Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality





Anuj Diwan, Layne Berry*, Eunsol Choi, David Harwath, Kyle Mahowald*

University of Texas at Austin

EMNLP 2022

Talk Overview

1. Background: Winoground
2. Models of Interest (CLIP, UNITER, LXMERT) and Winoground
3. Analyzing the dataset
4. Analyzing the evaluation criteria
5. Analyzing the models

Talk Overview

1. **Background: Winoground (Thrush et al., 2022)**

Background: The Winoground Visuolinguistic Compositionality Benchmark

T_0 : "An old person kisses a young person."

T_1 : "A young person kisses an old person."

Background: The Winoground Visuolinguistic Compositionality Benchmark

Text Score = $\mathbb{I}[M(I_0, T_0) > M(I_0, T_1)] \wedge \mathbb{I}[M(I_1, T_1) > M(I_1, T_0)]$

Image Score = $\mathbb{I}[M(I_0, T_0) > M(I_1, T_0)] \wedge \mathbb{I}[M(I_1, T_1) > M(I_0, T_1)]$

Group Score = Text Score and Image Score are both 1

I_0

T_0 "An old person kisses a young person."

I_1

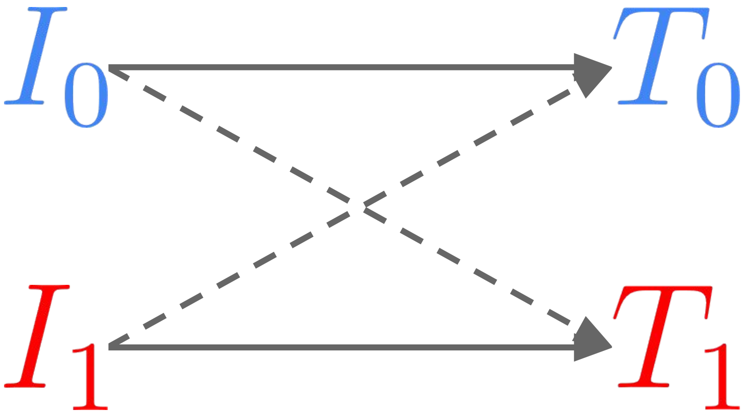
T_1 "A young person kisses an old person."

Background: The Winoground Visuolinguistic Compositionality Benchmark

Text Score = $\mathbb{I}[M(I_0, T_0) > M(I_0, T_1)] \wedge \mathbb{I}[M(I_1, T_1) > M(I_1, T_0)]$

Image Score = $\mathbb{I}[M(I_0, T_0) > M(I_1, T_0)] \wedge \mathbb{I}[M(I_1, T_1) > M(I_0, T_1)]$

Group Score = Text Score and Image Score are both 1



“An old person kisses a young person.”

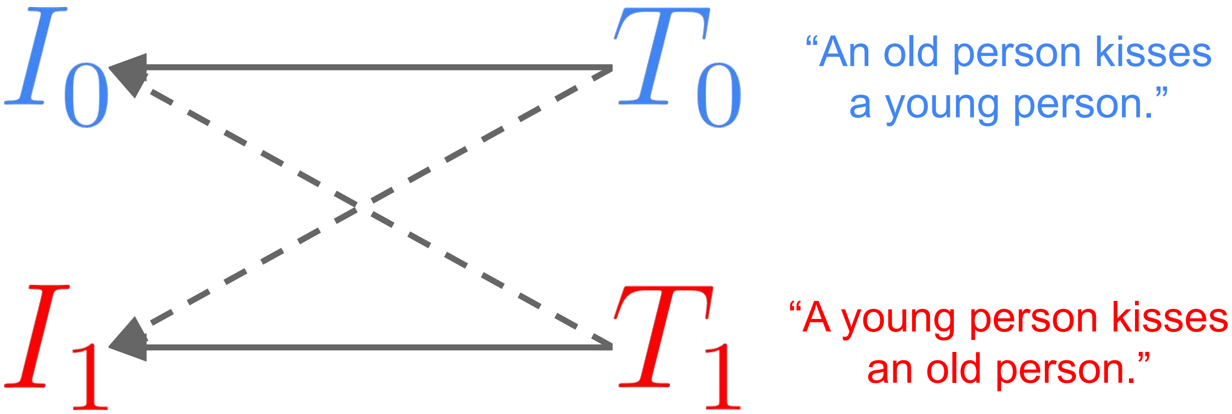
“A young person kisses an old person.”

Background: The Winoground Visuolinguistic Compositionality Benchmark

Text Score = $\mathbb{I}[M(I_0, T_0) > M(I_0, T_1)] \wedge \mathbb{I}[M(I_1, T_1) > M(I_1, T_0)]$

Image Score = $\mathbb{I}[M(I_0, T_0) > M(I_1, T_0)] \wedge \mathbb{I}[M(I_1, T_1) > M(I_0, T_1)]$

Group Score = Text Score and Image Score are both 1

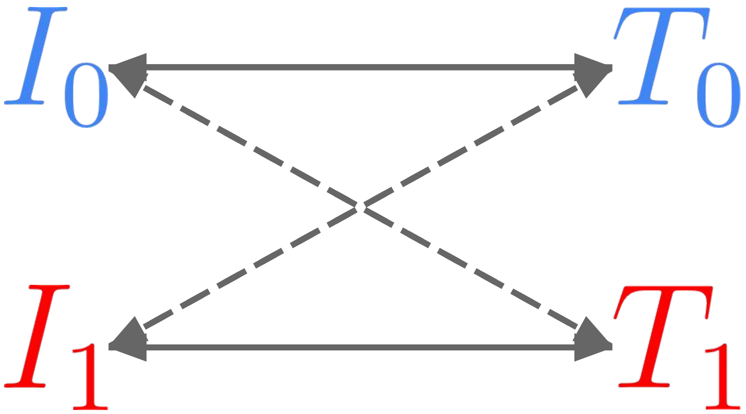


Background: The Winoground Visuolinguistic Compositionality Benchmark

Text Score = $\mathbb{I}[M(I_0, T_0) > M(I_0, T_1)] \wedge \mathbb{I}[M(I_1, T_1) > M(I_1, T_0)]$

Image Score = $\mathbb{I}[M(I_0, T_0) > M(I_1, T_0)] \wedge \mathbb{I}[M(I_1, T_1) > M(I_0, T_1)]$

Group Score = Text Score and Image Score are both 1



"An old person kisses a young person."

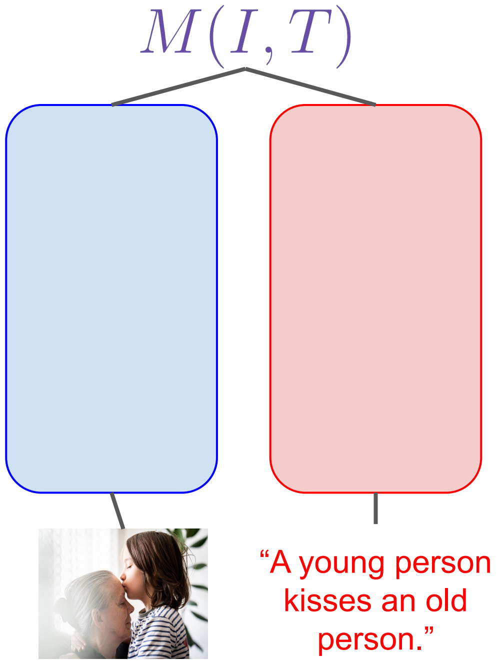
"A young person kisses an old person."

Talk Overview

1. Background: Winoground

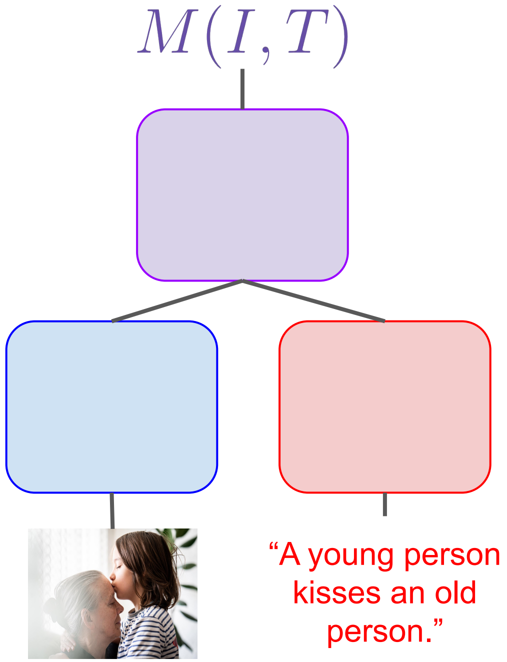
2. **Models of Interest (CLIP, UNITER, LXMERT) and Winoground**

Models of Interest



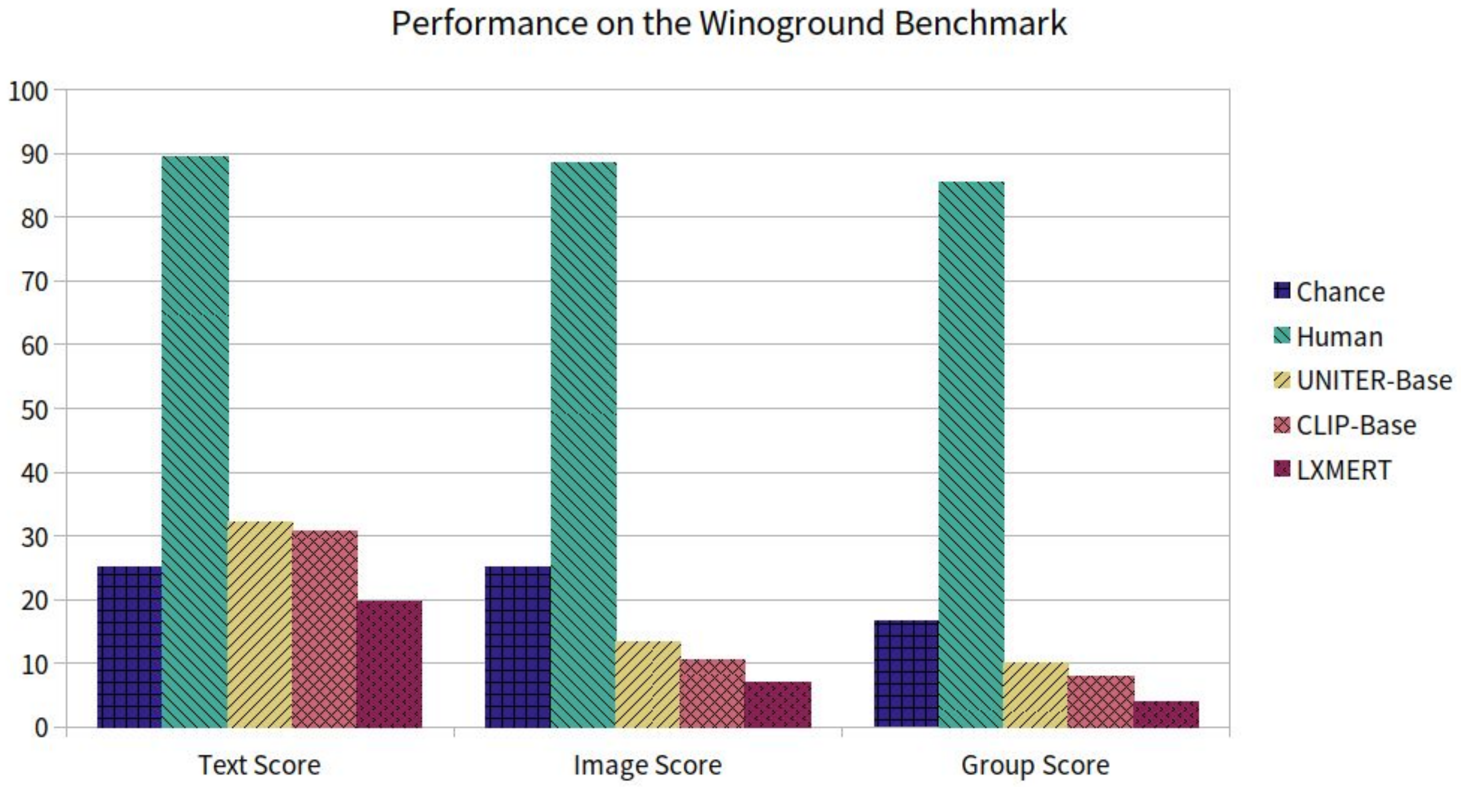
CLIP
151M parameters
400M image-caption pairs
(Radford & Kim et al., 2021)

UNITER
86M parameters
4.2M images; 9.58M captions
(Chen, Li, & Yu et al., 2020)



LXMERT
207M parameters
0.18M images, 9.18M captions
(Tan & Bansal, 2019)

SOTA VL Models Fail Miserably on Winoground



Talk Overview

1. Background: Winoground
2. Models of Interest (CLIP, UNITER, LXMERT) and Winoground

3. Analyzing the dataset

Analyzing the dataset: New annotated tags!

(A) The original Winoground task...

original

the cat on the left of the photo has its right paw ahead of its left

the cat on the left of the photo has its left paw ahead of its right

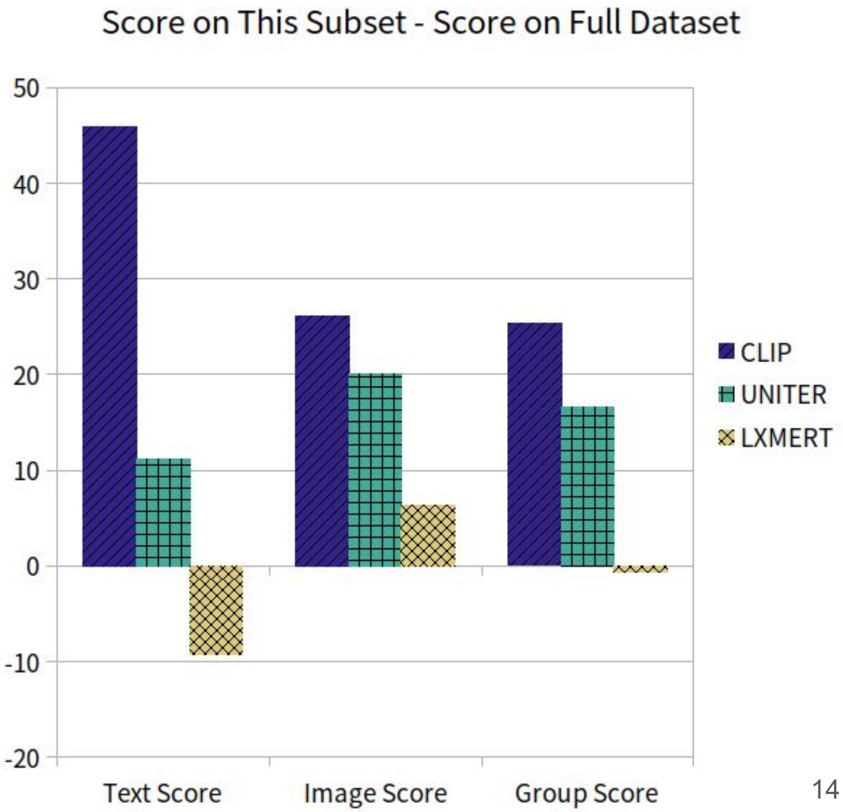
(B) With new tags

NonCompositional	
AmbiguouslyCorrect	
VisuallyDifficult	✓
UnusualImage	
UnusualText	
ComplexReasoning	✓

Non-Compositional Items (n=30)

“Shedding its leaves.”

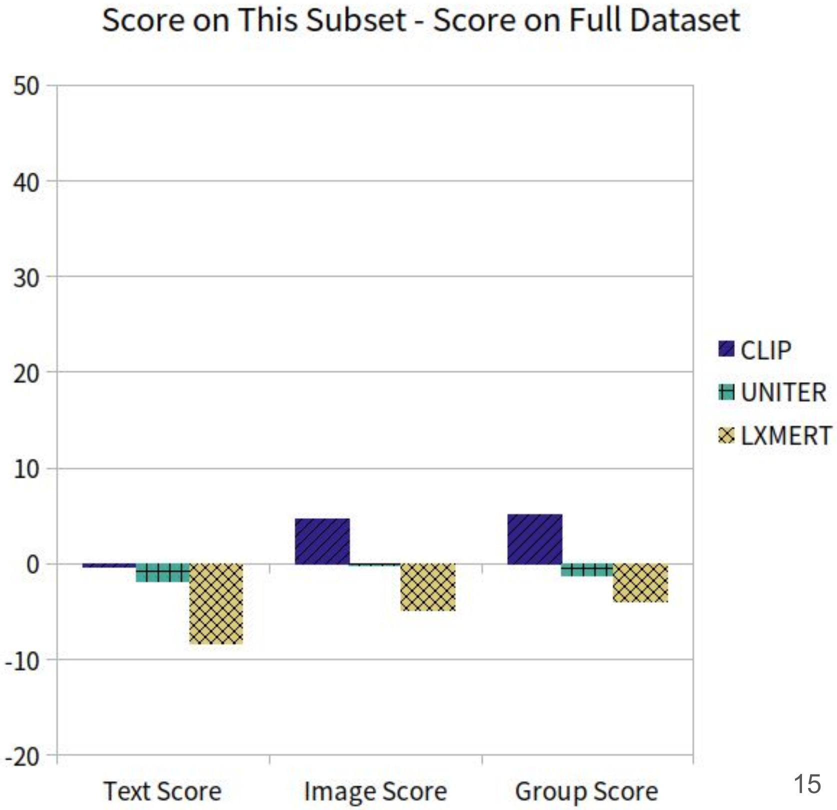
“Leaves its shedding.”



Ambiguously Correct Items (n=46)

“The person with the kids is sitting.”

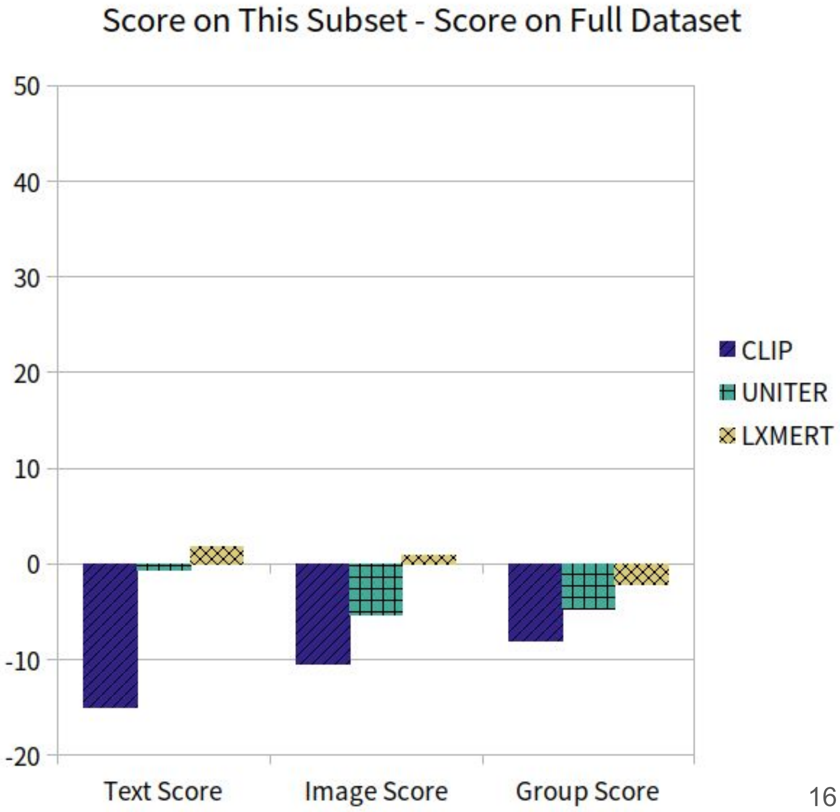
“The person is sitting with the kids.”



Visually Difficult Items (n=38)

"The person with hair to their shoulders has brown eyes and the other person's eyes are blue."

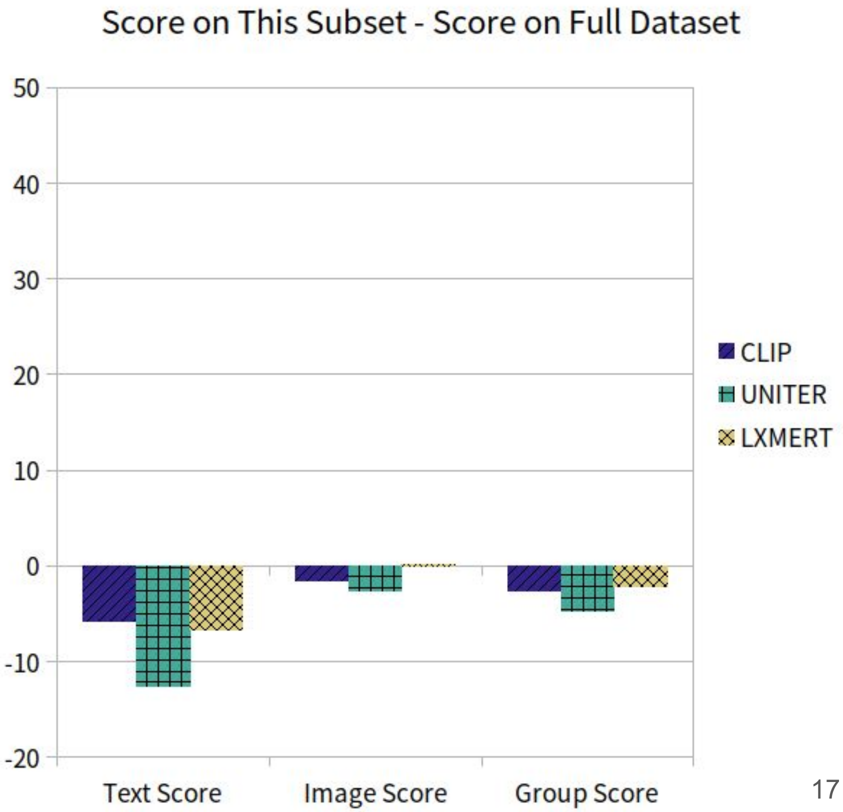
"The person with hair to their shoulders has blue eyes and the other person's eyes are brown."



Items with Unusual Images (n=56)

“The orange lollipop is sad and the red lollipop is surprised.”

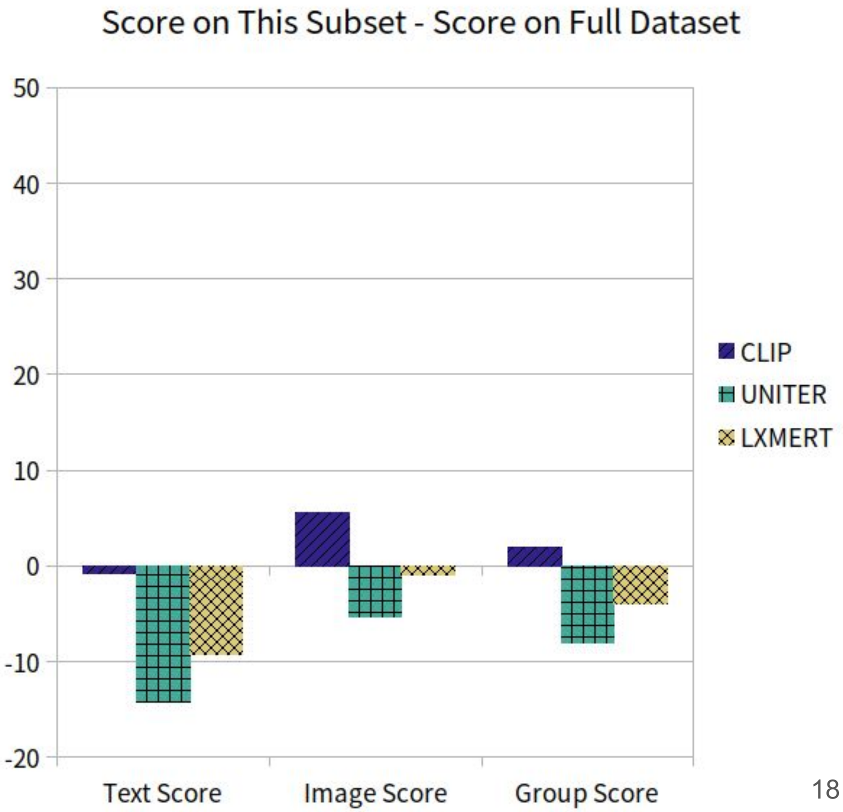
“The orange lollipop is surprised and the red lollipop is sad.”



Items with Unusual Text (n=50)

"The brave in the face of fear."

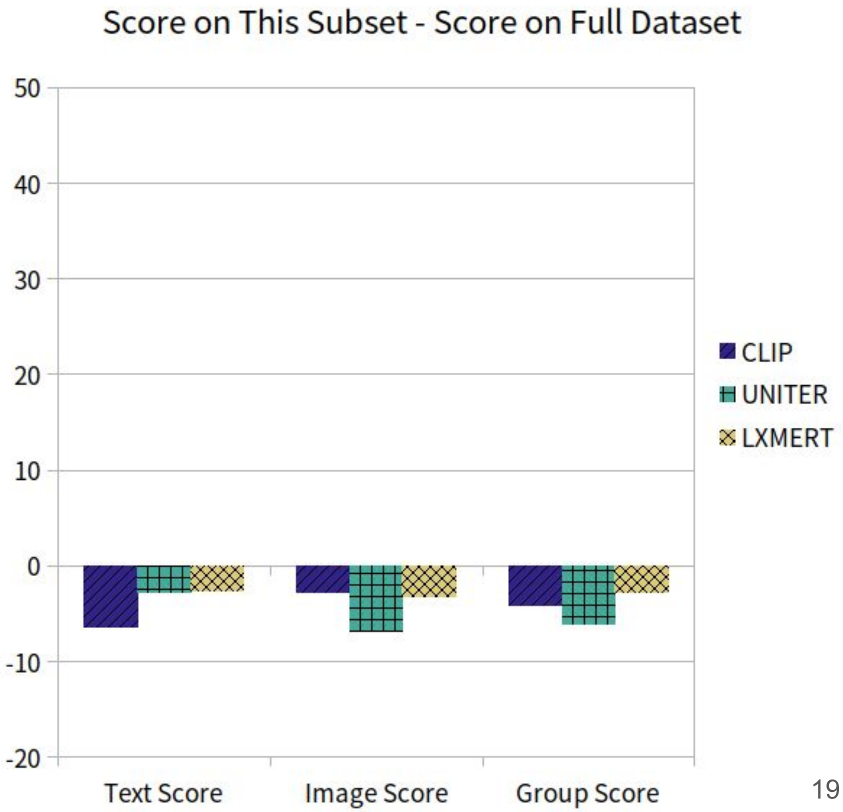
"Fear in the face of the brave."



Items Requiring Complex Reasoning (n=78)

"The cup on the left is filled first and the cup on the right is filled second."

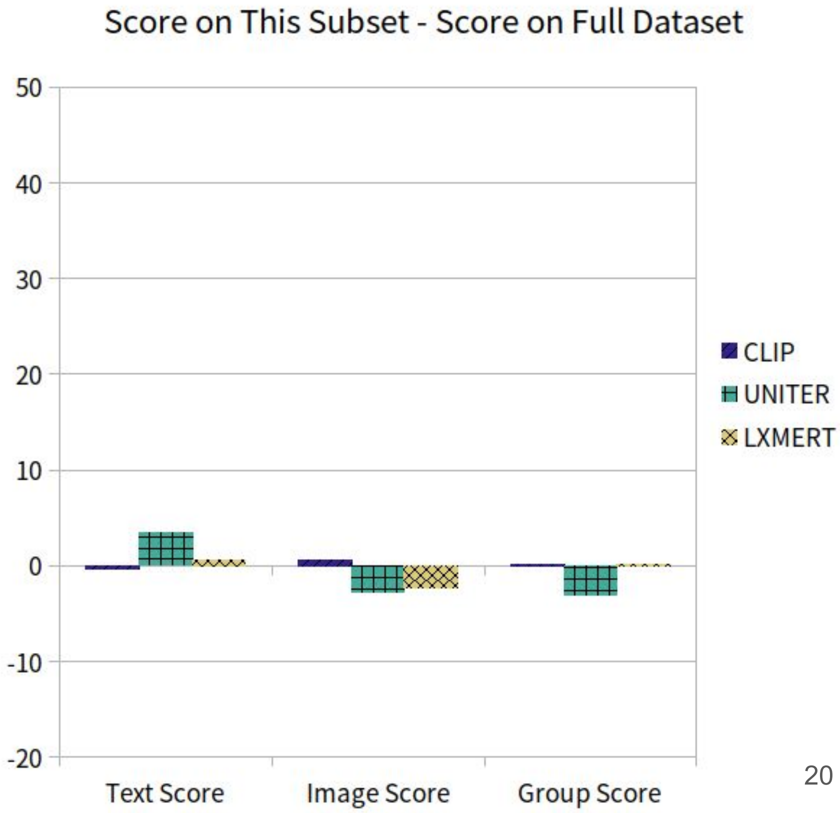
"The cup on the left is filled second and the cup on the right is filled first."



Items Directly Measuring Compositionality (n=171)

"There is a mug in some grass."

"There is some grass in a mug."



Talk Overview

1. Background: Winoground
2. Models of Interest (CLIP, UNITER, LXMERT) and Winoground
3. **Analyzing the dataset**
    a. **Takeaway**: Winoground dataset measures harder/different abilities than just compositionality

Talk Overview

1. Background: Winoground
2. Models of Interest (CLIP, UNITER, LXMERT) and Winoground
3. Analyzing the dataset

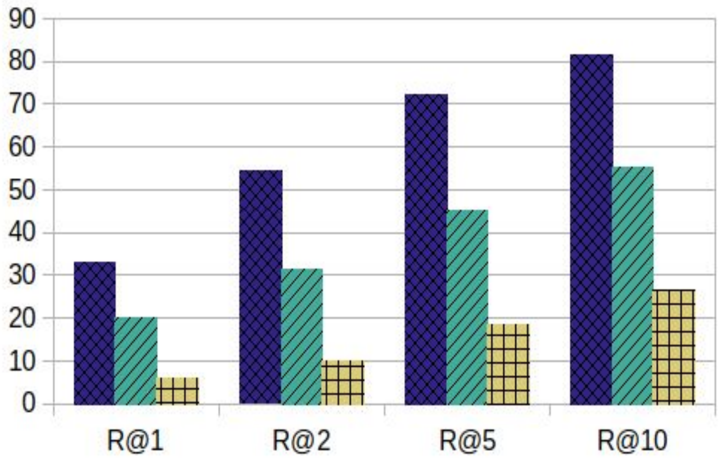
4. Analyzing the evaluation criteria

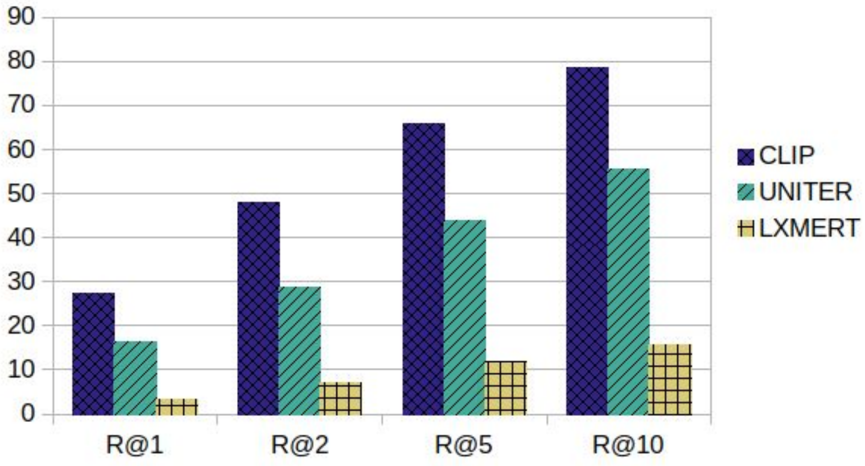
Analyzing the evaluation criteria

We relax evaluation criteria in two ways; 1. **Recall @ k** and 2. **Finetuning probes**

1. Instead of picking I_0 over I_1 conditioned on T_0 ("Image score"), can the model simply retrieve I_0 from the dataset, conditioned on T_0? (Recall @ k)
2. Models only see one image-text pair at a time when outputting score $M(I, T)$ and can't *compare* across pairs. Does training a probe on Winoground that has such access help?

Retrieval: Recall @ k





Recall @ *k* (T2I) = % of texts for which the correct image match is in the top *k* retrievals

Recall @ *k* (I2T) = % of images for which the correct text match is in the top *k* retrievals

Training a probe on Winoground

Target task: Train a single **non-linear** binary classification probe that takes two inputs:

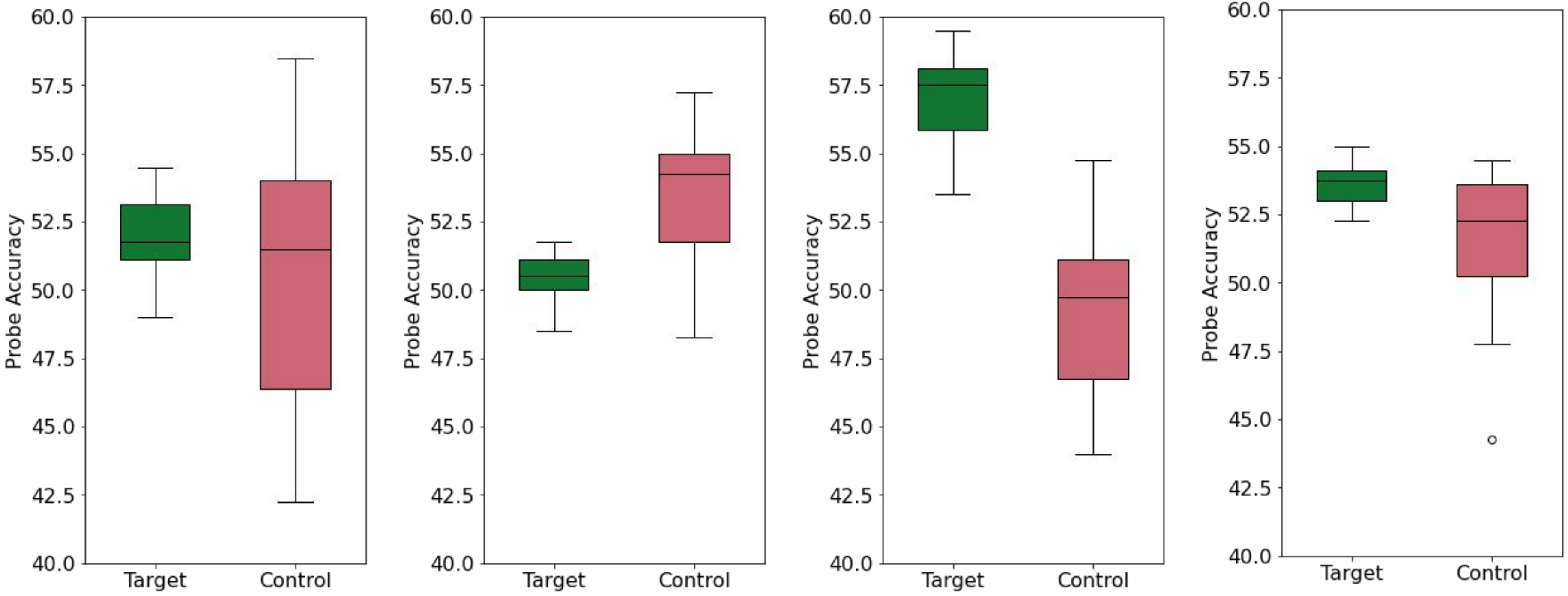
1. Joint embedding of Correct Pair (e.g. I_0, T_0)
2. Joint embedding of Incorrect Pair (e.g. I_1, T_0)

and must output the correct choice (class 0 here)

Control task ('Random baseline'): Same as above but trained with labels swapped for a random 50% of the dataset

Dataset: Winoground (400 examples) split into train set (300) and test set (100)

Training a probe on Winoground: Results (11 trials)



Text Score Probe | Image Score Probe | Text Score Probe | Image Score Probe

LXMERT | UNITER

Talk Overview

1. Background: Winoground
2. Models of Interest (CLIP, UNITER, LXMERT) and Winoground
3. Analyzing the dataset
4. **Analyzing the evaluation criteria**
    a. **Takeaway 1**: Relaxing the strict matching criterion in Winoground reveals new, interesting differences between models
    b. **Takeaway 2**: Surprisingly, training probes on Winoground doesn't seem to help performance

Talk Overview

1. Background: Winoground
2. Models of Interest (CLIP, UNITER, LXMERT) and Winoground
3. Analyzing the dataset
4. Analyzing the evaluation criteria

5. **Analyzing the models**

Analyzing the models

One potential hypothesis is that the text branch of V-L models is *confused* by these minimal textual pairs and cannot semantically distinguish them.

By using *semantics-preserving augmentations* of each text, we found that

1. The text branch actually *can* distinguish these pairs, but
2. Explicitly using this information still doesn't help performance on Winoground

Semantics-preserving augmentations

- We manually select 9 augmentation strategies from **NLAugmenter** (Dhole et.al 2021) that we found are most likely to preserve caption semantics
- Augmented captions i.e. caption variants are *no longer* minimal textual pairs.

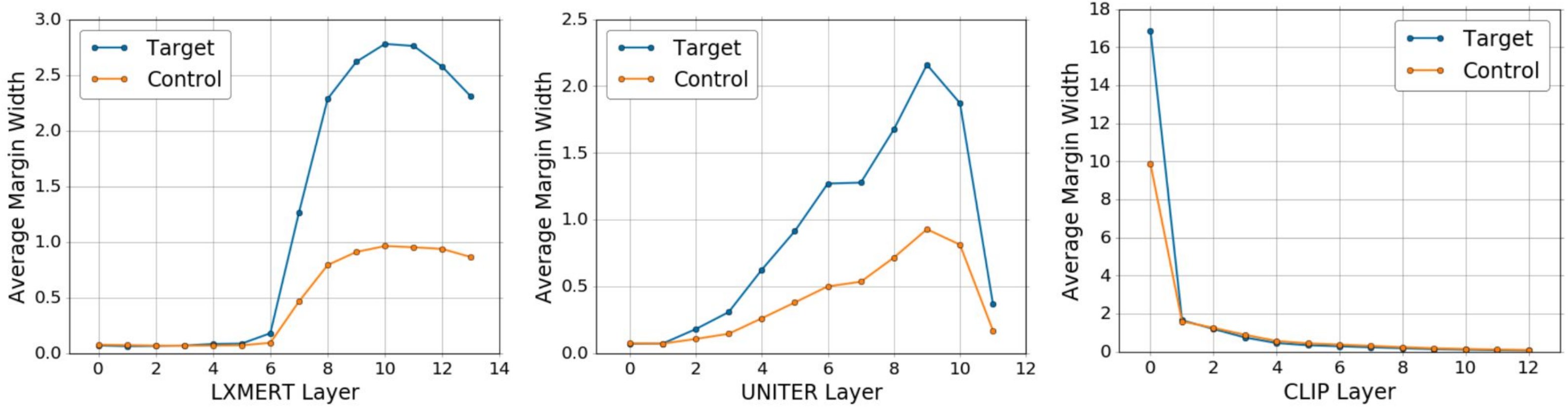
Augmentation	Example Sentence
Original Sentence (1): no changes from Winoground	a human viewing a cat on a screen
Hyponyms (2): replace noun with hyponym, from CheckList (Ribeiro et al., 2020)	a human viewing a lion on a screen
Hypernyms (2): replace noun with hypernym, from CheckList (Ribeiro et al., 2020)	a human viewing a device on a screen
SynonymSubstitution (3): replace word with WordNet (Miller, 1998) synonym	a human view a cat on a screen
Slangificator (3): replaces a word with a slang word from a curated word list	a human viewing a moggie on a screen
Backtranslation (1): translate to German and back using FSMT (Ng et al., 2019)	a human looking at a cat on a screen
DiverseParaphrase (3): diverse paraphrases (Kumar et al., 2019)	what is it like to look at a cat on screen
ProtAugmentDiverseParaphrase (5): diverse paraphrases (Dopierre et al., 2021)	a person who looks at a cat on a screen
Syntactic (3): use hardcoded syntactic rules to generate text with a new word order but same semantics using the AllenNLP of SRL BERT (Shi and Lin, 2019)	a human viewing on a screen a cat

Can models distinguish caption variants?

*Per-item **linear** separability using SVMs*

For each Winoground example (400 in total), learn a **separate** SVM linear classifier...

- Target task: between embeddings of caption 0 variants and caption 1 variants
- Control task: between 2 random, disjoint subsets of the union of caption 0 and caption 1 variants

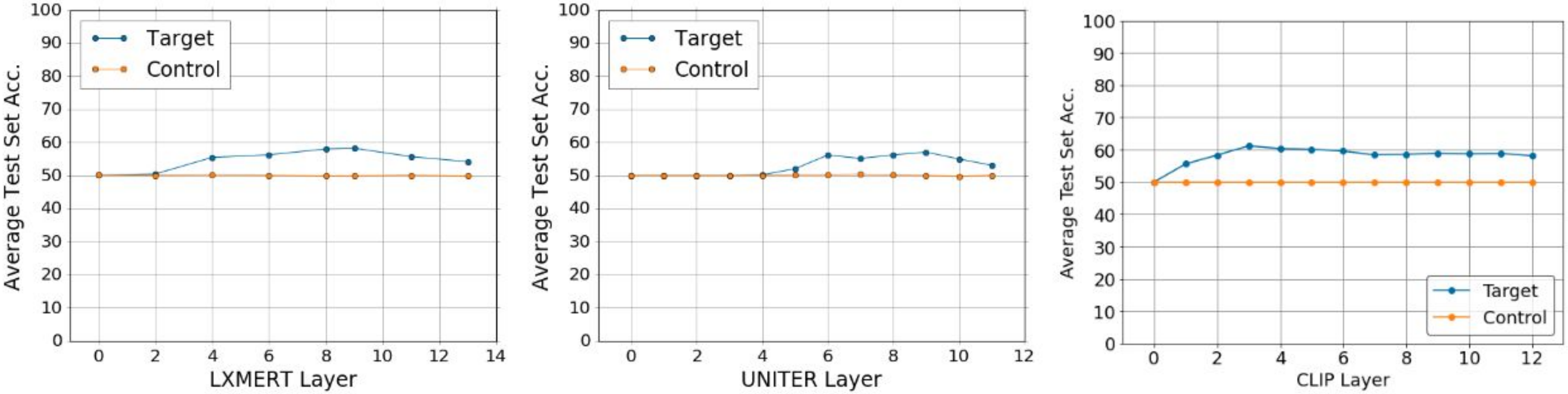


Can models distinguish caption variants?

*All-item **non-linear** separability using probes*

Target: Train a single **non-linear** probe that is given three inputs: a) 2 text embeddings of variants *X* and *Y* of the same caption and b) a text embedding of variant *Z* of a different caption and must correctly choose Y over Z.

Control: The same as above, but train it with 50% of the above matchings swapped



Using Caption Variants to Help Models

- If models can tell caption variants apart, maybe that information can be **used?**
- Use similarity scores between images and *caption variants* to aid models:
  - Given a caption T and its variants $\{T_1, T_2, \ldots, T_n\}$ compute new similarity score

$$S(I,T) = (1-\lambda)S(I,T) + \lambda \text{agg}(S(I,T_i))$$

weighting *original score* *mean/max of new scores*

- This doesn't change text/image/group scores by much, implying that good *semantic distinguishability* may not be sufficient to achieve good *image-text matching*

Talk Overview

1. Background: Winoground
2. Models of Interest (CLIP, UNITER, LXMERT) and Winoground
3. Analyzing the dataset
4. Analyzing the evaluation criteria
5. Analyzing the models
    a. **Takeaway 1**: Models' text branches can semantically distinguish the minimal textual pairs, but
    b. **Takeaway 2**: Models don't seem to be able to use this to do Winoground-style image-text matching

Summary

- We created new annotations that revealed that more abilities are needed to succeed on Winoground than just compositionality

Summary

- We created new annotations that revealed that more abilities are needed to succeed on Winoground than just compositionality
- We relaxed evaluation criteria using a) Recall @ k, revealing interesting differences between the 3 models and b) training probes, that didn't help

Summary

- We created new annotations that revealed that more abilities are needed to succeed on Winoground than just compositionality
- We relaxed evaluation criteria using a) Recall @ k, revealing interesting differences between the 3 models and b) training probes, that didn't help
- We finally showed that models are able to semantically distinguish the two captions using caption variants and linear/non-linear probes, but are likely unable to use such knowledge to succeed on Winoground

Recommendations for the Future

- To get a better idea of model performance, evaluate separately on each of our tag's subsets

Recommendations for the Future

- To get a better idea of model performance, evaluate separately on each of our tag's subsets
- In our model analysis, we only showed that the text branch is able to encode semantic distinctions; outstanding questions include

Recommendations for the Future

- To get a better idea of model performance, evaluate separately on each of our tag's subsets
- In our model analysis, we only showed that the text branch is able to encode semantic distinctions; outstanding questions include
    - Does the image branch encode semantic distinctions?

Recommendations for the Future

- To get a better idea of model performance, evaluate separately on each of our tag's subsets
- In our model analysis, we only showed that the text branch is able to encode semantic distinctions; outstanding questions include
    - Does the image branch encode semantic distinctions?
    - Is the image-text matching score capable of making finegrained distinctions to succeed on Winoground?

Recommendations for the Future

- To get a better idea of model performance, evaluate separately on each of our tag's subsets
- In our model analysis, we only showed that the text branch is able to encode semantic distinctions; outstanding questions include
  - Does the image branch encode semantic distinctions?
  - Is the image-text matching score capable of making finegrained distinctions to succeed on Winoground?
  - How can we train these pretrained models to be better at Winoground?

Recommendations for the Future

- To get a better idea of model performance, evaluate separately on each of our tag's subsets
- In our model analysis, we only showed that the text branch is able to encode semantic distinctions; outstanding questions include
  - Does the image branch encode semantic distinctions?
  - Is the image-text matching score capable of making finegrained distinctions to succeed on Winoground?
  - How can we train these pretrained models to be better at Winoground?

Github:
ajd12342/why-winoground-hard

ArXiv:
2211.00768