How Pre-trained Word Representations Capture Commonsense Physical Comparisons

Pranav Goel, Shi Feng, Jordan Boyd-Graber

November 3, 2019

Computer Science at University of Maryland

Motivation and Context

Why Should You Care?

1. Understanding common sense is important for effective natural language reasoning.

Why Should You Care?

1. Understanding common sense: important for effective natural language reasoning.
2. Pre-trained word representations: ubiquitous in NLP.

Why Should You Care?

1. Understanding common sense: important for effective natural language reasoning.
2. Pre-trained word representations: ubiquitous in NLP.
3. Pre-trained word representations: probed for linguistic properties.

Why Should You Care?

1. Understanding common sense: important for effective natural language reasoning.
2. Pre-trained word representations: ubiquitous in NLP.
3. Pre-trained word representations: probed for linguistic properties.
4. **What do these representations learn about the world?**

Why Should You Care?

1. Understanding common sense: important for effective natural language reasoning.
2. Pre-trained word representations: ubiquitous in NLP.
3. Pre-trained word representations: probed for linguistic properties.
4. **What do these representations learn about the world? - Probe them for commonsense reasoning!**

Goals

1. Do pre-trained representations encode common sense?[1]
2. If yes, how?

[1] A specific type...

Our Focus: Setting the Scope

1. Type of Common Sense:
    - **Commonsense Physical Comparisons**: How two objects compare on physical properties such as size and weight.

Our Focus: Setting the Scope

1. Type of Common Sense:
    - **Commonsense Physical Comparisons**: How two objects compare on physical properties such as size and weight.
    - 'Is a house bigger than a person?'

Our Focus: Setting the Scope

1. Type of Common Sense:
    - **Commonsense Physical Comparisons**: How two objects compare on physical properties such as size and weight.
    - 'Is a house bigger than a person?'
2. Pre-trained Word Representations:

Our Focus: Setting the Scope

1. Type of Common Sense:
    - **Commonsense Physical Comparisons**: How two objects compare on physical properties such as size and weight.
    - 'Is a house bigger than a person?'
2. Pre-trained Word Representations:
    - GloVe
    - ELMo
    - BERT

Probing Task and Dataset

Probing Task and Data

Dataset and Setup: Verb Physics [Forbes and Choi, 2017]

Probing Task and Data

Dataset and Setup: Verb Physics [Forbes and Choi, 2017]

Given: pair of words/objects; **Predict:** $word_1 < / > / \approx word_2$
(compared for a specific attribute)

Probing Task and Data

Dataset and Setup: Verb Physics [Forbes and Choi, 2017]

Given: pair of words/objects; **Predict:** $word_1 < / > / \approx word_2$ (compared for a specific attribute)

$bed >^{weight} hand$, $mouth \approx^{size} fist$, etc.

Verb Physics

Attributes: *size*, *weight*, *strength*, *rigidness*, and *speed*

Verb Physics

Attributes: *size*, *weight*, *strength*, *rigidness*, and *speed*

Split: (5:45:50); training $\sim$ 100 comparisons, dev $\sim$ 1000 comparisons

Generalization Test

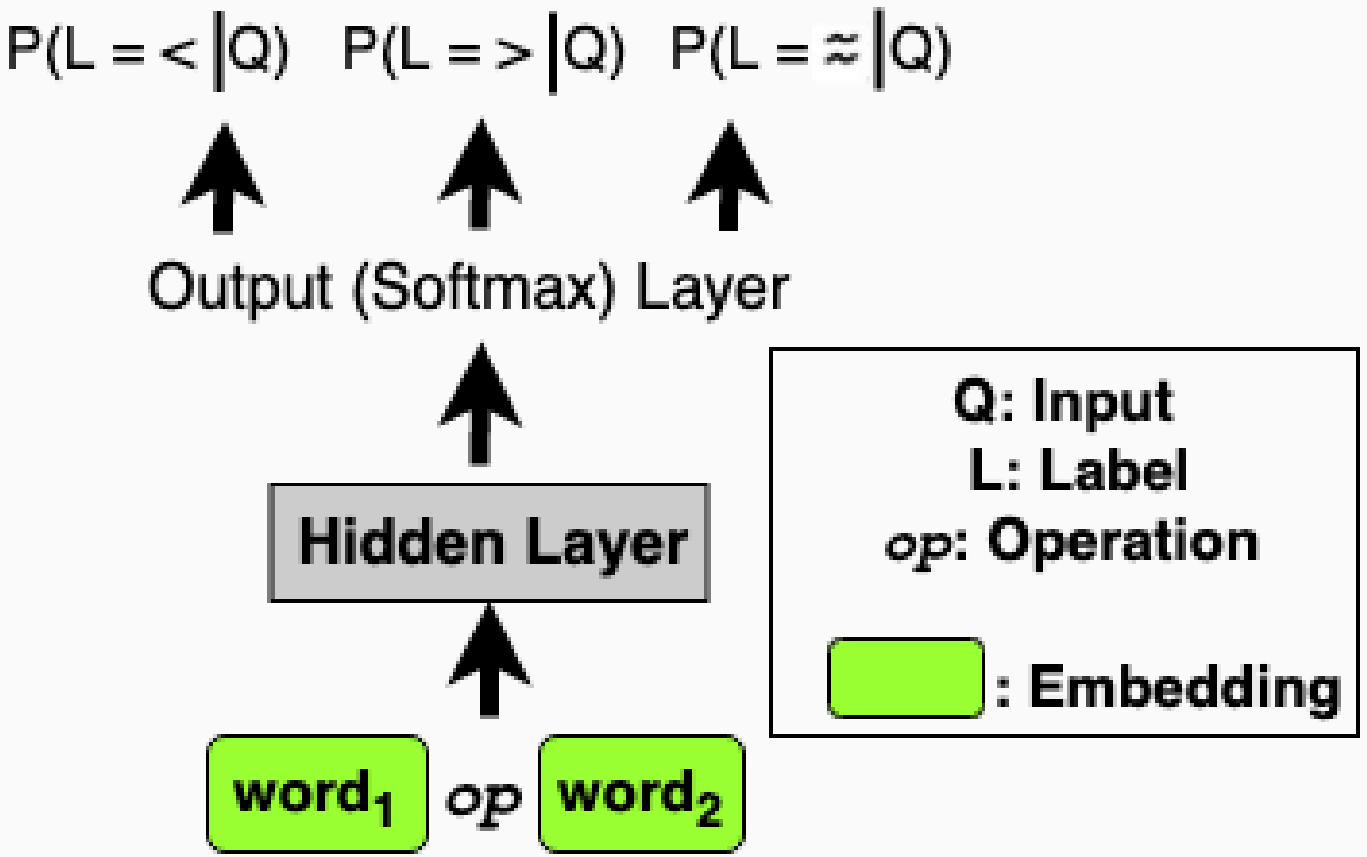
To test generalization to words not seen during training: A different evaluation [Bagherinezhad et al., 2016] - 486 size-based comparisons of objects.

Methodology

Our Probing Model

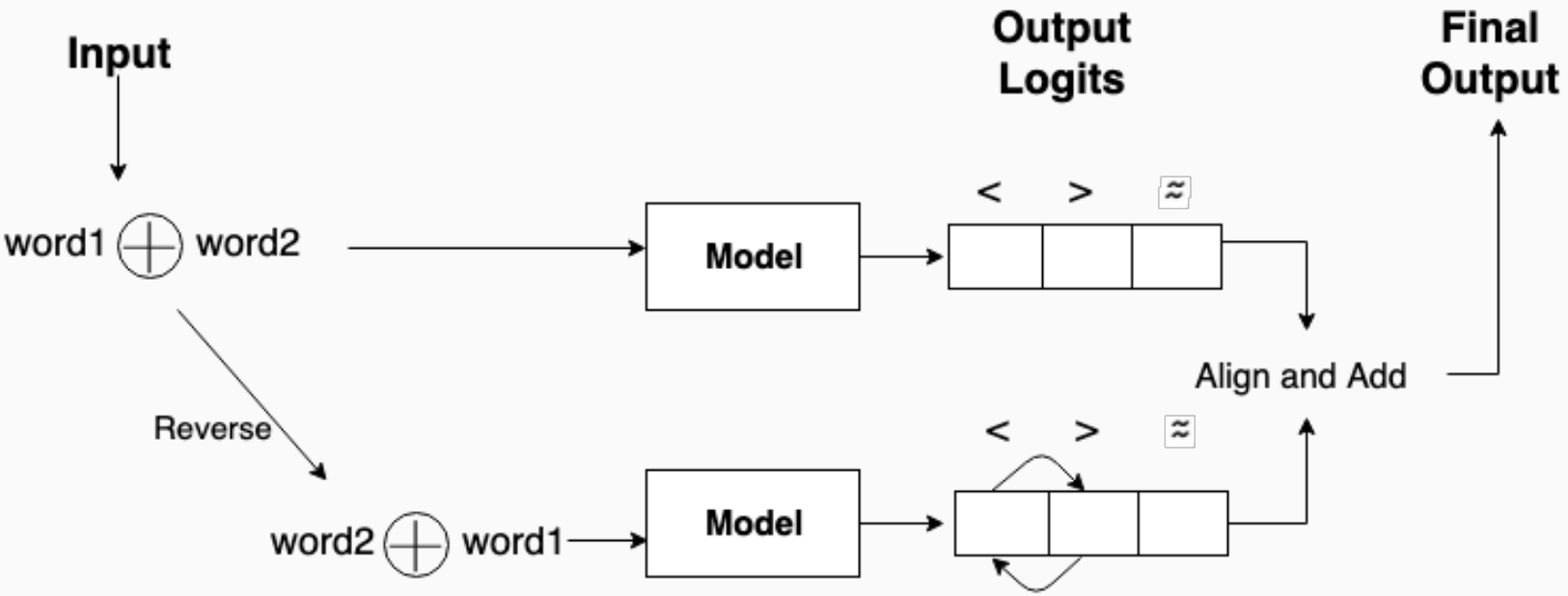
A simple setup to assess if pre-trained representations capture physical object comparisons.

Our Probing Model



The Probing Model: We combine the pre-trained word embeddings of the two words being compared (via concatenation or subtraction) and pass it though zero (linear) or one hidden layer.

Our Probing Model



The Reversal Trick: At test time, the reversed embedding is also passed through the network and the output logits for both pairs ($word_1$ concatenated with $word_2$ and $word_2$ concatenated with $word_1$) are aligned and combined for the final output (following [Yang et al., 2018]). We try doing this at training time as well which leads to an improvement in accuracy.

Baselines

Baseline 1: Majority Class

Baselines

Baseline 1: Majority Class

Predict based on the *highest-frequency label* for both the words (using *training set*).

Baselines

Baseline 1: Majority Class

Predict based on the *highest-frequency label* for both the words (using *training set*).

Case 1: The two labels *agree*

elephant $>^{size}$ *x* AND *mouse* $<^{size}$ *y* for most *x* and *y* in training set $\implies$ *elephant* $>^{size}$ *mouse*

Baselines

Baseline 1: Majority Class

Predict based on the *highest-frequency label* for both the words (using *training set*).

Case 2: The two labels *DISAGREE*

$bed >^{weight} x$ AND $hand >^{weight} y$ for most x and y in training set:

Compute ratios: $X_1 = \frac{\sum_{(bed,x) \in training} bed >^{weight} x}{\sum_{(bed,z) \in training}}; X_2 = \frac{\sum_{(hand,y) \in training} hand >^{weight} y}{\sum_{(hand,z) \in training}}$

$X_1 > X_2 \implies bed >^{weight} hand$

Baselines (Previous SOTA on Verb Physics)

Baseline 2: Verb-centric Frame Semantics (F&C)

Baselines (Previous SOTA on Verb Physics)

Baseline 2: Verb-centric Frame Semantics (F&C)

[Forbes and Choi, 2017, **F&C**]

- Probabilistic graphical modeling
- Joint inference over objects AND actions/verbs
- 'x *entered* y' $\implies y >^{size} x$

Baselines (Previous SOTA on Verb Physics)

Baseline 3: Property Comparisons from Embeddings (PCE)

Baselines (Previous SOTA on Verb Physics)

Baseline 3: Property Comparisons from Embeddings (PCE)

[Yang et al., 2018, **PCE**]

- Similar to our model (one-layer neural network over concatenated word embeddings)
- Important Difference: Compare the projection with the embeddings of 'poles': words exemplifying a physical relation ('big', 'small' for size; 'fast', 'slow' for speed, etc.).
- Classification is the closest 'pole'. *Use of poles is the main difference with our approach.*

Major Differentiating Point for Our Probing Model

Major Differentiating Point for Our Probing Model

Our probing model uses ONLY the two words being compared...

... Previous Approaches use more/extra information!

Results

Results: Accuracy On the Verb Physics Dataset

	Majority Class Baseline	F&C	PCE	Probing Model (GloVe)	Probing Model (ELMo)	Probing Model (BERT-base)
Size	0.66	0.75	0.80	**0.82**	**0.82**	0.80
Weight	0.67	0.74	0.81	**0.82**	**0.82**	0.80
Strength	0.66	0.71	0.77	0.78	**0.79**	0.75
Rigidness	0.60	0.68	0.71	0.71	**0.72**	0.71
Speed	0.59	0.66	0.72	0.72	**0.76**	0.71
Overall	0.64	0.71	0.76	0.77	**0.78**	0.75

Table 1: The simple probing model achieves better accuracy than previous approoaches which use extra information in addition to the words being compared. *This indicates pre-trained representations capture commonsense physical comparisons.*

Do pre-trained representations REALLY capture commonsense physical comparisons?

Generalization to New Objects

Handling Unseen Words: The Need

Verb Physics: ∼99% of the words or objects involved in comparisons in the dev set are seen in the training set.

Handling Unseen Words: The Need

Verb Physics: ∼99% of the words or objects involved in comparisons in the dev set are seen in the training set.

If word embeddings capture common sense well, they should compare two words not seen during training

Handling Unseen Words: Setup

Training: Verb Physics training set for the ‘size’ attribute.

Handling Unseen Words: Setup

Training: Verb Physics training set for the ‘size’ attribute.

Evaluation: A different test set [Bagherinezhad et al., 2016]: **EB evaluation set**.

Only ∼33% of the words are seen during training.

Handling Unseen Words: Results

Model	Accuracy
The Visual+Textual Model by [Bagherinezhad et al., 2016]	0.835
Probing Model (GloVe)	0.879
Probing Model (ELMo)	**0.905**
Probing Model (BERT)	0.893

Table 2: The probing model trained on the Verb Physics size dataset and evaluated on [Bagherinezhad et al., 2016]. **Only ∼33% of the objects in this test set are present in training set**

Handling Unseen Words: Results

Model	Accuracy
The Visual+Textual Model by bagherinezhad2016elephants	0.835
Probing Model (GloVe)	0.879
Probing Model (ELMo)	**0.905**
Probing Model (BERT)	0.893

Table 2: The probing model trained on the Verb Physics size dataset and evaluated on [Bagherinezhad et al., 2016]. **Unlike [Bagherinezhad et al., 2016] who use visual and textual cues, our model use only pre-trained text representations.**

Do pre-trained representations REALLY capture commonsense physical comparisons?

Lexical Memorization?

Baselines Using Just One Word: The Need

[Levy et al., 2015]: For hypernymy detection,

Accuracy(using both words) - Accuracy(using just one word) < 10% !

Baselines Using Just One Word: The Need

[Levy et al., 2015]: For hypernymy detection,

Accuracy(using both words) - Accuracy(using just one word) < 10% !

Prototypical hypernyms: single word in a pair that models can latch onto to detect hypernymy.

Baselines Using Just One Word: The Need

[Levy et al., 2015]: For hypernymy detection,

Accuracy(using both words) - Accuracy(using just one word) < 10% !

Prototypical hypernyms: single word in a pair that models can latch onto to detect hypernymy.

Unsupervised Baseline: Cosine similarity of the two words.

What about Commonsense Comparisons and our Probing Model?

Baselines Using Just One Word: Results

Using the given Verb Physics training set	$word_1$ - $word_2$	ONLY $word_2$ Baseline	Unsupervised Baseline
GloVe	0.78	0.66	0.49
ELMo	0.78	0.67	0.52
BERT	0.75	0.66	0.52

Table 3: Accuracy of probing models (averaged across the five attributes) on the Verb Physics dev sets. Using just one word when training and evaluating helps investigate possible lexical memorization.

Baselines Using Just One Word: Results

Using the given Verb Physics training set	$word_1$ - $word_2$	ONLY $word_2$ Baseline	Unsupervised Baseline
GloVe	0.78	0.66	0.49
ELMo	0.78	0.67	0.52
BERT	0.75	0.66	0.52

Table 3: Accuracy of probing models (averaged across the five attributes) on the Verb Physics dev sets. Only $word_2$ seems to be a strong baseline (much like the majority class baseline for this dataset), but the drop in accuracy is higher than 10% for GloVe and ELMo: *Our model is not simply relying on lexical memorization.*

Baselines Using Just One Word: Results

Using the given Verb Physics training set	$word_1$ - $word_2$	ONLY $word_2$ Baseline	Unsupervised Baseline
GloVe	0.78	0.66	0.49
ELMo	0.78	0.67	0.52
BERT	0.75	0.66	0.52

Table 3: Accuracy of probing models (averaged across the five attributes) on the Verb Physics dev sets. Unsupervised baseline takes cosine similarity of the embeddings and uses a threshold tuned on the dev set to classify - low accuracy suggests *supervision is helpful.*

Baselines Using Just One Word: Results

On the Complete EB Evaluation Set; ∼33% 'overlap'	$word_1$ - $word_2$	$word_1$	$word_2$
GloVe	0.88	0.74	0.73
ELMo	0.89	0.74	0.72
BERT	0.87	0.65	0.68

Table 4: Evaluation on [Bagherinezhad et al., 2016]. **Accuracy drops by 15 to 20% when compared with the only one word baselines.**

Pre-trained representations capture commonsense physical comparisons: HOW?

‘Local’ ordering and the potential role of logit scores

Local Ordering formed via Logit Difference

A particular word gets compared with many other words in data.

chair $<$ room, chair $<$ house

chair $>$ head, chair $>$ knee

Local Ordering formed via Logit Difference

A particular word gets compared with many other words in data.

chair $<$ room, chair $<$ house

chair $>$ head, chair $>$ knee

How to use all the comparisons for ‘chair’ and form a (local) ordering around this word?

Local Ordering formed via Logit Difference

A particular word gets compared with many other words in data.

chair $<$ room, chair $<$ house

chair $>$ head, chair $>$ knee

How to use all the comparisons for 'chair' and form a (local) ordering around this word?

Intuition: *Humans are* ***more confident*** *about a comparison when the* ***difference in objects in terms of the property is large*** *(a house is bigger than a chair).*

Local Ordering formed via Logit Difference

Intuition: *Humans are **more confident** about a comparison when the **difference in objects in terms of the property is large** (a house is bigger than a chair).*

For the model? Larger difference in output logits (for label 0 ($<$) and 1 ($>$)) $\implies$ more model confidence $\implies$ objects being farther apart in an ordering.

For e.g.,

Input = (chair, room), Prediction = $<$, Logit Score Difference ('$<$' - '$>$') = 0.8

Input = (chair, house), Prediction = $<$, Logit Score Difference = 0.95

Input = (chair, head), Prediction = $>$, Logit Score Difference ('$>$' - '$<$') = 0.95

Input = (chair, knee), Prediction = $>$, Logit Score Difference = 0.85

Ordering: head $<$ knee $<$ chair $<$ room $<$ house

Local Ordering formed via Logit Difference

Consistency of A Local Ordering:

Orderings are *consistent* if the same pair of words in different local orderings hold the same relationship.

Local Ordering formed via Logit Difference

Examples of Orders Formed Around a Word
head $<$ knee $<$ meal $<$ ***chair*** $<$ back $<$ place $<$ street $<$ world $<$ *gate* $<$ air $<$ floor $<$ *room*
eye $<$ *chair* $<$ child $<$ king $<$ daughter $<$ wife $<$ boy $<$ messenger $<$ father $<$ coach $<$ horse $<$ door $<$ house $<$ ***gate*** $<$ train $<$ *room* $<$ sun

Table 5: Two examples for orderings formed around the words ***chair*** and ***gate*** for the size attribute using GloVe. Comparisons between words occurring in both these orderings (italicized) are consistent.

Local Ordering formed via Logit Difference

All the local orderings formed around all words on Verb Physics are completely consistent for GloVe and BERT. For ELMo, more than 90% comparisons were usually consistent across any two orderings.

Local Ordering formed via Logit Difference

All the local orderings formed around all words on Verb Physics are completely consistent for GloVe and BERT. For ELMo, more than 90% comparisons were usually consistent across any two orderings.

Models seem to learn to arrange all the words in some sort of consistent ordering.

Pre-trained representations capture commonsense physical comparisons: HOW?

‘Global’ ordering on all words using learned weights of probing model

Global Ordering over all Words Using Learned Weights

We use a *linear* model to order all the objects in one of the Verb Physics dev sets (we confirmed linear modela are almost at par (accuracy within 1%) with shallow fully connected neural networks on the Verb Physics dev set).

Global Ordering over all Words Using Learned Weights

We use a *linear* model to order all the objects in one of the Verb Physics dev sets (we confirmed linear modela are almost at par (accuracy within 1%) with shallow fully connected neural networks on the Verb Physics dev set).

A score for a word is its embedding multiplied with the weight learned for mapping the input to the label 1 which would be higher if $word_1 > word_2$. We use this score to rank the objects.

Global Ordering over all Words Using Learned Weights

scissors < beard < spoon < hair < knife < finger < lip < purse < chin < goose < vial < eye < nose < bow < fist < piece < ash < glass < chair < skirt < grass < picture < head < face < hat < gulp < bag < ear < hand < strap < dress < bottle < torso < elbow < edition < mouth < pocket < arm < shoulder < rope < magazine < tear < seal < hedge < effect < violin < tree < knee < lamp < cup < pedestal < throat < book < coal < object < suit < button < ball < chest < magistrate < newspaper < fox < ice < candidate < harlot < basin < mosquito < meal < bower < foot < shirt < step < child < stone < body < anchor < clothes < seed < exile < shippe < dinner < trench < element < lung < light < block < poet < sink < king < stair < breath < fool < phone < coward < banker < result < base < response < sip < bench < end < lock < victim < source < torrent < brick < sail < daughter < master < watch < gully < cross < scene < disciple < lady < food < direction < teacher < boy < middle < boat < messenger < parent < precipice < person < call < window < shore < wife < vessel < horse < temple < servant < piano < bed < patient < side < something < parcel < back < way < position < wall < place < lover < wind < state < corner < office < father < prison < worker < volunteer < street < abode < coach < flood < doorway < anything < someone < ground < front < brother < world < horseback < shop < current < city < energy < reservation < friend < camp < store < bank < factory < gentleman < rain < lad < deck < soul < home < beach < everything < floor < clock < car < house < door < ship < heaven < truck < air < system < barn < stream < mountain < restaurant < road < river < sea < bay < gate < hill < coast < farm < town < train < sun < room

Table 6: Example of an Ordering Over all Words in a Set: The trained weights of the linear probing model multiplied with the embedding of a word can help form a 'global' ordering over all the words. This particular example is when the weight corresponding to the label 1 is used for the words in the Verb Physics size dev set with GloVe embeddings.

Global Ordering over all Words Using Learned Weights

Using this ordering to classify the comparisons of pair of words achieves accuracy at par with the original models on a subset of the dev set containing only 0/1 labels.

This suggests the models assign an absolute value to every word to rank all the objects and then *use this global ranking* to compare any two objects.

Global Ordering over all Words Using Learned Weights

Using this ordering to classify the comparisons of pair of words achieves accuracy at par with the original models on a subset of the dev set containing only 0/1 labels.

This suggests the models assign an absolute value to every word to rank all the objects and then *use this global ranking* to compare any two objects.

An ordering can be used directly for $>$ or $<$ comparisons but is not that indicative for $\approx$ comparisons. **This might explain the relative struggles GloVe, ELMo, and BERT face classifying comparisons labeled 2.**

Error Analysis

	0 ($<$)	1 ($>$)	2 ($\approx$)
GloVe	0.79	0.77	0.33
ELMo	0.81	0.80	0.18
BERT	0.77	0.78	0.12

Table 7: Label-Wise Accuracy:
The GloVe, ELMo, and BERT representations (fed to a linear model) struggle to capture the relationship $word_1 \approx word_2$ (label 2).
Possible Reason 1: Class imbalance in the dataset.
Possible Reason 2: The representations seem to learn an ordering over all the words and use it to compare objects (using a global ordering). Judging $\approx$ relationship between words is hard while the $<$ or $>$ relation can be inferred directly from an ordering.
Accuracies here are averaged across the results for all the five attributes.

Conclusion

Takeaways

A linear or a small fully connected neural network probing model can compare two words on commonsense physical attributes using frozen pre-trained representations (GloVe, ELMo, and BERT) of the words alone with higher accuracy than previous approaches which use extra information in addition to the objects being compared.

Takeaways

They also generalize to objects not seen during training (and in fact do better than previous approach using visual as well as textual cues)

Takeaways

Using both the words gets significantly higher accuracy than using just one word: not doing just lexical memorization.

Takeaways

Pre-trained Embeddings seem to encode physical common sense.

Takeaways

Models learn an ordering over of all the words involved in the comparisons and embeddings could be using this ordering to compare any two objects.

Takeaways

The difference in the output logit values corresponding to the labels serves as a surprisingly good proxy to form completely consistent orderings around different words.

Thank you! Questions? Please feel free to reach out at pgoel1@cs.umd.edu

Bagherinezhad, H., Hajishirzi, H., Choi, Y., and Farhadi, A. (2016).
Are elephants bigger than butterflies? reasoning about sizes of objects.
In *Thirtieth AAAI Conference on Artificial Intelligence*.

Forbes, M. and Choi, Y. (2017).
Verb physics: Relative physical knowledge of actions and objects.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276.

Levy, O., Remus, S., Biemann, C., and Dagan, I. (2015).
Do supervised distributional methods really learn lexical inference relations?
In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976.


Yang, Y., Birnbaum, L., Wang, J.-P., and Downey, D. (2018).
Extracting commonsense properties from embeddings with limited human guidance.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 644–649.