Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Haohan Shi April 8, 2021

What Transformers Might Know About the Physical World: T5 and the Origins of Knowledge

By
Haohan Shi

Dr. Phillip Wolff
Advisor

Department of Psychology

Dr. Phillip Wolff
Advisor

Dr. Jinho Choi
Committee Member

Dr. Philip Kragel
Committee Member

2021

What Transformers Might Know About the Physical World: T5 and the Origins of Knowledge

By

Haohan Shi

Dr. Phillip Wolff

Advisor

An Abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelors of Arts with Honors

Department of Psychology

2021

Abstract

What Transformers Might Know About the Physical World: T5 and the Origins of Knowledge

By Haohan Shi

We find that knowledge of animals and objects' physical properties does not depend on direct or indirect perceptual experience. Rather, such knowledge can emerge from inferential processes driven by the statistical properties of language. Here we investigate the latent knowledge of the T5 encoder-decoder model with respect to various physical properties of animals and objects. Such networks represent model organisms for the origins of knowledge in a learning system without innate knowledge or access to perceptual information. We proposed and evaluated three hypotheses about what T5 might know about the physical world: 1) that T5 might understand physical dimensions much like humans understand these dimensions, 2) that it has no understanding of the perceptual dimensions of experience, and 3) that it understands some dimensions of experience better than others and potentially uses the better-understood dimensions to understand the less well-understood dimensions. The results from Study 1-4 show that knowledge of the size, weight, and shape—but not color—agrees closely with that of humans. Moreover, agreement with human judges increased with the network's size, suggesting that disagreement with humans might ultimately disappear as the size of the networks is increased. However, Study 5 shows that T5's understanding of size might rely on its understanding of weight, which supports our third hypothesis regarding T5's understanding of dimensions in physical world.

What Transformers Might Know About the Physical World: T5 and the Origins of Knowledge

By
Haohan Shi

Dr. Phillip Wolff
Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelors of Arts with Honors

Department of Psychology

2021

Acknowledgements

I want to thank my parents who have helped and supported me throughout the four years in Emory University. They let me explore my own interest and offered maximum mental and financial support for my education. Their selfless support led to my accomplishment of the degree and the thesis.

Also, I would like to thank my advisor, Dr. Phillip Wolff, who introduced me to research in linguistics, psychology, and data science, and I'm grateful for all the advice and resources he offered for my thesis. He have inspired me and cemented my research interest in the interdisciplinary field of computational science, psychology, and linguistics.

I want to thank my committee members, Dr. Philip Kragel and Dr. Jinho Choi. They gave me valuable advice from their own fields' view, which expanded my horizons and helped me to notice many research questions needing to be addressed in the future.

Finally, I want to thank my friends including Jeffrey Li, Yongji Wu, and Xueting Yang, who have helped me during the four years in Emory University, especially during the hard time under COVID-19.

Table of Contents

Chapter 1

Introduction

There are essentially three accounts of the origin of knowledge (Anderson, 1989). According to *Nativism*, concepts and beliefs are hard-wired into the brain (Samuels, 2002). According to *Empiricism*, knowledge is acquired from the storage of perceptual experience. A third possibility, *Rationalism*, holds that knowledge is acquired from reasoning, that is, mental operations acting on stored representations that derive implications from that knowledge. In human learning, the mental operations implied by these positions are not mutually exclusive. Knowledge of the physical world may emerge from perceptual experience constrained by innate processing capabilities and extended through reasoning. However, certain knowledge would seem to depend on direct perceptual experience, such as knowledge of an object's appearance. Recent evidence suggests that this may not always be the case.

Kim, Elli, and Bedny (2019) investigated congenitally blind individuals' knowledge of the visual properties of animals. Surprisingly, blind individuals' judgments about the relative size, height, shape, and skin texture of a wide range of animals largely agreed with that

of sighted individuals. A possible explanation for the findings is that the bind people relied on sighted people's language about animals to make judgments about their perceptual properties. This possibility predicts that agreement between blind and sighted people should be highest for perceptual properties that are relatively easy to verbalize, like color. As it turns out, it was for this very perceptual property—color— that agreement between the sighted and non-sighted participants was lowest, suggesting that the relatively high agreement between blind and sighted individuals was not simply due to remembering comments of sighted people.

Kim et al. raise another possible explanation for the relatively high agreement between blind and sighted individuals. In particular, blind people might use knowledge of an animal's taxonomic category—which they learned through language—to draw inferences about an animal's perceptual properties. The explanation is supported by findings showing that children can use knowledge of natural kinds to draw inferences about invisible properties, such as animals' insides (Gelman & Wellman, 1991; Keil; 1989). While this hypothesis seems entirely reasonable, it leaves unexplained the difference in performance between properties like shape and color. An inference account still depends on language. If language is used to acquire knowledge of taxonomic categories, then why would it not be used for answering questions about color? Perhaps, as suggested by Kim et al., and claimed by John Locke (1924), blind individuals know more about properties such as shape than color because blind individuals have perceptual experience with shape though the sense of touch. Perhaps it is this indirect experience with the perceptual world that allows blind individuals to be able to draw successful inferences from about shapes, sizes, heights, and textures, and its absence that explains why inferences about color are far less accurate.

We evaluate the knowledge about physical world that can be drawn from language using a remarkable new class of language learning models that not only acquire a sophisticated

understanding of language, such as syntax (Ganesh, Sagot & Saddah, 2019; Goldberg, 2019; Hewitt & Manning, 2019; Peters et al., 2018; Tenney, et al. 2019), but also seem to acquire general knowledge about the world (Petroni et al., 2019; Da & Kasai, 2019). In this research, we used the recently introduced Text-To-Text Transfer Transformer (T5) (Raffel, et al. 2020).

Here we propose three hypotheses about what T5 might know about the physical world. In the first hypothesis, T5 understands the physical dimensions just like humans do. In this case, T5 not only can assess how different entities differ along one dimension, but it also understands differences between different dimensions. An understanding of how dimensions differ is a relatively deep idea. For example, a system like T5 might indicate that a plate weighs more than an acorn, and it might also judge that a plate is larger than an acorn. Knowing the difference between weight and size depends on recognizing that weight and size are entirely different aspects of the physical world. Based on the data above, we do not know whether T5 actually differentiates these dimensions. It is possible that it simply associates plates with "more" of any dimension, rather than specifically "more weight" (heavier) or "more size" (larger). To test whether a system understands the difference between dimensions, it is necessary to not only assess whether it is able to correctly rank order entities with respect to various dimensions. It also requires investigating whether the entities are correctly rank ordered when the values of the dimensions are pitted against one another. For example, while a balloon is larger than an apple, an apple weighs more than a balloon. Presenting a system with such carefully constructed pairs can allow us to fully evaluate the extent of its knowledge. In the second hypothesis, T5 has no understanding of physical dimensions like size and weight. This would be implied if the rank order of entities in terms of these dimensions did not differ from chance rank orderings. In the last hypothesis, T5 understands some physical dimensions better than others, and it may use the better-understood dimensions to understand the less well-understood dimensions. This would be indicated by situations in which the system appears to

correctly rank order entities with respect to one dimension, but it ultimately fails to do so when controlling for another dimension.

1.1 Related Works

Understanding common sense has always been an important goal for natural language processing (NLP) researchers, and many datasets have been created for language models to trained on to reach this goal. PIQA is a dataset created for physical commonsense reasoning in natural language (Bisk et al., 2019). It includes statements about the physical word, such as *To separate egg whites form the yolk using a water bottle, you should…* A model is expected to choose the most sensible solution, specifically a) *Squeeze the water bottle and press it against the yolk. Release, which creates suction and lifts the yolk.* b) *Place the water bottle and press it against the yolk* (Bisk et al., 2019). T5-11B has an outstanding performance on the dataset after fine-tuning. ReCoRD is a reading comprehension dataset requiring commonsense reasoning (Zhang et al., 2018), and T5-11B has 93.4% accuracy rate on the dataset (Raffel et al., 2019). Note that all of the existing commonsense datasets were able to focus on complex commonsense reasoning and require extra training the pre-trained language models on downstream tasks.

Shortly after Kim, Elli, and Bedny's study on congenital blind adults' knowledge in animal appearance, Lewis, Zettersten, and Lupyan (2019) used vector distances between representations of animal words and related target words (e.g., shark-skin vs. shark-feathers) to mimic tests performed in the human study. They found that the semantic representations learned from language significantly correlate with human judgements on animal shape, skin texture, and color (Lewis, Zettersten, & Lupyan., 2019).

In this paper, we aim to explore T5's knowledge in basic commonsense like

size, weight, shape, and color, and we want to see its pre-trained performance without being further finetuned on the task and the dataset.

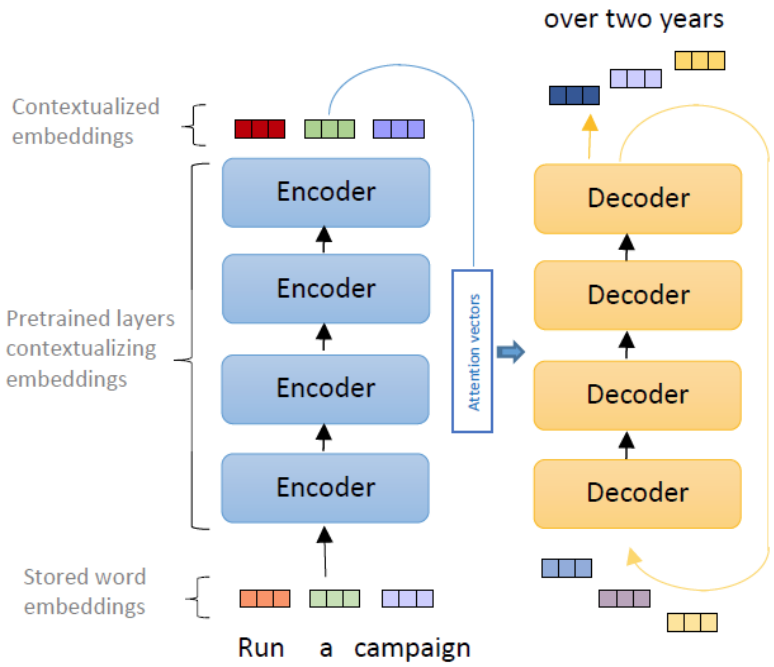

Figure 1: Two main parts of a transformer: Stacks of encoders and decoders that contextualize word embeddings. Words are processed in parallel in the encoder and sequentially in the decoder.

1.2 Transformers

A transformer architecture is a type of statistical learning system that includes multiple layers of feed forward networks preceded by an attentional mechanism that determines the degree to which the "meaning" of a word in a sequence of words is "colored" by the meaning of other words in the sequence. Training is driven by ambiguity resolution: words in the input sequence are masked and the network attempts to predict the word or set of words behind the mask. Training is computationally intensive and can take several weeks to complete. Fortunately, previously trained networks can be freely downloaded.

In this research, we used the recently introduced Text-To-Text Transfer Transformer (T5) (Raffel, et al. 2020). T5 is in a certain sense an old model because it uses the transformer

architecture first proposed in Vaswani et al. (2017). In its original formulation, a transformer has two main parts: a sequence of encoders and a sequence of decoders. The idea behind an encoder and decoder is made most transparent in language translation problems. The encoders are used to read in and "comprehend" the sentences from one language (e.g., French), while the decoders "generate" text in the second language (e.g., English). One of the many innovations in T5 is that it extends this basic idea to a range of language tasks: translation, grammaticality judgments, sentence similarity assessment, summarization, and question answering. It can also complete a sentence, as shown in Figure 1.

The processing begins with the encoder and involves accessing embeddings for each of the input words (or word tokens). A word embedding is essentially a point in a semantic space. Over the layers, the embeddings are modified by the encoders depending on the other words in the input. For example, in the phrase *Run a campaign*, contextualization allows the meaning of the word *run* to take on the meaning suited to campaigns, as opposed to, for example, miles. The ability to contextualize word embeddings is one of the key ways these models differ from those that produce static, context-free embeddings, such as Word2vec (Mikolov, et al., 2013) or Glove (Pennington, Socher, & Manning, 2014). The contextualization process occurs in real-time and is bidirectional, entailing that the process uses words on both sides of the target word. Contextualization depends on self-attentional mechanisms present in each encoder (and decoder). Self-attention reflects the degree to which a word is linked to other words in a sentence. The strength of this connection determines the impact each word will have on a word embedding. Several self-attention weight matrices are learned for each encoder. The self-attention matrices direct the encoder to weigh connections between verbs and their complements, pronouns and their referents, and ambiguous nouns and other nouns. Having multiple self-attention matrices—or heads—allows the encoder to capture the full range of dependencies in a sentence.

The encoders map an input sequence into a continuous abstract representation. The decoders then take that continuous representation and generate words in a step-by-step manner, using the previous step's output as input on the current step. In addition, the decoders' output is constrained by attentional vectors formed from the output of the top encoder. The inclusion of both encoders and decoders is one of the ways T5 differs from several other recently transformer-based models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), which include only the encoder part of the transformer. T5 model comes in several sizes, as specified in Table 1.

Table 1 shows the five T5 models, along with BERT-Large and RoBERTa-Large for comparison.

Model	Parameter	# layers	# heads
T5-Small	60M	6	8
T5-Base	220M	12	12
BERT-Large	336M	24	16
RoBERTa-Large	355M	24	12
T5-Large	770M	24	16
T5-3B	3B	24	32
T5-11B	11B	24	128

Table 1: Model size variants

T5-small is significantly smaller than BERT-Large and RoBERTa-Large. T5-Large is approximately the same size as these two other networks. T5-11B is quite large, containing 11 billion parameters and requires approximately 40GB of memory on a GPU. The model can also be run on a CPU on a system having 120GB of RAM.

A second major innovation of T5 is the manner in which it is trained. In BERT and RoBERTa, the target for an individual mask is associated with a single word piece. In T5, the target for a mask can be several words. For example, given the original sentence is *An elephant is larger than a goat*, T5 might be presented with the string *An elephant is <X> a goat*, with

the target being several words, namely <*X*> *larger than* <*Y*>. Multiword targets are made possible through the use of the stack of decoders.

All versions of T5 were trained on a cleaned version of the common crawl called the Colossal Cleaned Common Crawl (C4). The training corpus is over two times larger than Wikipedia. The largest version of the model, T5-11B, achieved state-of-the-art performance results on the GLUE, SuperGLUE, SQUAD, and benchmarks, which involve natural language processing tasks such as sentiment analysis, question answering, grammaticality judgments, paraphrase detection, selection of plausible causes and results, textual entailment detection, intended meaning detection, and reading comprehension with commonsense reasoning (Raffel, et al. 2020). As already stated, the reason why T5 succeeds on these tasks is likely due to its ability to acquire knowledge of both language and the world.

If we find that a model like T5 is unable to learn physical perceptual properties of the world, it will provide modest support to the proposal that a learning system devoid of any perceptual senses is unable to capture perceptual properties of the world. On the other hand, if T5 is able to capture certain properties of the physical world, it would suggest that physical senses are not a prerequisite for perceptual knowledge and that properties of the perceptual world can be acquired from language alone, contra Lock.

We investigated T5's knowledge of animal size, weight, shape, and color, as well as objects size and weight. It is certainly possible that T5 could show limited evidence of such knowledge, but far below that of human judgments. Under these circumstances, it would be good to know if the limit is a fundamental property of these statistical systems or else, possible, simply a function of the size of the system. To address this question, T5's knowledge was investigated for different model sizes. If the knowledge increases with model size, it would

suggest that any limits in its knowledge might be a simple matter of the model's capacity rather than an inherent limitation of the architecture.

Chapter 2

Experiment

2.1 Study 1: T5's Knowledge of Animal Size

In Study 1 we attempted to replicate judgments of humans about the relative size of animals reported in Kim, Elli, and Bedny (2019). In Kim et al., blind and sighted participants were presented with index cards with the names of animals on them. For the bind participants, the names were printed in Braille. Their task was to order the cards from smallest to largest. The relative ordering of the blind and sighted individuals were nearly identical. In the current study, we assessed T5's knowledge of size by generating cross-entropy loss scores to statements such as *A cat is smaller than a bear* and compared these scores to those in which the order of the animals was reversed, e.g., *A bear is smaller than a cat*. To the extent that T5 has knowledge of size, then loss scores should be lower for orderings that agree with the relative ordering of sizes made by humans. To help establish the generalizability of T5's knowledge, comparative size statements were expressed using three different comparative size adjectives: *smaller, larger,* and *bigger*.

In addition, we investigated the impact of providing a small amount of linguistic context

on T5's judgments. In the context condition, the key sentence was preceded with a few sentences introducing the topic of animals. Specifically, T5 was first presented with the sentences *Animals live outside. They breathe and drink water. They also differ in size*, before being presented with the comparative statement *A cat is smaller than a bear*. The context was included because prior pilot work with another transformer, XLNet (Yang, et al., 2020), suggested that performance of these networks may be improved when context is provided to "warm them up. "

This study would help us to investigate if T5 have at least some knowledge about how entities differ along one dimension, which could help us to evaluate the hypothesis that T5 does not have any knowledge about physical dimensions.

2.1.1 Methods

Materials The list of animals (n = 15) was the same as those investigated in (Kim, Elli, & Bedny, 2019) as shown in Table 2.

Animal Names
mosquito
bee
butterfly
toad
pigeon
raven
cat
koala
turkey
sheep
donkey
cow

bear
rhino
elephant

Table 2: List of animals (from the smallest to the largest)

```
import torch
from transformers import T5Tokenizer, T5Config, T5ForConditionalGeneration


T5_PATH = 't5-11b'


DEVICE = torch.device('cpu')


t5_tokenizer = T5Tokenizer.from_pretrained(T5_PATH)
t5_config = T5Config.from_pretrained(T5_PATH)
t5_mlm = T5ForConditionalGeneration.from_pretrained(T5_PATH,
config=t5_config).to(DEVICE)


input_ids = t5_tokenizer(text = 'A cat is <extra_id_0> than a bear',
return_tensors='pt').input_ids.to(DEVICE)
labels = t5_tokenizer('<extra_id_0> smaller <extra_id_1> </s>',
return_tensors='pt').input_ids.to(DEVICE)
outputs = t5_mlm(input_ids, labels = labels)


print(outputs.loss.item())
```

Figure 2: A python implementation printing out cross-entropy loss of *A cat is smaller than a bear* with a mask on the token *smaller*

Procedure Comparative sentences (n = 210) were generated with different pairs of animals and three kinds of degree adjectives (smaller, larger, bigger), e.g. *A cat is smaller than a bear*. T5 was credited with understanding the relative size difference of two animals when the cross-entropy loss was lower for the correct ordering. A snippet of python code printing out cross-entropy loss is shown in Figure 2. Correct understanding was coded with a 1 and incorrect understanding with a 0. In this and the following analyses, we used the HuggingFace

implementation of T5 (Wolf, et al., 2020).

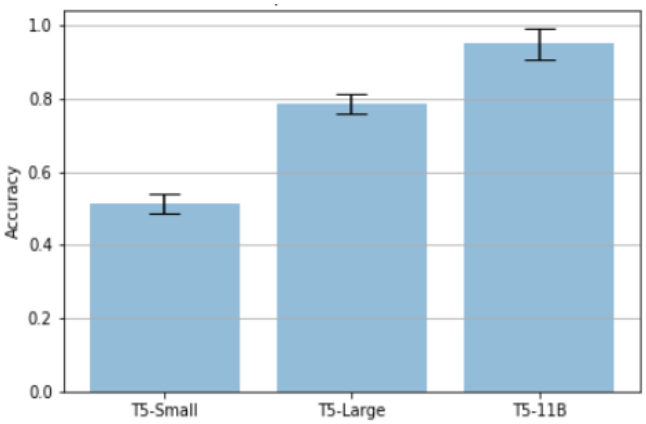

Figure 3: Accuracy of T5 models on animal size with context sentences. Error bars are 95% confidence intervals.

2.1.2 Results and Discussion

The results indicate that T5 has knowledge of the relative size of animals. The overall mean accuracies for different versions of T5 are shown in Figure 3. The results suggest that T5-large and T5-11B have access to perceptual information about size, as indicated by accuracies that differed from chance by binomial test, p = 5.16e-25 and p = 9.13e-69. However, there is no evidence that T5-small has access to this information, p = .652. Difference in performance across the models was confirmed by a main effect of model-type, $F(1,312)=4257$, $p < .0001$. Interesting, accuracy was higher when there was a preceding context than not, $F(1,312)=128$, $p < .0001$, suggesting that such systems may benefit from a short priming of the topic. There was also an effect of adjective type, $F(2,624)=7.51$, $p = .001$. However, this effect occurred only in the absence of context, with accuracy being higher for the *small* than *large* and *big*. Overall, we conclude that the results were largely the same across adjectives, especially when there was a preceding context, suggesting that the knowledge is not tied to a particular linguistic expression. Crucially, knowledge of relative size of animals is very robust in T5-11B, implying that networks with transformer architectures may be able to approach blind individuals' understanding of this dimension of experience. The results suggest that T5

has some knowledge about the physical world, and thus the second hypothesis that T5 has no understanding of the physical world is rejected.

2.2 Study 2: T5's Knowledge of Animal Weight

In Study 2, we investigated if T5 understands another dimension – weight. If T5 does have knowledge about weight, then we can construct pairs with items having inconsistent weight and size to evaluate the first hypothesis. The design and implementation of Study 2 was analogous to Study 1, except that instead of size, we investigated T5's knowledge of weight, a perceptual property that presumably depends on haptics, but is likely also informed by size. Weight was not examined in Kim, Elli, and Bedny (2019), but we could use relative size as an indicator of the correct responses.

2.2.1 Methods

Material The analyses used the same list of 15 animals as in Study 1.

Procedure We ranked the animals in the list from the lightest to the heaviest based on their size. The texts were generated in the same way as in Study 1 except that we replaced the adjectives with "weighs more" and "weighs less". A context was preceding each text, specifically: *Animals live outside. They breathe and drink water. They also differ in weight*. Afterwards, T5 was presented with sentences like *A raven weighs more than a bee*.

The procedure of comparing cross-entropy loss in Study 1 was performed to generate accuracy scores.

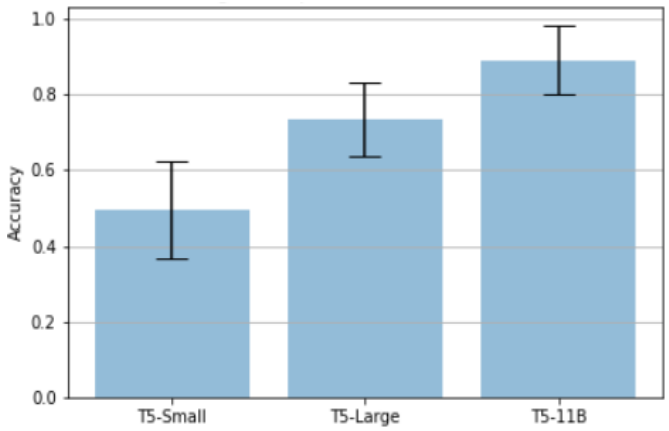

Figure 4: Accuracy of T5 models on animal weight with context sentences. Error bars are 95% confidence intervals.

2.2.2 Results and Discussion

The results indicate that the larger versions of T5 have world knowledge about the relative weight of animals. The overall mean accuracies for the differently sized models are shown in Figure 4. The results were highly similar to those of size. Accuracy scores for T5-large and T5-11B differed from chance, p = 3.47e-33 and p = 9.03e-12, but accuracy scores for T5-small did not, p = .945. Weight represents a type of force. The results raise the possibility that large transformer models like T5-11B may learn visual information about the world and invisible information like forces, which could prove crucial in identifying and reasoning about causes and results (Wolff & Shepard, 2013). As stated above, we cannot conclude that T5 understands different dimensions yet since animal size and weight are linearly correlated – a heavier animal will have a larger size. It is possible that T5 only understands that one animal is "more" than the other.

2.3 Study 3: T5's Knowledge of Animal Shape

In Study 3 we investigated T5's knowledge of animal shape. Unlike size and weight which can be reduced to a single dimension and referred to by an adjective, shape is inherently

multidimensional and more difficult to describe. Instead of focusing on a single dimension, T5 was presented with sentences describing each animal with respect to a range of shape-related adjectives like *long, short, thin* and *thick*. For example, T5 evaluated the acceptability of sentences like *A giraffe is long* or *A sloth is thick*. The result was an animal-by-shape-dimension matrix of cross-entropy scores. The dimensionality of the matrix was reduced and submitted to *k*-means clustering. In Kim et al. (2019), people's sorts of animal names were used to place 30 animals into one of 8 categories with respect to shape. We created 8 clusters from the animal-by-shape-dimension matrix and evaluated the degree to which these clusters agreed with those produced by participants in Kim et al. Given the superior performance of T5-11B, we focused on only that variant for the following analyses.

2.3.1 Methods

Materials The analyses used the list of 30 animals that were used in animal shape card sorting task (Kim, Elli, & Bedny, 2019). Their names are shown in Table 3.

Animal Names
dolphin
shark
killer whale
bat
pigeon
crow
swan
flamingo
beaver
skunk
sloth
panda
polar bear
grizzly

gorilla
mammoth
elephant
hippo
rhino
pig
boar
sheep
goat
zebra
deer
llama
giraffe
lion
panther
cheetah

Table 3: List of animal names used in Study 3 and Study 4.

Dimension	Pole 1	Pole 2
length	*long*	*short*
straightness	*straight*	*curved*
thinness	*thin*	*thick*
squareness	*circular*	*square*
orientation	*vertical*	*horizontal*
roundness	*rounded*	*angular, pointed*

Table 4: List of animal shape adjectives.

Procedure We introduced 13 different adjectives describing animal shape as shown in Table 4. The texts were generated in the pattern that contains a noun (animal) and an adjective (shape), e.g. *The shape of a giraffe is vertical*. With 30 animals, 30 texts were generated for each adjective. A context preceded each text: specifically, *Animals live outside. They breathe*

and drink water. They also differ in shape. We used T5 to predict the adjectives and took the cross-entropy loss as a raw score for each text. The dimensionality of resulting cross-entropy matrix was reduced to three dimensions using the IVIS dimensionality reduction framework (Szubert et al., 2019), which does a better job of preserving both local and global structure than linear projection methods such as Principal Components Analysis (PCA) and other non-linear dimensionality reduction methods such as t-SNE (Maaten & Hinton, 2008). The reduced space was partitioned into 8 clusters using the k-means++ algorithm available in scikit-learn (Pedregosa, et al., 2011).

Figure 5: Text plot of animals clustered with respect to shape as determined by T5.

2.3.2 Results and Discussion

The results indicate that T5-11B has knowledge of animal shapes. Agreement between

the eight human and T5 clusters was achieved by identifying the most common cluster label generated by T5 for each human category, counting the number of times that label occurred within each human category and dividing by the sum of the counts by the number of animals. This measure indicated that the clusters produced from T5 overlapped 70% with those of humans. Assuming a chance level of agreement of .399 (based on simulations), the clusters produced from T5's scores differed from chance by binomial test, $p = .001$. A textplot of one of the solutions produced from T5 is shown in Figure 5.

2.4 Study 4: T5's Knowledge in Animal Color

In this study we investigate T5's knowledge of animal colors. Unlike shape, languages like English make it relatively easy to refer to color. However, animals are not uniform in color, as exemplified by zebras, killer whales, and giraffes. Given the multi-dimensional nature of animal colors, Kim, et al. (2019) measured knowledge of animal colors in terms of eight color-combination categories, as they had done with shape. As a consequence, we measured knowledge of animal colors in the same way we did in Study 3. T5 evaluated sentences such as *A gorilla is yellow* or *A gorilla is black*. The resulting matrix of cross-entropies was reduced in dimensionality and partitioned into eight clusters. Finally, we measured the degree of overlap between the clusters formed by humans in Kim et al. (2019) and those produced by T5.

2.4.1 Methods

Materials The analyses used the same list of 30 animals as in Study 3, and the animal names are shown in Table 3.

Procedure We introduced 6 different adjectives that can be used to describe the color of the animals in the list (*black, white, brown, grey, yellow, blue*). The texts were generated in the same way as in Study 3, except that the adjectives describing shapes were replaced with adjectives describing colors. With 30 animals, 30 texts were generated for each adjective. We

used T5 to predict the adjectives and took cross-entropy loss as a raw score for each text.

Figure 6: Text plot of animals clustered with respect to color as determined by T5.

2.4.2 Results and Discussion

T5 demonstrated less knowledge about color than the other perceptual dimensions. Agreement between the eight human and eight T5 clusters was 60%, differed from chance (.399), p = .021, but not as strongly as what was observed for the other perceptual dimensions. A textplot of the solution produced by T5 with respect to color is shown in Figure 6. Several of the clusters seem intuitive, such as the cluster containing animals that are black (*bats, panthers, crows,* and *gorillas*) and white (*swan, polar bear*), but several others are less clear, such as the cluster containing *hippos, sloths,* and *pigeons*. We conclude that T5 has relatively limited knowledge of animal colors.

2.5 Study 5: T5's Knowledge in Object Weight and Size

In study 1 and study 2, T5 demonstrated its knowledge in animal's weight and size. In the case of animals, weight and size are linearly correlated, and thus a larger animal will have a larger weight. However, the linear correlation does not hold in many other cases. For example, a balloon can be larger than an apple, but an apple is heavier than a balloon. This raises the question whether T5 really knows about animal size and weight or it just compares objects based on a general sense of "more". In Study 5, we created a list of pairs of objects based on the contrast of their weight and size. For example, *magnet* and *leaf* are a pair: *magnet* is "more" in terms of weight, while *leaf* is "more" in terms of size. For each pair, an analysis similar to Study 1 and Study 2 were conducted to see if T5 had knowledge in the objects' weight and size.

2.5.1 Methods

Material A list of 15 pairs of objects was created as shown in Table 3.

Larger Lighter	Smaller Heavier
straw	*fork*
napkin	*watch*
paper	*knife*
bed sheet	*laptop*
leaf	*magnet*
feather	*ice cube*
sail	*car*
beach ball	*bowling ball*
tissue	*spoon*
flag	*laptop*
parachute	*motorcycle*
spider web	*apple*
balloon	*brick*
envelope	*stone*

shirt	*glass*

Table 4: Pairs of objects.

Procedure Comparative sentences (n = 120) were generated with different pairs of objects and six kinds of degree adjectives, namely *larger*, *smaller*, *(weighs) more*, *(weighs) less*. T5 was credited with understanding the relative size or weight differences of two objects when the cross-entropy loss was lower for the correct ordering. Correct understanding was coded with a 1 and incorrect understanding with a 0. A context sentence was added preceding each comparative sentence, specifically *Different objects have different sizes* for *larger/smaller* comparison and *Different objects have different weights* for *weighs more/weighs less* comparison.

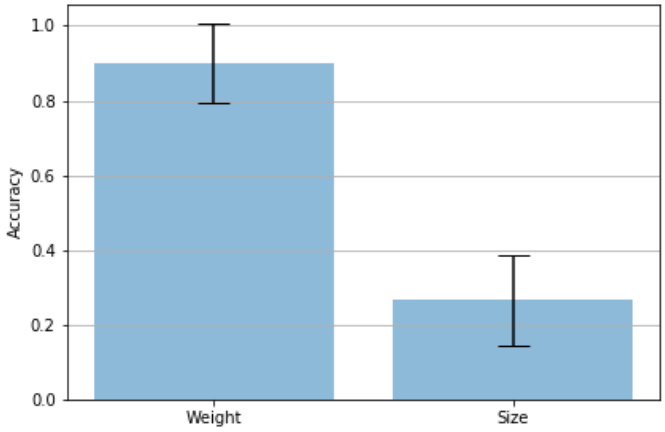

Figure 7: Accuracy of T5-11B on object weight and size with context sentences. Error bars are 95% confidence intervals.

2.5.2 Results and Discussion

The overall mean accuracies for weight and size are shown in Figure 7. The results indicate that T5 has knowledge of the relative weight but poor knowledge of the relative size of the objects with inconsistent weight and size. The results suggest that T5-11B has access to perceptual information about weight, as indicated by accuracies that differed from chance by

binomial test, p = 8.43E-06. However, T5-11B understands perceptual information about size in an opposite and incorrect way, p = .01612. Overall, we conclude that the knowledge of relative weight of objects is very robust in T5-11B, but its knowledge of relative size of objects seem to rely on its knowledge of relative weight. This can be an evidence opposing to our conclusion in study 1 that T5-11B has knowledge of relative size of animals, which rejects the first hypothesis that T5 understands different dimensions. This leaves us the third hypothesis that T5 understands certain physical dimensions better than other dimensions, and it potentially uses the better understood dimensions to understand the less well understood dimensions.

Chapter 3

General Discussion

In this research we investigated the perceptual knowledge of a model organism having extensive experience with language, but no direct or indirect experience with physical quantities in the world. Remarkably, this fully disembodied organism demonstrated high levels of sensitivity to physical features of the world. The results establish that the information needed to learn perceptual features of animals is present in language. The results suggest that the model's success is not due to the simple retrieval of pieces of statements about the perceptual characteristics of animals. If performance was merely a matter of memory retrieval, the systems' awareness of the colors of animals should have been much stronger than was observed. Seemingly, the results largely mirror the results found in Kim, Elli, and Bedny (2019), with agreement between sighted and blind participants highest for size, then shape, and then relatively poor for color. Results from Study 1 and Study 2 show that T5 can correctly rank order animal size and weight. However, these results can only indicate that T5 understands

some dimension of the physical world, which rules out the second hypothesis that T5 does not understand any dimension in physical world. Results from Study 5 show that T5

cannot correctly rank order object size if the objects have inconsistent weight and size. This implies that T5 is not able to assess different entities along different dimensions, which rules out the first hypothesis and leaves us the third hypothesis – T5 understands some dimensions better and uses the better-understood dimensions to understand the less well understood dimensions. As the accuracy of size ordering in Study 5 is low and statistically different from chance, we can reasonably imply that T5's perception of size is based on its perception of weight which is shown to be T5's robust physical world knowledge in Study 2 and Study 5. The pattern of results suggests that statistical systems such as T5 may offer a powerful model organism for mining the conceptual system.

In this research, we aimed to replicate some of the studies in Kim et al.'s work (2019), but we failed to do so in Study 3 and Study 4 due to T5's incapability of doing classification without being fine-tuned. In Study 3, we used adjectives describing shape to evaluate T5's knowledge of the shape of different animals. However, the subjects did not have access to these adjectives but directly sorted the animals based on their shape. Similarly, T5 did not sort the animals but evaluated each color associated with different animals. In future studies, human data should be collected on tasks similar to T5's tasks, so that the potential confounding variables can be eliminated. In Study 5, we used only 15 pairs of objects, which is a small dataset compared to the 105 animal pairs used in Studies 1 and 2. We would like to create more such pairs for T5 investigate the generalizability of the results. Throughout the research, we only tested T5's knowledge in animals and objects' physical properties, but there exists many other outstanding language models. Even though T5 fails on the object size task in Study 5, other models might succeed. Therefore, more models should be tested in future work.

Why does T5 understand weight better than size? One possibility is that weight is a dimension that enters into our understanding of "power" and "stability", and that these notions

might have more causal impact than relative size. Another possible account is that weight is more reliable than size. In general, weight reflects quality of an object, and quality is a relatively stable dimension. It is possible to "make up" size without modifying quality, but it would be impossible to "make up" quality. Take the sponge for example: one can squeeze a sponge and make it much smaller than the original size, but its quality would not change.

The current study provides evidence showing that knowledge of the physical world does not inclusively come from perceptual experience. Instead, physical properties like weight and shape can be derived from language alone.

References

Anderson, J. R. (1989). A theory of the origins of human knowledge. *Artificial Intelligence, 40,* 313-351.

Bisk, Y., Zellers, R., Gao, J., & Choi, Y. (2020, April). Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 7432-7439).

Da, J., & Kasai, J. (2019). Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pp 1-12.

Ganesh, J., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Gelman. S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. Cognition, 38, 213-244.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. Computing Research Repository, arXiv:1901.05287.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

Keil, F. C. (1989). Concepts, kinds, and conceptual development. MIT Press: Cambridge, MA.

Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences*, *116*(23), 11213-11222.

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237-19238.

Locke, J. (1924). An essay concerning human understanding. Nature 114, 462.

Mikolov, T., Chen, K., Corrado, D., & Dean, J. (2013). Efficient estimation of word representations in vector space. In International Conference on Learning Representations (ICLR) 2013. Retrieved from https://sites.google.com/site/representationlearning2013/workshop-proceedings

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word

representation. In Proceedings of the 2014 Conference on EMNLP (pp. 1532–1543). New York, NY: Association for Computational Linguistics.

Peters, M. E., Neuman, M., Zettlemoyer, L., & Yih, W. (2018). Dissecting contextual word embeddings: Architecture and representation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 – November 4, 1499-1509.

Petroni, F., Rocktaschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language, pages 2463 – 2473.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv:1910.10683[cs.LG]

Samuels, R. (2002). Nativism in cognitive science, *Mind & Language, 17*, 233-265.

Szubert, B., J. E. Cole, C. Monaco and I. Drozdov (2019). Structure-preserving visualisation of high dimensional single-cell datasets. Sci Rep 9(1): 8914.

Tenny, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S., Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In International Conference on Learning Representations.

Vaswani, A., Shazeer, N., Parmar, N., Uszhoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). arXiv:1706.03762v5 [cs.CL]

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

Wolf, T., et al. (2020). HuggingFace's transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771v5.

Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. P*sychology of Learning and Motivation, 58*, 167 – 202.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q.V. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., & Van Durme, B. (2018). Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.