Do Neural Language Representations Learn Physical Commonsense?

Maxwell Forbes†, Ari Holtzman†‡, and Yejin Choi†‡
{mbforbes, ahai, yejin}@cs.washington.edu
†Paul G. Allen School of Computer Science and Engineering, University of Washington
‡Allen Institute for Artificial Intelligence

Abstract

Humans understand language based on the rich background knowledge about how the physical world works, which in turn, allows us to reason about the physical world through language. In addition to the *properties* of objects (e.g., *boats require fuel*) and their *affordances*, i.e., the actions that are applicable to them (e.g., *boats can be driven*), we can also reason about *if–then* inferences between what properties of objects imply the kind of actions that are applicable to them (e.g., that *if we can drive something then it likely requires fuel*).

In this paper, we investigate the extent to which state-of-the-art neural language representations, trained on a vast amount of natural language text, demonstrate physical commonsense reasoning. While recent advancements of neural language models have demonstrated strong performance on various types of natural language inference tasks, our study based on a dataset of over 200k newly collected annotations suggests that neural language representations still only learn associations that are explicitly written down.[1]

Keywords: physical commonsense, natural language, neural networks, affordances

Introduction

Understanding everyday natural language communication requires a rich spectrum of physical commonsense knowledge. Consider the example dialog sketched in Figure 1. A simple observation that, *"The blender is broken again!"* triggers myriad pieces of implied understanding (e.g., that something which requires electricity will only work with a source of power). Such knowledge is rarely stated explicitly (Van Durme, 2010), and instead can be inferred on-the-fly as needed.

In this paper, we study physical commonsense knowledge underlying natural language understanding, organized as interactions among three distinct concepts: (i) objects, (ii) their attributes (properties), and (iii) the actions that can be applied to them (affordances) (Figure 1, bottom). The premise of our study is that language models trained on a sufficiently large amount of text can recover a great deal of physical commonsense knowledge about each of these concepts. However, aspects of this knowledge may only be implicit in natural language utterances. For example, answering a question from the Winograd Schema Challenge (Levesque, Davis, & Morgenstern, 2012)—"The trophy would not *fit* in the brown suitcase because it was too *big*. What was too big?"—implicitly requires the physical commonsense reasoning that *"in order to fit X in Y, X should be relatively smaller compared to Y"*, which essentially requires reasoning about the affordances of objects (fit X in Y) from their attributes (relative size of X and Y).

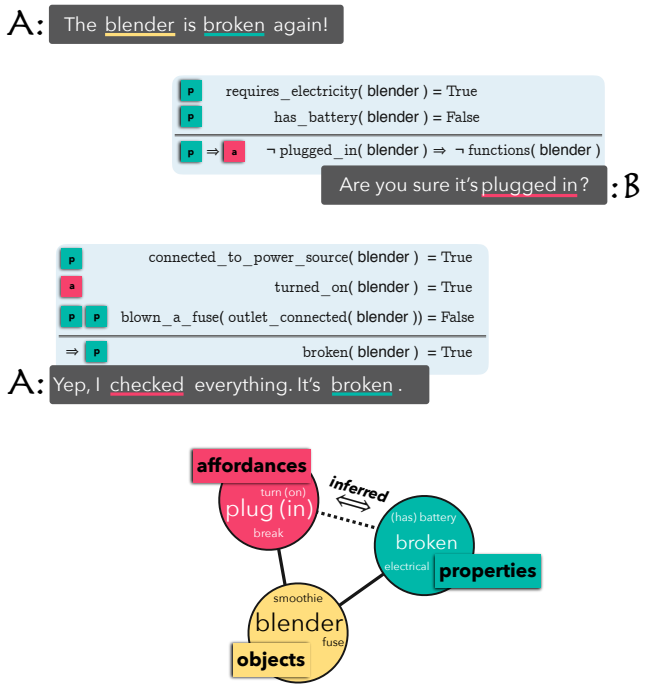

Figure 1: Natural language communication often requires reasoning about the affordances of objects (i.e., what actions are applicable to objects) from the properties of the objects (e.g., what are the size, weights, material of the objects) and vice versa. We study the extent to which neural networks trained on a large amount of text can recover various aspects of physical commonsense knowledge.

In this paper, we investigate the extent to which neural language models trained on a massive amount of text demonstrate various aspects of physical commonsense knowledge and reasoning. Our analysis includes word embeddings such as GloVe (Pennington, Socher, & Manning, 2014), as well as

[1] Visit `https://mbforbes.github.io/physical-commonsense` for our data, code, and more project information.

more recent contextualized representations like ELMo (Peters et al., 2018) and BERT (Devlin, Chang, Lee, & Toutanova, 2018). Such models are trained without supervision by exposing them to billions of words, and allowing them to extract patterns purely from token prediction tasks that can be derived directly from raw text. These language representation models have established unprecedented performance on a wide range of evaluations, including natural language inference and commonsense reasoning.

How much do these large, unsupervised models of language learn about physical commonsense knowledge? Some recent work has studied the capabilities of word embeddings to predict an object's properties (Rubinstein, Levi, Schwartz, & Rappoport, 2015; Lucy & Gauthier, 2017). Motivated by these efforts to understand language representations, we present several contributions. We contribute two datasets: the *abstract* dataset, a refreshed version of the McRate dataset (McRae, Cree, Seidenberg, & McNorgan, 2005), pruned and densely annotated to eliminate false negatives present in previous work; and the *situated dataset*, with annotations for objects' properties and affordances in real-world images sampled from the MS COCO dataset (Lin et al., 2014). As in previous work, we consider the prediction task of linking objects and their properties (O$\longleftrightarrow$P), but with our new situated dataset, we are also able to study the connection between objects and their affordances (O$\longleftrightarrow$A), as well as between affordances and properties (A$\longleftrightarrow$P). We also study the latest models from the natural language processing community (ELMo, BERT) using in-context word representations, and present results for all of our proposed datasets and tasks. Our analysis suggests that current neural language representations are proficient at guessing the affordances and properties of objects, but lack the ability to reason about the relationship between affordances and properties itself.

Characterizing Objects through Properties and Affordances

Properties

We use the term *properties* to refer to the static characteristics of objects. They encompass our commonsense understanding of what something is like. For example, we might say that an *apple* has the property of being *edible*, or that a *plant* is *stationary*.

As with McRae et al. (2005), properties capture the general perception of a thing. Exceptions naturally arise. For example, specific instances can violate the general properties of an object, such as the inediblilty of a rotten apple. Additionally, subtypes can diverge from the exemplar of a category, as with the Venus flytrap, a plant with the ability to move.

Affordances

We express an object's actions with verbs. One way to focus on understanding the actions of objects is to focus on their *affordances*. Coined by Gibson (1966), this term initially described animal-perceived uses for an object, but has since come to mean the perceived uses of an object in a given environment (Norman, 1988; Gaver, 1991).

Here, we take a simpler, human-centric definition. We consider an object's affordances to be, "what actions do humans take with an object?" For example, *boots* commonly afford *wear*, *kick off*, *lace up*, and *put on*.

Inference Between Affordances and Properties

Affordances and properties exhibit a surprising connection. As humans, we are able to infer many of an object's affordances based on its properties (A$\leftarrow$P). The same is also true in the reverse (A$\rightarrow$P).

Consider an exchange: *"You think you could fit that boulder in your truck?" "No way! That thing was so big you could go for a hike on it."* We might sketch out some of this information as:

$$\textit{fit } x \textit{ into } y \implies x <^{\text{size}} y$$
$$\text{hike(x)} \implies x \gg^{\text{size}} \textsc{Human}$$

While the above information only concerns a property's relative value (comparative size), all kinds of information traverse this edge implicitly:

*She **plugged in** her robot.*
$$\text{plug-in}(x) \implies \text{uses-electricity}(x)$$
*He **poured** coffee into the cup*
$$\text{pour-into}(x) \implies \text{holds-liquid}(x)$$
*It **shattered** on the floor.*
$$\text{shatter(x)} \implies \text{rigid}(x)$$

The implications ($\implies$) should be taken with a probabilistic grain of salt. However, they capture our intuitions about what we expect to be true. Wouldn't it be surprising to shatter something that isn't rigid, or plug-in something that doesn't take power?

Humans use the link between affordances and properties to recover information. Can machine learning models do the same? It is is difficult to model these implications based on text alone because there is no direct evidence for the implied information. Any implication that can be trivially understood by a person is precisely the kind of information left unsaid. Who would write, *"If I can walk inside my house, I know that my house is bigger than I am?"* Nevertheless, we naturally understand that: x *walk-inside* $y \implies x <^{\text{size}} y$.

Directly attacking the link between affordances and properties requires access to implications across the edges. Without such information, we can use objects as a proxy to understand how much modern neural networks know about this edge. For example, taking an object like *boots*, and using only its top affordances *wear*, *kick off*, and *lace up*, can we predict its properties?

Statistics

	Total	Statistics
Abstract		
Objects	514	411 train / 103 test
Properties	50	obj/prop: 60 median (3 min, 302 max) prop/obj: 8 median (1 min, 23 max)
Annotations	77,1000	3 anns/datum
Situated		
Objects	1,024	80 unique, split: 64 train / 16 test
Properties	50	
Affordances	3072	3 affordances / object (by design)
Annotations	156,672	3 anns/datum

Examples

Objects	Properties	Affordances
harmonica, van	*expensive, squishy*	*pick up, remove*
potato, shovel	*used as a tool for cooking*	*pet, talk to*
cat, bed	*decorative, fun*	*cook, throw out*

Table 1: Statistics and examples for the proposed abstract and situated datasets (based on (McRae et al., 2005) and (Lin et al., 2014)).

Experiments

Tasks

As shown at the bottom of Figure 1, our problem space naturally defines three edges in a graph. A property prediction task may attempt to produce the human-labeled set of properties given a new object (O→P) (Lucy & Gauthier, 2017). Predicting affordances can be done similarly: given a new object, can its top affordances be distinguished from others (O→A)? And finally, the troublesome but fertile edge between properties and affordances: can a model predict the set of properties compatible with an affordance (A→P)?

We frame each scenario as a series of joint reasoning tasks. Given two instances (e.g., an object and a property), a model must make a binary decision as to whether they are compatible. For example, predicting which properties out of a total of k are compatible with an object o will be set up as k compatibility tasks $(o, p_i) \rightarrow \{0, 1\}$. We denote the tasks as object-property (O⟷P), object-affordance (O⟷A), and affordance-property (A⟷P).

Data

To fuel experiments in these three tasks, we introduce two new datasets. The first we call the *abstract* dataset, which is a set of judgements elicited from only the name of the object (e.g., *wheelbarrow*) and property (e.g., *is an animal*). The second is the *situated* dataset, where properties and affordances are annotated on objects in the context of real-world pictures.[2]

Abstract Dataset Several lists of properties (McRae et al., 2005), categorization schemes (Devereux, Tyler, Geertzen, & Randall, 2014), and quantification layers (Herbelot & Vecchi, 2015) have been proposed. We take the set of objects and properties from McRae et al. and perform filtering and preprocessing similar to Lucy and Gauthier (2017). We also include the set of objects from the MS COCO dataset (Lin et al., 2014), collapse similar objects (e.g., many bird species) and add seven new properties (such as *man-made* and *squishy*). We end up with a set of 514 objects and 50 properties. We re-annotate all 25,700 object-property pairs to eliminate false negatives from the original McRae data collection process and provide labels for new entries. We annotate each pair three times for a total of 77,100 annotations, and keep only labels with $\geq 2/3$ agreement.

Situated Dataset We also annotate instances of objects situated in photographs. Images have the great advantage of resolving visual ambiguities of appearance, shape, and form. For example, a *bottle* has different properties if it is a glass beverage container or plastic shampoo tube. Only a few non-visual properties (e.g., *smelliness*) must then be inferred from the environment.

To build the an experimental situated testbed, we sample images from the MS COCO dataset (Lin et al., 2014). We constrain each image to have between three and seven objects to avoid scenes that are too sparse (often portraits) or dense (cluttered collections). We also ensure that we have at least five samples of each of the 80 unique object categories in the dataset. We end up with 1,024 objects across 220 images. We then annotate all 50 properties (introduced in the abstract dataset) for each object, annotating each three times for a total of 153,600 labels. We filter using the same scheme ($\geq 2/3$ agreement).

In addition to the properties, we also collect annotations of the affordances for all objects in the situated dataset. We allow annotators to choose from the 504 verbs from the imSitu dataset (Yatskar, Zettlemoyer, & Farhadi, 2016). We provide common variants of each verb that include particles, allowing annotations such as *pick up* and *throw out*. Annotators select the top three to five affordances that come to mind when they see the selected object in the context of its photograph. We again perform this annotation three times for each object, and aggregate the verbs chosen to pick the top three most common affordances for each object. We end up with a set of sparsely labeled affordances for each situated object. We perform balanced negative sampling by selecting $k = 3$ affordances for each datum and setting their labels to zero.

Detailed statistics and examples for both datasets are shown in Table 1.

[2]Annotations for both datasets are performed by workers on Amazon Mechanical Turk.

	Abstract				Situated											
	O ⟷ P				O ⟷ P				O ⟷ A				A ⟷ P			
	obj	*prop*	*μF1*	*sig*	*obj*	*prop*	*μF1*	*sig*	*obj*	*aff*	*μF1*	*sig*	*aff*	*prop*	*μF1*	*sig*
RANDOM	0.25	0.26	0.26	***	0.24	0.25	0.22	***	0.53	0.62	0.51	***	0.24	**0.26**	0.23	***
MAJORITY	0.34	0.11	0.31	***	0.16	0.05	0.17	***	0.82	0.68	0.82	***	0.18	0.05	0.17	***
GLOVE	0.63	0.47	0.63	*	0.55	0.39	0.57	**	**0.85**	**0.73**	**0.86**	←	0.27	0.13	0.29	
DEP-EMBS	0.62	0.42	0.60	**	0.54	0.36	0.54		0.84	0.67	0.84		0.26	0.12	0.28	
BERT	0.62	0.48	0.60	***	0.53	0.38	0.56		**0.85**	0.70	0.85		0.26	0.12	0.28	**
ELMO	**0.67**	**0.55**	**0.67**	←	**0.58**	**0.44**	**0.58**	←	0.84	0.71	0.85		**0.31**	0.17	**0.34**	←
HUMAN	0.78	0.80	0.67		0.70	0.69	0.61		0.83	0.93	0.80		0.65	0.67	0.40	

Table 2: Macro F1 scores per category (object, property, affordance) and micro F1 score (*μF1*) on both the abstract and situated test sets. Highest model values are bolded. Statistical significance (*sig*) is calculated with McNemar's test, comparing the best-scoring model (by *μF1*, denoted ←) with each other model. Stratified p-values are shown, with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. Human performance is estimated by 50 expert-annotated random samples from the test set (no McNemar's test).

Models

Word embeddings We consider four representations of the words involved in the tasks. Two of the representations are word embeddings. These map single words to vectors in $\mathbb{R}^d$. We use GloVe embeddings (Pennington et al., 2014) as they have proven effective at object-property tasks in the past (Lucy & Gauthier, 2017). We also use Dependency Based Word Embeddings (Levy & Goldberg, 2014), as they may more directly capture the relations between objects and their affordances. In both cases, $d = 300$, and we use the GloVe embedding variant with the largest amount of pretraining (840 billion words).

Contextualized representations The other two representations are ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), which are contextualized. These require full sentences (as opposed to single words) to compute a vector, but in turn produce results more specific to words' linguistic surroundings. For example, ELMo and BERT produce different representations for *book* in *"I read the book"* versus *"Please book the flight,"* while word embeddings have only a single representation.

To account for this, we generate sentences using the relevant objects, properties, and affordances for the task at hand. For example, to judge *accordion* and *squishy*, we would generate *"An accordion is squishy."*

For ELMo, we then take the final layer representations for the two compared words, each of which is a $d = 1024$ length vector. For BERT, we take the overall sentence representation and sum across the final four layers, which produces a single $d = 1024$ vector.

Finetuning Given the word representations above, we finetune each of the models by adding trainable multilayer perceptron (MLP) after the input representations. This allows models to learn interrelations between the two categories at hand, essentially calibrating the unsupervised representations into a compatibility function. We use a single hidden layer in the MLP, and train using mean squared error loss with L2 regularization.

To summarize, for two words (w_i, w_j) which can be written together in a sentence $s = w_1 ... w_n$, we have for a model m,

$$r(w_i, w_j) = \begin{cases} \langle m(w_i), m(w_j) \rangle & \text{if } m \in \{\text{GL., D.E.}\} \\ m^{-1}_{\{i,j\}}(s) & \text{if } m = \text{ELMO} \\ \sum_{\ell \in \{-4 ... -1\}} m^{\ell}(s) & \text{if } m = \text{BERT} \end{cases}$$

$$\hat{y}_{w_i,w_j} \propto \sigma(\mathbf{w_2^T} a(\mathbf{w_1^T} r(w_i, w_j) + \mathbf{b_1}) + \mathbf{b_2})$$

$$\mathcal{L}(w_i, w_j, y, \theta, \lambda) = (y - \hat{y}_{w_i,w_j})^2 + \lambda \|\theta\|_2^2$$

where $m(\cdot)_i^{\ell}$ is an embedding of the ith token in the layer ℓ, a is a nonlinear activation function, $y \in \{0, 1\}$ is the ground truth label, $\theta = \{\mathbf{w_1}, \mathbf{w_2}, \mathbf{b_1}, \mathbf{b_2}\}$ are trainable parameters, and λ is the regularization strength.

We optimize all models using gradient descent, and tune all hyperparameters using k-fold cross validation with $k = 5$.

Baselines We compare performance for these models against two simple approaches. The *random* baseline simply flips a coin for each compatibility decision. The *majority* baseline uses the per-class majority label for the training set, aggregating by property for the $O \longleftrightarrow P$ and $A \longleftrightarrow P$ tasks, and by affordance for the $O \longleftrightarrow A$ task.

Human performance Finally, we estimate human performance on this task. We sample 50 samples at random from the test set for each task, and have an expert annotate them. For fairness to the models, we do not show the expert the photographs or exact instance from which the situated examples are drawn.

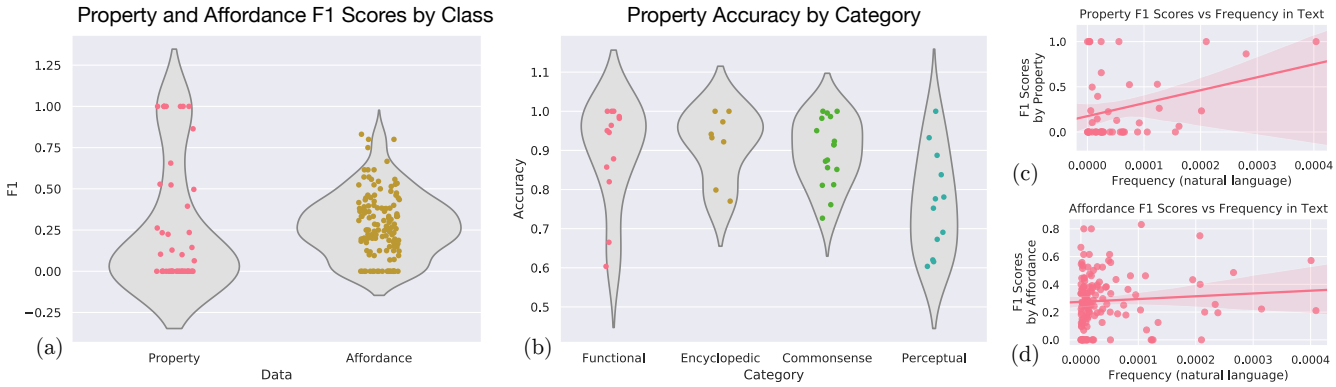

Figure 2: Detailed results of top performing model (ELMo) on the affordance-property compatibility task (A⟷P) in the situated dataset. (a) F1 scores are plotted per property (left) and affordance (right). (b) Properties are divided into four categories and plotted by accuracy. (c), (d) Both property and affordance F1 plotted against word frequency in natural language text.

Results

A summary of all model performances is shown in Table 2. Consistent with prior work that has studied object and property compatibility (Lucy & Gauthier, 2017), we find good but not perfect performance (close to 0.70 F1 scores) on the abstract dataset (task $O \longleftrightarrow P$). Models fare slightly worse on the situated $O \longleftrightarrow P$ task, with the best performance below 0.60 F1. This effect is consistent in the human expert scores as well. Though this dataset is larger, the introduction of context allows for greater variance in the properties of an object.

The object-affordance compatibility task ($O \longleftrightarrow A$) yields significantly higher numbers. Not only is this task statistically easier (as demonstrated by the strong majority baseline), but this edge is the only one directly observed in language. All models pretrained on text have been exposed to many instances of likely verbs for each object considered. In fact, all pretrained models perform in the same range as human ability, and there is no statistically significant difference between the models for this task.

However, all models struggle with the affordance-property task ($A \longleftrightarrow P$). The highest F1 scores are in the 0.30s, with the random baseline achieving the highest macro F1 score by property. While this task is also the most difficult for humans, their macro F1 scores for both affordances and properties are more than double that of the best performing models. We posit that the inference between affordances and properties requires multi-hop inference that is simply not present in the pretraining of large text-based models. We provide further analysis in the following section.

Analysis

Models achieve reasonable performance predicting the compatibility of both properties and affordances with objects. However, the task requiring inference between affordances and properties (A⟷P) confounds even the strongest models.

We explore this result through a detailed analysis of the top performing model. Figure 2 presents a breakdown of ELMo's results on the affordances-property compatibility task (A⟷P) on the situated dataset. From the leftmost graph (a), we observe that a per-property analysis shows a largely bimodal split between properties that are fully predicted (1.0 F1), and went completely unmodeled (0.0 F1). Affordances, on the other hand, lie more evenly across the F1 range. Because the task involved the compatibility between properties and affordances, mass for correct predictions must be shared between the two data groups. That so few properties achieved a high F1 score suggests that many affordances rely on only a few properties for accurate prediction.

We perform further analysis to investigate which kinds of properties yielded better affordance-property modeling. We categorize each property into four coarse classes: functional (e.g., *is used for cooking*), encyclopedic (e.g., *is an animal*), commonsense (e.g., *comes in pairs*), and perceptual (e.g., *is smooth*). Figure 2 (b) shows a breakdown of property performance grouped by these four categories. (Here, we plot accuracy instead of the sharper F1 metric to better illustrate the spread of performance.) Functional properties exhibit the highest performance. This makes intuitive since, because functional capabilities are directly tied to an object's affordances. In contrast, perceptual properties exhibit generally lower and inconsistent performance than other categories. We suspect that perceptual observations observed in text are not expressed with affordances, making this connection difficult for models. Largely perceptual features can be written about with simple verbs (*hear, see, feel*), giving them less implicit evidence than more nuanced properties. Finally, encyclopedic and commonsense properties fall somewhere in the middle. These properties, which involve an object's general characteristics (like *requires gasoline, lives in water,* or *has a peel*), correlate with a variety of verbs. But they may only be directly expressed at a distance from a verb, making the inference between them still challenging.

Our final analyses in Figure 2 (c) and (d) investigate

whether there is a link between the predictive power of the model and how often a word is used in text. We compute the frequencies of all affordances and properties occurring in natural language using the Google Web 1T corpus, an n-gram corpus computed from approximately one trillion words (Brants & Franz, 2006). Figure 2(c) plots the F1 score of properties against how frequently they appear in natural language; 2(d) plots the same for affordances. We include a best-fit line along with confidence intervals shown as one standard deviation of the data. We do not observe a statistical correlation between how much affordances and properties are written about, and how well neural models are able to connect their effects; a single confidence interval spans both positive and negative slopes. This lack of clear correlation is surprising, because large state-of-the-art neural textual models generally improve with repeated exposure to instances of words. Except for the three most common words measured by property F1 score, the rest of the data shows a strikingly uniform distribution of F1 scores for any choice of frequency in natural language. This suggests that current neural models are fundamentally limited in their capacity for physical reasoning, and that only new designs—not more data—can allow them to acquire this skill.

Discussion

Despite being able to associate a considerable range of information with the names of objects, neural models are not able to capture the more subtle interplay between affordances and properties. In some sense, this result is unsurprising. Collecting information around an object can be informed largely by the co-occurrence of words around that object's various mentions. Affordances that imply properties (and the reverse) are rarely mentioned together; their mutual connotation naturally renders joint expression redundant. Hence, priorless models that learn from statistical associations falter. Given the depth of the networks used in models such as ELMo and BERT, complex inter-parameter structure arises, but the latent semantic patterns that describe physical commonsense are much weaker than more superficial patterns that arise due to grammar or domain, making it difficult to capture.

This evidence feeds into theories of embodied cognition (Gover, 1996; Wilson, 2002), which suggest that the nature of human cognition depends strongly on the stimuli granted by physical experience. If this is so, then how is information encoded in our physical experience such that we can make predictions? If we assume a form of mental simulation, then what are the mental limits on its reliability? From an artificial intelligence perspective, the more interesting proof is in the principles of creating such a mental simulator. If we are to simulate human capacity for thought, how actually must we simulate elements of the physical world?

With the rise of physics engines, our ability to model physical inferences grows (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). However, while this may make us better at anticipating human predictions about physical situations through perceptual stimuli (Gerstenberg, Zhou, Smith, & Tenenbaum, 2017), there is still a long way to go before we understand the inferences that are being made through more symbolic stimuli, such as language. Exploring the mechanisms underlying this communication using an implicit shared world model will require us to either develop access to such a world model, or expose algorithms to predictions of that world model by directly querying humans. Bridging the inductive biases learned from simulation (Battaglia, Hamrick, & Tenenbaum, 2013) and those discovered by scientists (Lake, Linzen, & Baroni, 2019) to make inferences implicit in text will lead to a more cohesive model of commonsense physics. We expect such a model to bear great fruit in studies of communication rich with physical implications.

Acknowledgments

This work was supported by NSF grants (IIS-1524371, 1637479, 1703166), NSF Fellowship, the DARPA CwC program through ARO (W911NF-15-1-0543), and gifts by Google and Facebook. The views and conclusions contained herein are those of the authors and should not be interpreted as representing endorsements of the funding agencies.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Brants, T., & Franz, A. (2006). Web 1t 5-gram version 1.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, *46*(4), 1119–1127.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gaver, W. W. (1991). Technology affordances. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 79–84).

Gerstenberg, T., Zhou, L., Smith, K. A., & Tenenbaum, J. B. (2017). Faulty towers: A hypothetical simulation model of physical support. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Gibson, J. J. (1966). The senses considered as perceptual systems.

Gover, M. R. (1996). The embodied mind: Cognitive science and human experience (book). *Mind, Culture, and Activity*, *3*(4), 295–299.

Herbelot, A., & Vecchi, E. M. (2015). From concepts to models: some issues in quantifying feature norms. In *Lilt* (Vol. 2).

Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. In *Proceedings of the thirteenth international conference on principles of knowledge representation and reasoning* (pp. 552–561). AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=3031843.3031909

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 302–308).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. *arXiv preprint arXiv:1705.11168*.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

Norman, D. (1988). *The design of everyday things: Revised and expanded edition*. Constellation.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (Vol. 1, pp. 2227–2237).

Rubinstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015). How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (Vol. 2, pp. 726–730).

Van Durme, B. D. (2010). *Extracting implicit knowledge from text*. University of Rochester.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).

Yatskar, M., Zettlemoyer, L., & Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5534–5542).