Predicting Word Learning in Children from the Performance of Computer Vision Systems

Sunayana Rane
Department of Computer Science
Princeton University
srane@princeton.edu

Mira L. Nencheva
Department of Psychology
Princeton University
nencheva@princeton.edu

Zeyu Wang
Department of Electrical and Computer Engineering
Princeton University
zeyuwang@princeton.edu

Casey Lew-Williams
Department of Psychology
Princeton University
caseylw@princeton.edu

Olga Russakovsky
Department of Computer Science
Princeton University
olgarus@princeton.edu

Thomas L. Griffiths
Departments of Psychology and Computer Science
Princeton University
tomg@princeton.edu

Abstract

For human children as well as machine learning systems, a key challenge in learning a word is linking the word to the visual phenomena it describes. We explore this aspect of word learning by using the performance of computer vision systems as a proxy for the difficulty of learning a word from visual cues. We show that the age at which children acquire different categories of words is correlated with the performance of visual classification and captioning systems, over and above the expected effects of word frequency. The performance of the computer vision systems is correlated with human judgments of the concreteness of words, which are in turn a predictor of children's word learning, suggesting that these models are capturing the relationship between words and visual phenomena.

Introduction

Both humans and machines face the problem of establishing the relationship between visual and linguistic information. In humans, this process is known as word learning, and has been extensively studied by developmental scientists. In machines, linking visual features with words is a key part of several tasks studied by computer vision researchers, including object classification and image captioning. In this paper, we explore the extent to which the solutions to these problems found by humans and machines are related by predicting the time course of word learning in human children from the performance of computer vision systems.

Developmental scientists have long been interested in understanding how infants and young children learn new words (Bloom, 2002; Brown, 1973; Golinkoff et al., 2000; Quine, 1960; Wojcik et al., 2022), often framing the problem as one of establishing reference between words and their corresponding objects, events, or properties (Markman, 1990; Schwab & Lew-Williams, 2016). While the trajectory of word learning varies across children, there is at least some consistency in the rates at which different kinds of words are learned (Frank et al., 2021). For example, children learning English (as well as many other languages) tend to learn words describing body parts (such as "eye" or "nose") earlier than they learn connecting words (such as "and" or "because"). Developmental scientists have looked for predictors of this pattern. For example, words that are more frequent in child-directed speech tended to be learned earlier (Swingley & Humphrey, 2018). However, the investigation of these predictors has been limited to quantities that can be measured from linguistic input (such as word frequency) or by adults making an intuitive judgment about the properties of words (such as a word's "concreteness" or "abstractness").

Previous work has not made use of predictors that directly measure the correspondence between a word and the visual phenomena it describes. Visual aspects of reference pose a challenge for the child learner because scenes vary in complexity (Quine, 1960) and because the referents of words can be highly variable (e.g., "dog" can refer to both chihuahuas and Bernese mountain dogs). Relatedly, some words and word categories refer to concrete objects (e.g., "cup"), but others do not (e.g., "more" or "fine"), a dimension known to shape age of acquisition (AoA) (Bergelson & Swingley, 2013; Swingley & Humphrey, 2018). Prior experimental work has begun to understand how visual context supports word learning. For example, young children can track word/object co-occurrence statistics over time to disambiguate the meanings of novel labels in complex visual scenes (Chen et al., 2018; Yurovsky et al., 2013), and infants who see more variable views of objects show more rapid vocabulary growth later on (Slone et al., 2019). Although the ease with which a word can be mapped to a concrete visual referent affects children's noun learning, developmental scientists have not formalized how the infant mind may process and create representations of the statistics of visual scenes and labels.

In this paper, we investigate whether we can capture the visual difficulty of learning words by examining the performance of classification and image captioning systems. Since these systems need to solve similar problems to children, they may face the same difficulties. We look at how well object classification and image captioning systems perform for different categories of words (such as animals vs. furniture), and use the resulting performance measures to predict children's word learning. Our results show – across different tasks and architectures – that the difficulty with which machines learn words in different categories is a good predictor of the difficulty with which children learn words in those categories, and that including this measure improves prediction of children's word learning over just using word frequency. We also show that the performance of the computer vision systems is correlated with human judgments of the "concreteness" of a word, which is known to predict AoA. Computer vision models thus

provide an automated measure of this subjective quantity.

While human children and deep neural networks for object classification and image captioning are presumably quite different kinds of systems, discovering parallels in their performance suggests that some aspects of the difficulty with which different kinds of words are learned is a consequence of the nature of the problem itself. Just as the statistics of the linguistic input to children play a key role in understanding language acquisition (Bergelson & Aslin, 2017; Laing & Bergelson, 2020; Montag et al., 2015), the statistics of correspondences between that linguistic input and the world that it describes are significant. Our results demonstrate how improvements in computer vision systems offer new opportunities for the scientific study of child development.

Datasets and models

To investigate model word learning in a way that is relevant to child word learning, we need two kinds of data:

1. Child word learning data, including which words (or word categories) are learned at various developmental stages.

2. Standardized image and natural language data, which can be used to train vision and language models to produce language comparable to child word production.

We address both these needs by working with multiple sources of data: WordBank for child word learning, and COCO for training our models.

Child language acquisition data: WordBank

We use data sourced from the WordBank child language database (Frank et al., 2017) to extract words commonly produced by toddlers between the ages of 16 and 30 months. Figure 1 gives an example of the type of data collected and tracked by WordBank. In particular, the data we use corresponds to which words are easily (and not easily) produced by toddlers of various ages. WordBank contains production percentiles for approximately 1200 words, which we use for our analysis of model word production.

In order to compare word learning at scale, we decided to investigate patterns in word categories instead of individual words. For an effective comparison to child word learning, we use word categories for which there exists parallel child data. Fortunately, the WordBank database contains such data. We extracted approximately 1200 frequently-produced words for toddlers, as listed in the WordBank database, and mapped them to the corresponding category. WordBank categories include people, toys, animals, etc. We remove sounds/sound effects (such as "cockadoodledoo") from these categories, because our models are restricted to vision and language.

Computer vision tasks: COCO

On the computer vision side we use the canonical COCO (**COCO**) image captioning dataset (the Karpathy split, to be precise (Karpathy & Fei-Fei, 2015)) for training and evaluating the models. The dataset contains 113,278

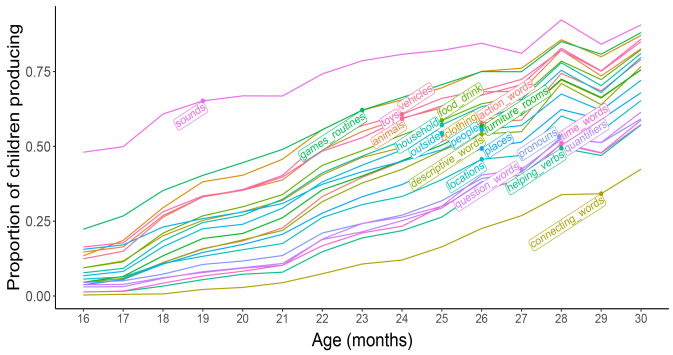

Figure 1: WordBank data tracking production of different categories across different ages. Each line represents the average proportion of children producing all words in a given category on the MCDI. The age at which half of children produce a given word (age of acquisition; AoA) is marked with a circle on each line. We aim to predict the order of acquisition of words in these categories using the performance of computer vision models for the corresponding words.

training and 5,000 validation images, each associated with 5 captions provided by human annotators. We use two computer vision tasks in our experiments: image captioning, where a model is trained to produce a natural language caption, as well as the simpler image classification task, where a model is trained to predict which visual categories are present in the image, without tying these categories to a natural language description. For image classification, we create (imperfect) binary labels comprising of 855 individual words which are in both COCO and WordBank. The binary label is determined by whether the word is mentioned in one of the captions associated with the image.

We run experiments with canonical computer vision models. Our goal is to verify that our findings hold across a range of standard setups. For image classification we use two different CNN architectures: VGGNet (Simonyan & Zisserman, 2014) and ResNet50 (He et al., 2016), both with and without pretraining on ImageNet (Russakovsky et al., 2015). For image captioning, we explore two more complex vision backbones: a ResNet101 CNN (He et al., 2016) or bottom-up features from Faster R-CNN (Anderson et al., 2018; Ren et al., 2015). We combine these models with one of two language models: the classic LSTM (Anderson et al., 2018) or the more recent Transformer (Vaswani et al., 2017).

We use open-source implementations of all models along with the recommended hyperparameters (Chollet et al., 2015; Luo et al., 2021; Luo et al., 2018). Each model is trained on a single GPU. For image classification, we train with an adaptive learning rate using the Adam (Kingma & Ba, 2014) optimizer and dropout until the loss converged. VGGNet trained from scratch proved difficult to train to convergence, despite performing grid search over initial learning rate, dropout, and

batch size, so only pretrained results are reported. In the captioning models, for the LSTM layers, we use an input encoding size of 1000, a hidden size of 512, a batch size of 10, and an adaptive learning rate. For the Transformer layers, we use 8 attention heads, 6 encoder and decoder layers, 512 hidden unit size, and a batch size of 10. We keep as much consistent as possible between the implementations, so that we can have a meaningful comparison.

Metrics

To quantitatively measure the correspondence between word learning in human children and computer vision systems, we adopted standard metrics used in the relevant fields: the median age at which children produce a word, and the word-level performance measures of AUC for classifiers and SPICE for captioning systems.

Metric for children: AoA (Age of Acquisition)

The Age of Acquisition (AoA) of a word is defined as the age at which at least 50% of children produce the word. WordBank includes the vocabularies of 5520 toddlers learning North American English assessed using parent report on the MacArthur Bates Communicative Development Inventory (MCDI) (Fenson et al., 2007). In the WordBank database, AoA can be calculated over the parent-reported scores for word learning for each child in the database. AoA has previously been shown to correspond with the difficulty of learning to read a word (Coltheart et al., 1988). We use this measure of AoA as a proxy for the difficulty of learning a word for a child. AoA was calculated for each word and then averaged within each of the WordBank categories: body parts, animals, vehicles, toys, household, outside, food/drink, furniture/rooms, clothing, locations, descriptive words, places, people, action words, pronouns, question words, quantifiers, helping verbs, time words, and connecting words.

Metrics for machines: AUC and SPICE

The metrics we use for our models are designed to measure a model's performance at the level of individual words. For classification models, we use AUC (the area under the receiver operating characteristic (ROC) curve (Freedman, 2009)) as a classification metric which is robust to class imbalance. Our multi-label, multi-class classification task was binary for each label, so a per-word AUC calculated over each label was an appropriate metric.

For captioning models, we use the Semantic-Propositional Image Caption Evaluation score (SPICE) (Anderson et al., 2016). SPICE is an automatic evaluation metric which uses scene graphs corresponding to the actual image to gauge semantic and propositional correctness, instead of just the textual n-gram comparison of previous metrics. From a caption like "woman sitting on a brown chair in a restaurant", SPICE produces a tuple-based scene graph containing tuples like "woman-sitting" and "sitting-on-chair." SPICE then calculates whether each produced tuple matches one of the tuples that appear in the ground truth manual captions. To get a score for each individual WordBank word, we then average the tuple-based scores of all the tuples where the word appears. The intuition for using this metric is that it is impossible to gauge whether a word like "sitting" is used correctly without looking at the other words around it.

For both AUC and SPICE, after calculating at a word-level (or a tuple-level, for SPICE) we then aggregate over WordBank categories for ease of comparison to AoA for those same categories.

Table 1: Classification model AUC correlation with AoA

Classification model setup					
Model	Training	Pearson	p	Spearman	p
VGG	Pretrained	-0.280	0.232	-0.311	0.182
ResNet50	From scratch	-0.138	0.562	-0.081	0.734
ResNet50	Pretrained	-0.531	0.016	-0.544	0.013

Table 2: Captioning model SPICE correlation with AoA

Captioning architecture					
Pretrained Visual Features	Language Layers	Pearson	p	Spearman	p
ResNet 101	LSTM	-0.515	0.034	-0.640	0.006
Bottom-up (Faster R-CNN)	LSTM	-0.617	0.004	-0.708	0.000
Bottom-up (Faster R-CNN)	Transformer	-0.565	0.012	-0.624	0.004

Results

As an initial analysis, we examined the raw correlation between AoA and the machine metrics. We then conducted a series of multiple regression analyses to determine whether computer vision systems can improve prediction of the time-course of child word learning over existing measures used in the child language acquisition literature.

Correlations

To compare the word category-level AoA to AUC/SPICE of the models, we report two types of correlation: the Pearson correlation coefficient, which assumes a linear relationship, and the Spearman rank-order correlation coefficient, which only assumes monotonicity (Freedman, 2009). The results are shown in Table 1 for classification and Table 2 for captioning, with corresponding scatterplots in Figure 2.

Four of our models showed statistically significant correlations with AoA: classification ResNet50 with pretrained features, and all three of the captioning models. In all four of these models, as performance (AUC or SPICE) increased, AoA decreased: categories of words that were easier for the models were acquired earlier by children. The remaining two classification models (VGG with pretrained features and ResNet50 trained from scratch) also showed correlations consistent with this relationship, but those correlations were not statistically significant. The correlation for ResNet50 trained from scratch was particularly weak, suggesting that pretrained features may be important. The correlations for the captioning systems were all of similar magnitude, suggesting that the specific architecture (including the choice of an LSTM or Transformer) may be less relevant than the cap-

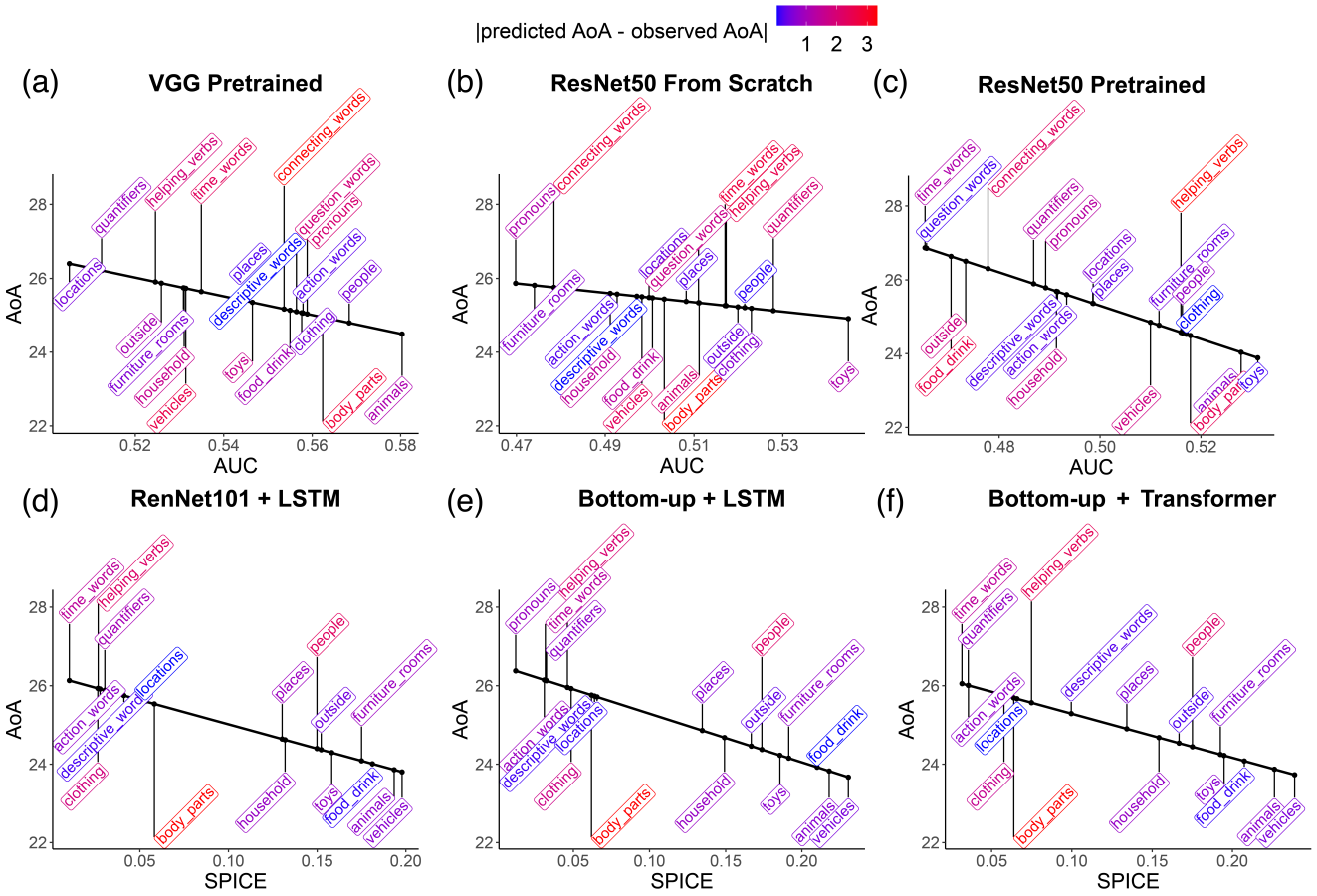

Figure 2: Regression of successful models' performance vs age of acquisition in months, per category. Each category label is placed so that its bottom left corner indicates the corresponding age of acquisition (AoA) and AUC (classification) or SPICE (captioning) values. The black regression line shows the AoA predicted by each model in the regression, and the vertical residual lines and category labels show the observed average AoA for the category.

tioning task itself.

Comparison with other predictors

As noted above, developmental psychologists have explored variables that predict children's word learning. One such variable is the frequency with which words appear in child-directed speech. We evaluated the correlation between word frequency (extracted from the TalkBank database (MacWhinney, 2007)) and AoA, finding a Pearson correlation of $r =$ -0.377 ($p = 0.092$) and a statistically significant Spearman correlation of $\rho =$ -0.494 ($p = 0.022$).

To determine whether our computer vision metrics (AUC and SPICE) are predictive of AoA even after accounting for word frequency, we conducted a multiple regression analysis where the independent variables are word frequency and AUC/SPICE, and the dependent variable is AoA. The coefficients of the multiple regression analysis show that AUC/SPICE across different successful models do indeed predict AoA over and above frequency in child-directed data. The results are shown in Table 3. The multiple regression analysis showed that all four predictors that originally resulted in a statistically significant correlation with AoA remained statistically significant when word frequency was taken into account. Notably, word frequency was no longer a statistically significant predictor in the resulting models.

Table 3: Multiple regression with TalkBank word frequency as a predictor in addition to AUC or SPICE

Regression Coefficients (predicting AoA)					
Model	TalkBank frequency	p	AUC / SPICE	p	R^2
VGG Pretrained	-1.891	0.128	-1.580	0.306	0.199
ResNet50 From Scratch	-2.160	0.087	-1.243	0.431	0.178
ResNet50 Pretrained	-1.286	0.269	-2.559	0.044	0.333
ResNet101 + LSTM	-1.297	0.267	-2.148	0.050	0.329
Bottom-up + LSTM	-1.225	0.229	-2.769	0.007	0.433
Bottom-up + Transformer	-1.552	0.153	-2.355	0.027	0.404

Another variable that has been shown to be a good predictor of AoA is the "concreteness" of words (Bergelson & Swingley, 2013; Swingley & Humphrey, 2018). Unlike word frequency, concreteness is not a property that can be mea-

Table 4: Multiple regression with concreteness judgments as a predictor in addition to AUC or SPICE

Regression Coefficients (predicting AoA)					
Model	AUC SPICE	p	Concreteness	p	R^2
VGG Pretrained	-0.128	0.884	-3.925	0.000	0.754
ResNet50 From Scratch	-0.409	0.631	-3.926	0.000	0.757
ResNet50 Pretrained	-0.940	0.213	-3.601	0.000	0.776
ResNet101 + LSTM	0.620	0.574	-3.996	0.002	0.631
Bottom-up + LSTM	0.488	0.668	-4.024	0.002	0.652
Bottom-up + Transformer	0.544	0.621	-3.897	0.002	0.629

Table 5: Correlation of AUC/SPICE with human judgments of concreteness

Model	Pearson	p	Spearman	p
VGG	0.303	0.195	0.277	0.238
ResNet50 From Scratch	0.092	0.701	0.056	0.816
ResNet50 Pretrained	0.459	0.042	0.421	0.064
ResNet101 + LSTM	0.733	0.001	0.598	0.011
Bottom-up + LSTM	0.744	0.000	0.659	0.003
Bottom-up + Transformer	0.690	0.002	0.574	0.016

sured directly from the linguistic input to children. Rather, it is typically measured by asking human raters to rate on a scale how "concrete" or "abstract" they consider a word to be, typically after providing a definition for concrete (e.g. can be experienced with the five senses) and abstract words (e.g. cannot be experienced through the five senses) (Brysbaert et al., 2014). Some work has expanded these rating lists by using supervised models trained directly to predict concreteness (Köper & im Walde, 2017). We ran a second multiple regression using a standard measure of concreteness (taken from Köper and im Walde, 2017) as a predictor. The results are shown in Table 6. In this model, concreteness was the only statistically significant predictor, with neither word frequency nor our measures being statistically significant. The same result is observed in a multiple regression incorporating only concreteness and AUC/SPICE as predictors 4.

Investigation of this result revealed that it is a consequence of substantial collinearity between the computer vision measures and concreteness ratings. The correlations between AUC/SPICE and concreteness are shown in Table 5. The four models that produced statistically significant correlations with AoA are all correlated with concreteness, with statistically significant correlations from all the captioning models.

The observed correlation between concreteness and AUC/SPICE makes sense: concreteness is people's judgment of how well a word corresponds with a visible or tangible thing in the world, and this is what our measures reflect. A high correlation with concreteness thus indicates that our models are capturing what we intended: the ease of relating a word to its visual referent(s). Importantly, the performance of a computer vision model is an objective quantity that can be estimated directly from a dataset, rather than a subjective quantity that requires additional judgments from people. Further, our approach captures meaningful variability within the visual contexts of concrete and abstract words. For example, the words "hello" and "economy" may be judged as equally abstract (Köper & im Walde, 2017), however, "hello" may be associated with a more consistent visual context as part of a routine (e.g., waving) compared to "economy". Similarly, concrete words like "spoon" may have a more consistent surrounding visual context (e.g., a kitchen), compared to words like "dog", which may be encountered in many different visual contexts. Future work can apply this approach to go beyond category-level estimates and capture the visual variability of different items, as well as individual differences in the visual contexts that different children experience in densely sampled child-view visual corpora (e.g. Sullivan et al., 2022). By providing a new way to directly measure the concreteness of words, our approach provides a novel metric that can be used in the broader investigation of language processing.

Discussion

We have shown that despite training on only standard machine learning datasets (ImageNet and COCO), several captioning models and one classification model successfully predict the age at which children acquire different categories of words. This result holds across multiple architectures, and for both simple and complex models. This indicates that these models effectively capture the visual difficulty of learning a word for a child. It also suggests that the underlying mechanisms of learning for models and children might be similar in ways that are not yet fully understood but result from the shared statistical structure of the problems they face.

Figure 2 provides some intuition for why visual difficulty goes beyond training data distribution: for example, while the categories 'food/drink' and 'descriptive words' occur much more frequently in child-directed speech than in the COCO training data, the models are nevertheless successfully predictive of AoA for those categories. This illustrates the value of ML approaches to concreteness, and provides some intuition for the commonalities in child and model learning. Certain categories are also difficult for both models and children, despite those categories being *over*represented in the training data. For example, quantifiers are difficult for both models and children to learn, despite being well represented in COCO training data and child-directed speech.

Pretraining seems to be one of the key differentiating factors between the models which showed substantial correlation and those which did not, such as the ResNet50 model pretrained and the same ResNet50 trained from scratch. There are several potential reasons for this. If pretraining (even on ImageNet alone) supports learning to extract visual features, that skill can be applied to more complex visual features than those in the training data. However, pretraining is not the whole story: the difference between pretrained VGGNet and ResNet50 classification models' correlation to AoA shows that architecture does contribute to the correlation as well.

The consistently high correlations for captioning models with language components support anecdotal evidence that these larger models combining vision and language modalities do indeed produce more human-like performance for vi-

Table 6: Multiple regression with all variables as predictors of AoA: TalkBank word frequency, AUC/SPICE, and concreteness

	Regression Coefficients (predicting AoA)						
Model	TalkBank frequency	p	AUC / SPICE	p	Concreteness	p	R^2
VGG Pretrained	-1.111	0.096	-0.015	0.986	-3.747	0.000	0.794
ResNet50 From Scratch	-1.170	0.077	-0.594	0.461	-3.704	0.000	0.801
ResNet50 Pretrained	-0.967	0.144	-0.661	0.374	-3.530	0.000	0.804
ResNet101 + LSTM	-1.469	0.076	0.901	0.383	-4.099	0.001	0.713
Bottom-up + LSTM	-1.239	0.126	0.647	0.552	-4.106	0.001	0.707
Bottom-up + Transformer	-1.445	0.082	0.782	0.447	-3.951	0.001	0.709

sual word learning. The high correlation across different architectures opens the door to future investigations as to why exactly this is the case – it is clearly not one particular element, such as a transformer language model, which yields this result. However, sophisticated language modeling with attention mechanisms, present in all the captioning models either through the LSTM or Transformer layers, may be important for producing these results.

Relationship to Previous Work

While there has been no previous work looking directly at predicting AoA from metrics derived from computer vision models, there is an extensive literature in cognitive science and computer vision examining different kinds of correspondences between humans and machines. For example, representations from image classification systems have been used to predict human judgments of image typicality (Lake et al., 2015), the similarity between images (Hebart et al., 2020; Jozwik et al., 2017; Peterson et al., 2018), image classification (Battleday et al., 2020; Sanders & Nosofsky, 2020), and neural responses to images (Schrimpf et al., 2020; Yamins & DiCarlo, 2016). Better capturing these aspects of human cognition has been shown to result in improvements in computer vision applications (Peterson et al., 2019). Developmental research has also previously explored the use of deep neural networks to capture aspects of children's language learning, particularly systems that are trained on data from cameras mounted on the heads of infants (Bambach et al., 2018; Orhan et al., 2020; Tsutsui et al., 2020). This research has primarily focused on visual object learning rather than predicting the timecourse of word-learning itself. Other work has looked at using multimodal neural networks to capture human performance in stylized word-learning settings (Vong & Lake, 2022). This work provides a converging perspective on how models from computer vision can be used to capture the relationship between linguistic and visual input.

Future Work

In demonstrating how vision and language models' effectively capture word learning difficulty in children, this work also opens the door to more behavioral comparisons of word learning in children and computer vision models. We have demonstrated this result on a standardized group of datasets, with standard pretraining protocols. An important future question is, to what extent particular architectural components (ResNet/Faster R-CNN visual features, or LSTM/Transformer language layers) are important for capturing different facets of child word learning. Is it the scale of larger captioning models which yields the robust similarity to child word learning? Or is it the attention mechanisms in the more sophisticated language components? Another important line of inquiry is how this result changes with datasets; it is surprising that this correlation exists although children and models are certainly exposed to different data. Training models on a child-directed dataset, such as SAYCam (Sullivan et al., 2022) is likely to strengthen the correlation to child word learning patterns. Our results lay the groundwork for further behavioral comparisons between models and child learning.

Conclusion

We have shown that the difficulty with which computer vision models learn different categories of words predicts the age at which children learn words in those categories. Although computer vision systems and human children potentially have significant differences in the mechanisms of learning, both face the challenge of relating a word to the visual phenomena it describes. The developmental parallels, which show that the difficulty of learning different categories of words is aligned for both machines and children, suggest that the structure of the learning problem itself may induce similarities in patterns of learning. We hope that these results open the door to new opportunities to model child development using machine learning systems for computer vision and language, and in turn help us to understand these machine learning systems better through their parallels to child development.

Acknowledgments. Mira Nencheva was supported by the ACM SIGHPC Computational & Data Science Fellowship. This material is based upon work supported by the National Science Foundation (grant number 2107048) and the National Institute of Child Health and Human Development (grant number R01 NICHD 095912). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies.

References

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. *European Conference on Computer Vision*, 382–398.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems*, *31*.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), 1–14.

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397.

Bloom, P. (2002). *How children learn the meanings of words*. MIT Press.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Chen, C.-h., Zhang, Y., & Yu, C. (2018). Learning object names at different hierarchical levels using cross-situational statistics. *Cognitive science*, *42*, 591–605.

Chollet, F., et al. (2015). Keras.

Coltheart, V., Laxon, V. J., & Keating, C. (1988). Effects of word imageability and age of acquisition on children's reading. *British Journal of Psychology*, *79*(1), 1–12.

Fenson, L., et al. (2007). *Macarthur-bates communicative development inventories*. Paul H. Brookes Publishing Company Baltimore, MD.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Freedman, D. A. (2009). *Statistical models: Theory and practice*. Cambridge University Press.

Golinkoff, R. M., Hirsh-Pasek, K., Bloom, L., Smith, L. B., Woodward, A. L., Akhtar, N., Tomasello, M., & Hollich, G. (2000). *Becoming a word learner: A debate on lexical acquisition*. Oxford University Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Köper, M., & im Walde, S. S. (2017). Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, 24–30.

Laing, C., & Bergelson, E. (2020). From babble to words: Infants' early productions match words and objects in their environment. *Cognitive Psychology*, *122*, 101308.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Luo, R., et al. (2021). An image captioning codebase [MIT License].

Luo, R., Price, B., Cohen, S., & Shakhnarovich, G. (2018). Discriminability objective for training descriptive captions. *arXiv preprint arXiv:1803.04376*.

MacWhinney, B. (2007). The talkbank project. In *Creating and digitizing language corpora* (pp. 163–180). Springer.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*(9), 1489–1496.

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, *33*, 9960–9971.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., & Russakovsky, O. (2019). Human uncertainty makes classification more robust. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.

Quine, W. V. O. (1960). *Word and object*. MIT Press.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal

networks. *Advances in Neural Information Processing Systems*, *28*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, *3*(3), 229–251.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Schwab, J. F., & Lew-Williams, C. (2016). Repetition across successive sentences facilitates young children's word learning. *Developmental Psychology*, *52*(6), 879.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental Science*, *22*(6), e12816.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2022). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, *5*, 20–29.

Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants' learning of specific english words. *Child Development*, *89*(4), 1247–1267.

Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D., & Yu, C. (2020). A computational model of early word learning from the infant's point of view. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vong, W. K., & Lake, B. M. (2022). Cross-situational word learning with multimodal neural networks. *Cognitive Science*, *46*(4), e13122.

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1596.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, *16*(6), 959–966.