# Evaluating compositionality in sentences embeddings

**Ishita Dasgupta**
Harvard University, Computational Cognitive Neuroscience Lab
CogSci 2018, Learning as program induction
July 25th, 2018

# What/why compositionality?

X is taller than me
⇒ I am not taller than X

X = The man
X = The thin man
X = The man with the red hat
X = The man who just ate the muffin
X = The thin man with the red hat who just ate the muffin
…

Need to understand the abstract / functional rules for how words combine.

Simple domain that utilizes these abstract rules?

# Natural Language Inference (NLI)

Pairs of sentences (Premise and Hypothesis) that are related by one of

1. Contradiction
2. Neutral
3. Entailment.

3-way discriminative classifier

# Compositionality in NLI

X is more Y than Z

*Contradicts:*

- Z is more Y than X
- X is less Y than Z
- X is not more Y than Z

*Entails:*

- Z is not more Y than X
- Z is less Y than X

X and Z can be any noun phrase, and Y can be any adjective, and the conclusion holds.

A good sentence representation should capture these rules.

# Questions of Interest

Given some sentence representation,

1. How do we test if specific abstract structure has been learned?
2. How can we better understand the rules that were learned?
3. Are there ways to have these architectures learn this abstract structure?

Today's talk: Present a **new comparisons NLI dataset** and elucidate how it helps answer some of these questions.*

*Related work: White et al. 2017., Pavlick & Callison-Burch. 2016., Ettinger et al. 2016.

# Questions of Interest

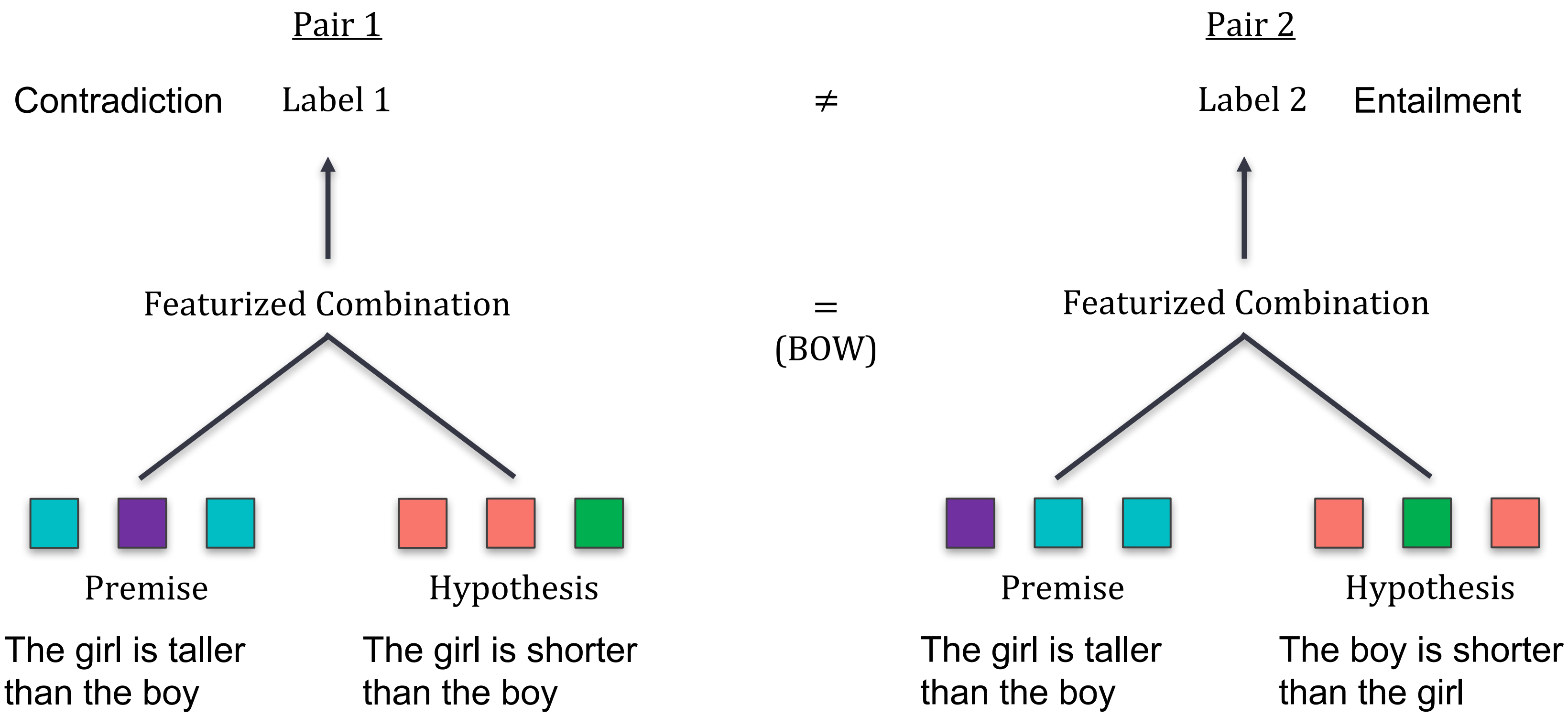
Given some sentence representation,

1. $\boxed{\text{How do we test if specific abstract structure has been learned?}}$
2. How can we better understand the rules that were learned?
3. Are there ways to have these architectures learn this abstract structure?

Today's talk: Present a **new comparisons NLI dataset** and elucidate how it helps answer some of these questions.*

*Related work: White et al. 2017., Pavlick & Callison-Burch. 2016., Ettinger et al. 2016.

# Comparisons NLI Dataset




Maximum BOW performance = 50%

# Only order change: Comparisons

```
A: The woman is more cheerful than the man
B: The woman is more cheerful than the man
ENTAILMENT
A: The woman is more cheerful than the man
B: The man is more cheerful than the woman
CONTRADICTION
```

# Order + one word: Comparisons (more/less type)

```
A: The woman is more cheerful than the man
B: The woman is less cheerful than the man
CONTRADICTION
A: The woman is more cheerful than the man
B: The man is less cheerful than the woman
ENTAILMENT
```

# Order + one word: Comparisons (not type)

```
A: The woman is more cheerful than the man
B: The woman is not more cheerful than the man
CONTRADICTION
A: The woman is more cheerful than the man
B: The man is not more cheerful than the woman
ENTAILMENT
```

# Comparisons NLI Dataset

Premise: X is more Y than Z

| Type | Entailment hypothesis | Contradiction hypothesis | # of pairs |
|---|---|---|---|
| Same | X is more Y than Z | Z is more Y than X | 14670 |
| More-Less | Z is less Y than X | X is less Y than Z | 14670 |
| Not | Z is not more Y than X | X is not more Y than Z | 14670 |

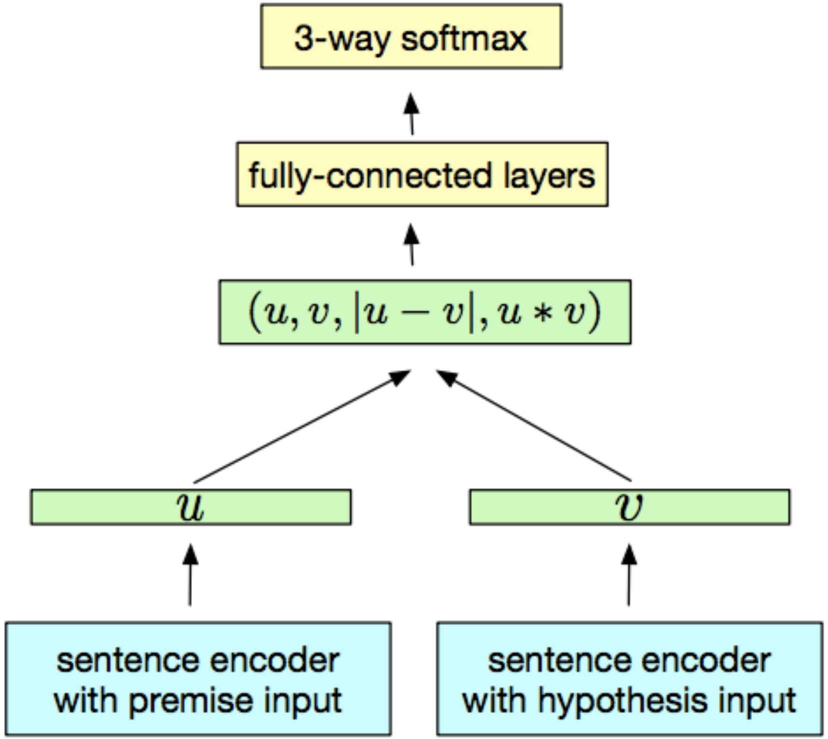
# Questions of Interest

Given some sentence representation,

1. How do we test if specific abstract structure has been learned?
2. How can we better understand the rules that were learned?
3. Are there ways to have these architectures learn this abstract structure?

Today’s talk: Present a **new comparsons NLI dataset** and elucidate how it helps answer some of these questions.*

# Example sentence embeddings: InferSent




SOTA on *transfer tasks* – embeddings perform well on tasks that they were not trained on.

1. What is the input to the sentence encoder? GLoVe embeddings.

2. How does it encode sentences? Recurrent neural networks.

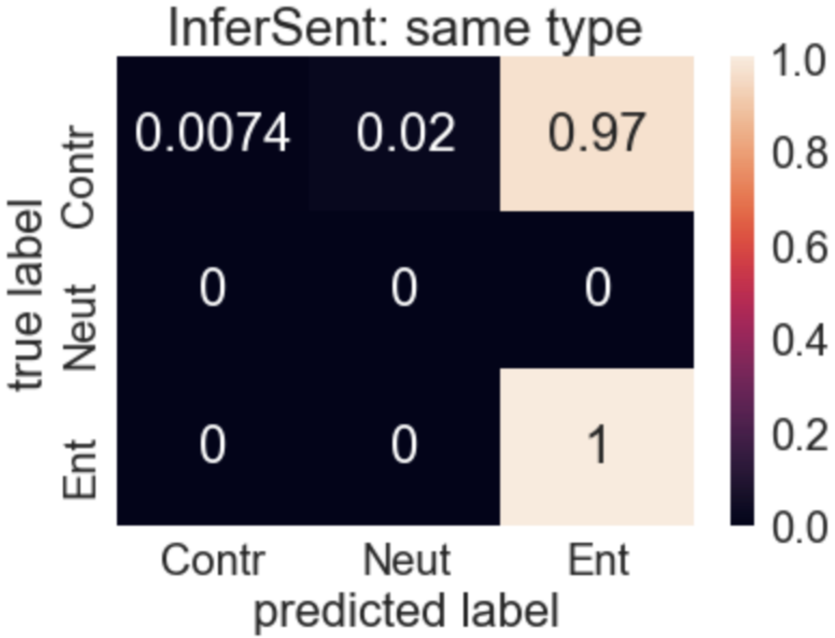
3. What is the labelled training set? Human generated pairs (SNLI)

*Conneau et al. arXiv:1705.02364 (2017).

# Performance of InferSent on Comp-NLI

| Type | BOW-MLP | InferSent |
|---|---|---|
| same | 50.0 | 50.37 |
| more/less | 30.24 | 50.35 |
| not | 48.98 | 45.24 |

# Performance of InferSent on Comp-NLI: same type




```
A: The woman is more cheerful than the man
B: The woman is more cheerful than the man
ENTAILMENT
A: The woman is more cheerful than the man
B: The man is more cheerful than the woman
CONTRADICTION
```

InferSent classifies close to all as entailment, despite half being true contradictions

Note: The premise and hypothesis here have very high word overlap.

# Performance of InferSent on Comp-NLI: same type

Hypothesis: InferSent disfavors contradiction for sentence pairs with high word overlap.

Is this supported by its training data?

Sort the SNLI dataset by extent of overlap, in decreasing order.

| Top | Entailment | Neutral | Contradiction |
|---|---|---|---|
| All | 33.4% | 33.3% | 33.3% |

# Performance of InferSent on Comp-NLI: more/less type

Hypothesis: InferSent favors contradiction for sentence pairs that differ by an antonym.

Is this supported by its training data?

Check for the presence of antonyms in sentence pairs in SNLI.

| | P(Antonym \| X) | P( X \| Antonym) |
|---|---|---|
| X = Contradiction | 12.2% | 61.2% |
| X = Entailment | 3.5% | 18.0% |

# Performance of InferSent on Comp-NLI: not type

Hypothesis: InferSent favors contradiction for sentence pairs that differ by a negation.

Is this supported by its training data?

Check for difference of negation in sentence pairs in SNLI.

| | P(Negation \| X) | P( X \| Negation) |
|---|---|---|
| X = Contradiction | 3.3 % | 58.4 % |
| X = Entailment | 1.1 % | 20.0 % |

# Questions of Interest

Given some sentence representation,

1. How do we test if specific abstract structure has been learned?
2. How can we better understand the rules that were learned?
3. Are there ways to have these architectures learn this abstract structure?

Today's talk: Present a **new comparsions NLI dataset** and elucidate how it helps answer some of these questions.*

# Training on the Comparisons NLI dataset

| | Train | Validation | Test |
|---|---|---|---|
| SNLI | 550,152 | 10,000 | 10,000 |
| Comp-NLI | 40,0010 | 2,000 | 2,000 |

| Training set | Test (Comp-NLI) | Test (SNLI) |
|---|---|---|
| SNLI | 45.36% | 84.84% |
| SNLI + Comp-NLI | 100.0% | 84.96% |

No loss in test performance on SNLI, and still achieves close to perfect on test sets from Comp-NLI dataset

# Compositionality in InferSent after training on Comp-NLI

X is more Y than Z

*Contradicts:*

- Z is more Y than X
- X is less Y than Z
- X is not more Y than Z

*Entails:*

- Z is not more Y than X
- Z is less Y than X

X and Z can be any noun phrase, and Y can be any adjective, and the conclusion holds**.

**Tested for X, Y and Z InferSent has seen before, but never in the same combination.

# Generalization: X, Y and Z not seen before

1. Random words that do not appear in SNLI / CompNLI.

2. Random GloVe vector – 300 dimensional uncorrelated Gaussian.

3. Divide CompNLI into “long” and “short” noun phrase types

   For example:

   short = *the man is more cheerful than the woman*

   long = *the man with a red hat is more cheerful than the woman with a blue coat*

   Train on only one sub-type, other sub-type is not seen before.

# Generalization: X, Y and Z not seen before

| | Additional training (Beyond SNLI) | | |
|---|---|---|---|
| **Test Set** | **Full CompNLI** | **Only Long** | **Only Short** |
| **Random word** | **83.7** | 72.9 | 82.0 |
| **Random vector** | **82.5** | 77.4 | 83.2 |
| **Only Long** | 100 | 100 | **91.1** |
| **Only Short** | 100 | **74.5** | 100 |

# Compositionality in InferSent after training on Comp-NLI

X is more Y than Z

*Contradicts:*

- Z is more Y than X
- X is less Y than Z
- X is not more Y than Z

*Entails:*

- Z is not more Y than X
- Z is less Y than X

X and Z can be any noun phrase, and Y can be any adjective, and the conclusion holds**.

**Even for X and Z InferSent has never seen before.

# Take-aways and future directions

1. The datasets on which NLP systems are evaluated do not test directly for structure – **Need datasets that test for specific abilities***.
2. These datasets can also be used as **diagnostic tools** to identify what these systems actually learn and accordingly suggest improvements.
3. Augmenting training with this dataset shows positive initial results on **learning abstract/functional rules**.
4. Future work: Is such data augmentation a **scalable tool** for *teaching* these systems more sophisticated forms of compositionality.
    a. Does learning one speed up learning others?
    b. Can we automate generating adversarial functional forms?
    c. How much data would we need?

*Related work: White et al. 2017., Pavlick & Callison-Burch. 2016., Ettinger et al. 2016.

# Acknowledgments


Demi Guo,
Harvard

Noah Goodman,
Stanford

Andreas Stuhlmüller,
Stanford

Sam Gershman,
Harvard


For more info:

1. Poster at the back of the room, and on Friday!
2. Evaluating Compositionality in Sentence Embeddings, arXiv:1802.04302.
3. github.com/ishita-dg/ScrambleTests