gold parse trees;
2) Self-consistency across different choice of random initialization;
3) Performance on different types of constituents;
4) Data efficiency.

Dataset: MS COCO (Lin et al, 2014), which contains 82,783 images for training, 1,000 for development, and another 1,000 for testing. Each image is associated with 5 captions.

Gold Parse Trees: We use **Benepar** (Kitaev and Klein, 2018), an off-the-shelf constituency parser trained on the Penn Treebank (PTB; Marcus et al., 1993) to parse the captions in the MSCOCO test set as gold constituency parse trees.

Baselines

In our experiment, we compare VG-SNL with the following 4 categories of baselines for unsupervised tree structure modeling of texts:

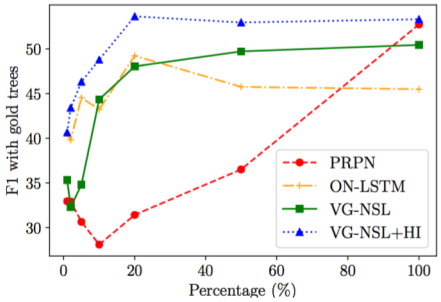
- Trivial Tree Structures
- Syntax Acquisition by Language Modeling and Statistics
- Syntax Acquisition from Downstream Tasks
- Syntax Acquisition from Concreteness Estimation

Experiment Results

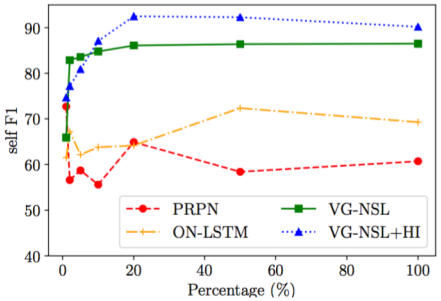
Model	NP	VP	PP	ADJP	Avg. F_1	Self F_1
Random	$47.3_{\pm 0.3}$	$10.5_{\pm 0.4}$	$17.3_{\pm 0.7}$	$33.5_{\pm 0.8}$	$27.1_{\pm 0.2}$	32.4
Left	51.4	1.8	0.2	16.0	23.3	N/A
Right	32.2	23.4	18.7	14.4	22.9	N/A
PMI	54.2	16.0	14.3	39.2	30.5	N/A
PRPN (Shen et al., 2018a)	$72.8_{\pm 9.7}$	$33.0_{\pm 9.1}$	$61.6_{\pm 9.9}$	$35.4_{\pm 4.3}$	$52.5_{\pm 2.6}$	60.3
ON-LSTM (Shen et al., 2019)	$74.4_{\pm 7.1}$	$11.8_{\pm 5.6}$	$41.3_{\pm 16.4}$	**44.0**$_{\pm 14.0}$	$45.5_{\pm 3.3}$	69.3
Gumbel (Choi et al., 2018)†	$50.4_{\pm 0.3}$	$8.7_{\pm 0.3}$	$15.5_{\pm 0.0}$	$34.8_{\pm 1.6}$	$27.9_{\pm 0.2}$	40.1
VG-NSL (ours)†	**79.6**$_{\pm 0.4}$	$26.2_{\pm 0.4}$	$42.0_{\pm 0.6}$	$22.0_{\pm 0.4}$	$50.4_{\pm 0.3}$	87.1
VG-NSL+HI (ours)†	$74.6_{\pm 0.5}$	$32.5_{\pm 1.5}$	**66.5**$_{\pm 1.2}$	$21.7_{\pm 1.1}$	**53.3**$_{\pm 0.2}$	**90.2**
VG-NSL+HI+FastText (ours)*†	$78.8_{\pm 0.5}$	$24.4_{\pm 0.9}$	$65.6_{\pm 1.1}$	$22.0_{\pm 0.7}$	**54.4**$_{\pm 0.4}$	89.8
Concreteness estimation–based models						
Turney et al. (2011)*	65.5	30.8	35.3	30.4	42.5	N/A
Turney et al. (2011)+HI*	74.5	26.2	47.6	25.6	48.9	N/A
Brysbaert et al. (2014)*	54.1	27.8	27.0	33.1	34.1	N/A
Brysbaert et al. (2014)+HI*	73.4	23.9	50.0	26.1	47.9	N/A
Hessel et al. (2018)†	50.9	21.7	32.8	27.5	33.2	N/A
Hessel et al. (2018)+HI†	72.5	**34.4**	65.8	26.2	52.9	N/A

Table 1: Recall of specific typed phrases, and overall F_1 score, evaluated on the MSCOCO test split, averaged over 5 runs with different random initializations. We also include self-agreement F_1 score (Williams et al., 2018) across the 5 runs. $\pm$ denotes standard deviation. * denotes models requiring extra labels and/or corpus, and † denotes models requiring a pre-trained visual feature extractor. We highlight the best number in each column among all models that do not require extra data other than paired image-caption data, as well as the overall best number. The Left, Right, PMI, and concreteness estimation–based models have no standard deviation or self F_1 (shown as N/A) as they are deterministic given the training and/or testing data.

Experiment Results



(a) The percent data-F_1 curves.



(b) The percent data-self F_1 curves.

Figure 4: F_1 score and self F_1 score with respect to the amount of training data. All numbers are averaged over 5 runs with different random initialization.

Model/method	VG-NSL	(+HI)
Turney et al. (2011)	0.74	0.72
Brysbaert et al. (2014)	0.71	0.71
Hessel et al. (2018)	0.84	0.85

Table 2: Agreement between our concreteness estimates and existing models or labels, evaluated via the Pearson correlation coefficient computed over the most frequent 100 words in the MSCOCO test set, averaged over 5 runs with different random initialization.

Model	**Criterion**	**Avg. F_1**	**Self F_1**
VG-NSL	Self F_1	**50.4** ± 0.3	**87.1**
VG-NSL	R@1	47.7 ± 0.6	83.4
VG-NSL+HI	Self F_1	**53.3** ± 0.2	**90.2**
VG-NSL+HI	R@1	53.1 ± 0.2	88.7

Table 3: Average F_1 scores and Self F_1 scores of VG-NSL and VG-NSL+HI with different model selection methods. R@1 denotes using recall at 1 (Kiros et al., 2014) as the model selection criterion. All hyperparameters are tuned with respect to self-agreement F_1 score. The numbers are comparable to those in Table 1.

Paper Three

Incorporating Visual Semantics into Sentence Representations within a Grounded Space

Patrick Bordes, Eloi Zablocki, Laure Soulier,
Benjamin Piwowarski, Patrick Gallinari

ACL 2019

Motivation

- Visual language grounding is an active research field aiming at enriching textual representations with visual information.
- In most existing works, textual and visual elements are embedded in the same representation space, which implicitly assumes a one-to-one correspondence between the two modalities.
- But this underlying assumption could be problematic …

Problems with Existing Approaches

- In most existing approaches, cross-modal projections are learned to incorporate visual semantics in the final representations (Lazaridou et al., 2015b; Collell et al., 2017; Kiela et al., 2018).

- Implicitly based on the hypothesis of a one-to-one correspondence between modalities: an image of an object univocally represents a word.

- However, no obvious reason implies that the structure of the two spaces should match. Collell and Moens (2018) even empirically show that cross-modal projection of a source modality does not resemble the target modality in terms of its neighborhood structure. This is especially the case for sentences, where many different sentences can describe a similar image.

- Therefore, learning grounded representations with cross-modal projections onto a visual space is not appropriate to incorporate visual semantics in text representations, especially for sentence representations

Proposed Approach

In order to overcome this issue, we propose an alternative approach where the structure of the visual space is partially transferred to the textual space. To achieve this, we distinguish between two types of complementary information sources:

- The **cluster information**: the implicit knowledge that sentences associated with the same image refer to the same underlying reality.
- The **perceptual information**: the information contained within high-level representations of images.

These two sources of information aim at transferring the **structure** of the visual space to the textual space. To preserve textual semantics and to avoid an over-constrained textual space, we propose to incorporate the visual information to textual representations using an intermediate representation space that we call **grounded space**, on which cluster and perceptual objectives are trained.

Inflexibility with One-to-One Correspondence

There are three major inflexibility issues with enforcing one-to-one correspondence between the visual space and the textual space:

- (1) A caption can have a wide variety of paraphrases and related sentences describing the same scene:
  e.g., 'The kitten is devouring a mouse.' versus 'a cat eating a mouse'

- (2) A caption can be visually ambiguous:
  e.g., 'A cat is eating.' can be associated with many different images, depending on the visual scene/context

- (3) A caption may carry non-visual information:
  e.g., 'Cats often think about their meals.'

Justification for the Proposed Approach

In order to overcome this lack of flexibility, we propose our approach as follow:

- To cope with (1), we encourage that sentences associated with the same image should be close ⇒ **cluster information**

- To cope with (2), avoid projecting sentences to a particular point of the visual space. Instead, we require that the similarity between two images in the visual space should be close to the similarity between their associated sentences in the textual space.
  ⇒ **perceptual information**

- To cope with (3), i.e. to preserve non-visual information in sentence representations, we make use of an intermediate space called **grounded space**, that allows textual representations to benefit from visual properties without degrading the semantics brought by the textual objective.

Model Overview

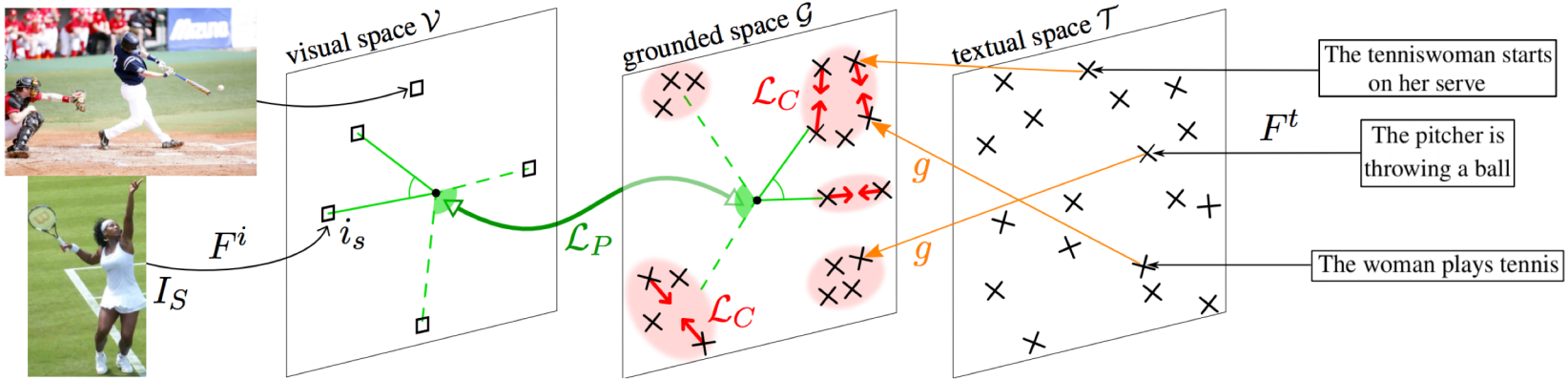


Figure 1: Model overview. Red circles indicate visual clusters. Red arrows represent the gradient of the cluster loss, which gathers visually equivalent sentences — the contrastive term in loss $\mathcal{L}_C$ is not represented. The green arrow and angles illustrate the perceptual loss, ensuring that cosine similarities correlate across modalities. The origin is at the center of each space.

The Model balances a textual objective $\mathcal{L}_\mathcal{T}$ with an additional grounding objective $\mathcal{L}_\mathcal{G}$:

$$\mathcal{L}(\theta^t, \theta^i) = \mathcal{L}_\mathcal{T}(\theta^t) + \mathcal{L}_\mathcal{G}(\theta^t, \theta^i)$$

Use the SkipThought model
(Kiros et al., 2015)

Model Objectives

The model assumes that the structure of the textual space might be partially modeled on the structure of the visual space. So instead of directly applying the grounding objectives on a sentence s embedding, we propose to train the grounding objective $\mathcal{L}_\mathcal{G}$ on an intermediate **grounded space** g(s), where g is a multi-layer perceptron. Then the objective functions of our model is formulate as:

- Cluster Loss: $\mathcal{L}_C = \sum_{(s,s^+,s^-)} \left[\gamma - \cos(g(s), g(s^+)) + \cos(g(s), g(s^-))\right]_+$

  where s^+ (resp. s^-) is a randomly sampled visually equivalent (resp. different) sentence to s.

- Perceptual Loss: $\mathcal{L}_\mathcal{P} = -\rho(\{sim^{\text{text}}_{k_1,k_2}\}, \{sim^{\text{im}}_{k_1,k_2}\})$

  where ρ is the Pearson correlation, $sim^{\text{text}}_{k_1,k_2} = \cos(g(s_{k_1}), g(s_{k_2}))$ and $sim^{\text{im}}_{k_1,k_2} = \cos(i_{k_1}, i_{k_2})$ are respectively textual and visual similarities computed over several randomly sampled pairs of matching sentences and images.

- Grounded Loss: $\mathcal{L}_\mathcal{G}(\theta^t, \theta^i) = \alpha_C \mathcal{L}_C(\theta^t, \theta^i) + \alpha_P \mathcal{L}_P(\theta^t, \theta^i)$

Experiment Results

- Textual Dataset: Toronto BookCorpus (Zhu et al., 2015)
- Visual Dataset: MS COCO (Lin et al., 2014)

	Structural measures				Semantic relatedness				
Model	mNNO	ρ_{vis}	C_{inter}	C_{intra}	STS/All	STS/Cap	STS/News	STS/Forum	SICK
T	10.0	4.1	54.2	70.1	30	41	36	21	51
CM (text)	24.2	12.8	41.7	74.8	52	76	42	**37**	55
$\mathbf{P}_{id}$	21.1	**37.9**	42.2	69.3	45	66	41	34	54
$\mathbf{C}_{id}$	27.5	10.5	**2.9**	**84.7**	60	83	45	20	55
$\mathbf{C}_{id} + \mathbf{P}_{id}$	**27.9**	25.8	6.7	82.6	**61**	**84**	**46**	28	**57**
CM (vis.)	27.1	19.2	1.5	85.8	56	78	40	34	55
$\mathbf{P}_g$	21.3	**32.4**	43.9	73.3	45	66	41	**37**	53
$\mathbf{C}_g$	28.6	9.4	**1.1**	**88.5**	62	83	46	29	59
$\mathbf{C}_g + \mathbf{P}_g$	**28.9**	29.1	4.7	87.5	**63**	**84**	**48**	33	**60**

Table 1: Intrinsic evaluations carried out on the grounded space for models with g = MLP; the textual space for **T**, **CM** (text) and models with $g = id$; and the visual space for **CM** (vis).

Experiment Results

Model		MR	CR	SUBJ	MPQA	MRPC	SST	SNLI	SICK	AVG
(Kiros et al., 2015)†	$\mathbf{T}_{1024}$	72.7*	75.2*	90.6*	84.7*	71.8* /79.2*	76.2*	68.8*	79.3*	77.4
(Kiela et al., 2018)†	GS-Cap	72.0*	76.8*	90.7*	85.5*	72.9/80.6	76.7*	73.7	**82.9**	78.4
(Kiela et al., 2018)†	GS-Img	74.5*	79.3*	90.8*	87.8*	73.0/80.3	80.0*	72.2*	80.9*	79.8
(Kiela et al., 2018)†	GS-Both	72.5*	75.7*	90.7*	85.4*	72.9/81.3	76.7*	72.2*	81.4*	78.4
(Kiros et al., 2015)†	**T**	75.9*	79.2*	92.0	86.7*	72.2/80.2	81.8*	72.0*	81.1*	80.1
(Lazaridou et al., 2015a)‡	**T** + **CM**	77.6	81.4	92.6	88.3	73.5/81.1	82.0*	73.0	81.4*	81.1
(Collell et al., 2017)‡	**SEQ**	76.1*	79.8*	92.5	86.7*	70.0* /79.5*	81.7*	67.3*	76.7*	78.9
Model scenarios	$\mathbf{T} + \mathbf{P}_{id}$	77.5	81.5	92.7	88.4	**73.7**/81.3	82.4	72.4	81.1	81.2
	$\mathbf{T} + \mathbf{P}_{g}$	**77.8**	**81.8**	**93.0**	88.1	73.3/**81.6**	**83.5**	72.8	82.2	**81.6**
	$\mathbf{T} + \mathbf{C}_{id}$	77.5	81.6	92.8	88.3	72.9/80.5	82.2	73.1	82.3	81.3
	$\mathbf{T} + \mathbf{C}_{g}$	77.3	81.5	92.8	**88.6**	73.6/81.1	82.6	**74.1**	82.6	**81.6**
	$\mathbf{T} + \mathbf{C}_{id} + \mathbf{P}_{id}$	77.3	81.2	**93.0**	88.4	73.0/80.6	82.5	73.5	82.1	81.4
	$\mathbf{T} + \mathbf{C}_{g} + \mathbf{P}_{g}$	77.4	81.5	**93.0**	88.1	73.2/80.9	82.7	73.9	**82.9**	**81.6**

Table 3: Extrinsic evaluations with SentEval. All models give sentences in dimension $d_t = 2048$ (except $\mathbf{T}_{1024}$). 'AVG' stands for the average accuracies reported in the other columns. '†': the model has been re-implemented (we obtained higher scores than the one given in the original papers). '‡': the baseline is an adaptation of the model to the case of sentences. '$*$': significantly differs from the best scenario among our models.

Experiment Results

Query: A woman sitting on stone steps with a suitcase full of books.

Q Query image

N Nearest image

Grounded model

Q A woman sitting on stairs has a suitcase full of books.

Q A woman reads a book while sitting on steps near a suitcase full of books.

Q The woman is setting on the steps with a case of books.

A woman sitting inside of an open suitcase.

N A woman sitting on the ground next to luggage.

Q A young woman sits near three suitcases of luggage.

Textual model

A young woman sitting cross legged on an apartment sofa.

A woman sitting on a couch in front of a laptop.

N A girl sitting next to three old suitcases.

A woman standing on a tennis court holding a racquet.

Q The woman is setting on the steps with a case of books.

A woman standing on a tennis court holding a racquet.

Figure 2: Nearest neighbors of a selected sentence in the validation set of MS COCO, for both grounded and purely textual models. Q is the query image, N is the nearest neighbor of Q in the visual space. Sentences that are caption of Q or N are prefixed with Q or N.

Query	Textual model	Grounded model
Two people are in **love**	Two people are fencing indoors	A couple just got married and are taking a picture with family
A man is **horrified**	A man and a woman are smiling	A teenage boy wearing a cap looks irritated
This is a **tragedy**	A group of people are at a party	Men doing a war reenactment

Table 2: Qualitative analysis: nearest neighbor of a given query (containing an abstract word) among Flickr30K sentences.

Visualization of Learned Sentence Embeddings

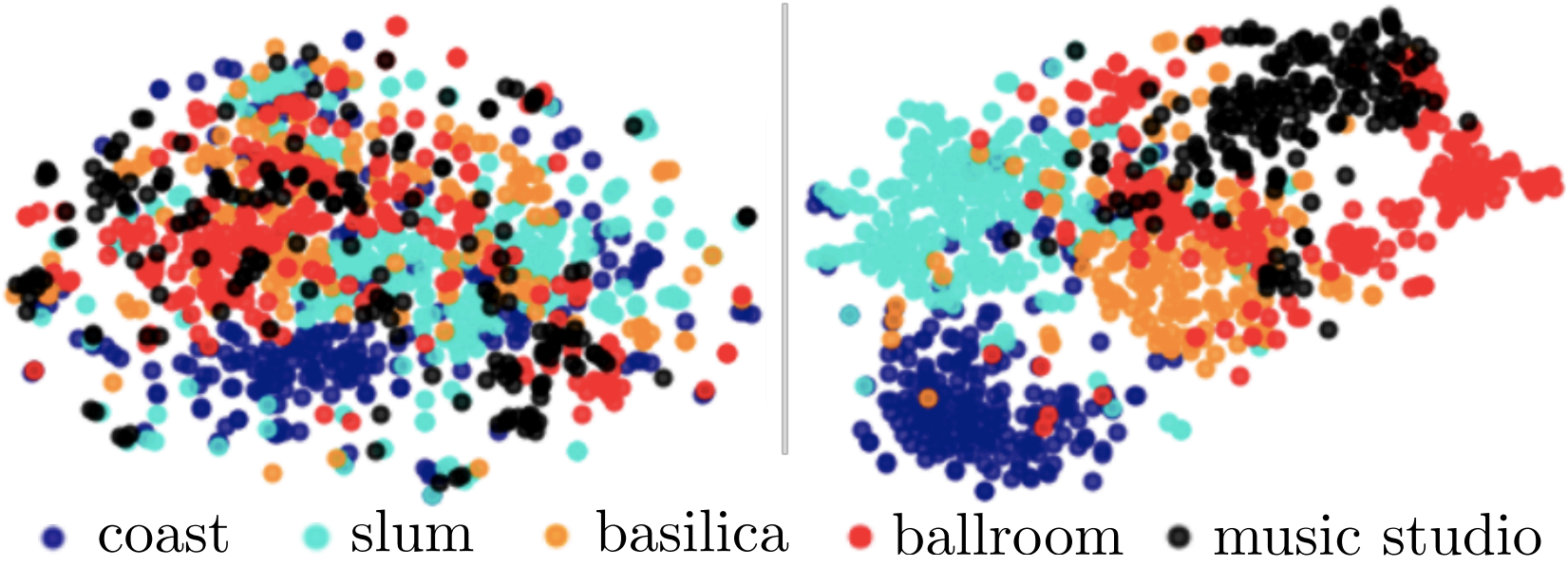
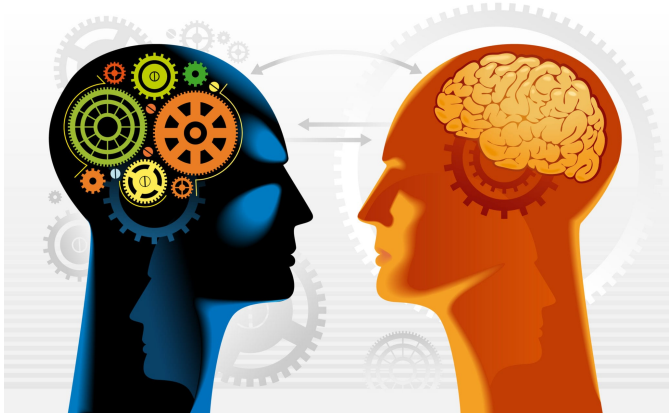


Figure 3: t-SNE visualization on CMPlaces sentences for a set of randomly sampled visual scenes. Left: textual model $\mathbf{T}$. Right: grounded model $\mathbf{C}_g + \mathbf{P}_g$.

3



Discussion Questions

Discussion Question 1:

What's the best way to ground abstract words, phrases and sentences in NLP?

Discussion Question 2:

Is there a way to perform ‘descriptive’ or ‘explanatory’ grounding on the words and phrases in natural languages? In other words, can we utilize the descriptive information contained in dictionaries and encyclopedias to learn better word embeddings and sentence representations? Why or why not?

Discussion Question 3:

How to effectively learn physical commonsense knowledge from visual language grounding?

Discussion Question 4:

What useful future applications of grounded NLP do you foresee? In particular, mobile applications.

Discussion Question 5:

What are some potential issues when multiple people interact with the same grounded system?

Thank you!