EFFECTS OF LINGUISTIC LABELS ON LEARNED VISUAL REPRESENTATIONS IN CONVOLUTIONAL NEURAL NETWORKS: LABELS MATTER!

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigated the changes in visual representations learnt by CNNs when using different linguistic labels (e.g., trained with basic-level labels only, superordinate-level only, or both at the same time) and how they compare to human behavior when asked to select which of three images is most different. We compared CNNs with identical architecture and input, differing only in what labels were used to supervise the training. The results showed that in the absence of labels, the models learn very little categorical structure that is often assumed to be in the input. Models trained with superordinate labels (vehicle, tool, etc.) are most helpful in allowing the models to match human categorization, implying that human representations used in odd-one-out tasks are highly modulated by semantic information not obviously present in the visual input.

1 INTRODUCTION

When compare to the performance of many classification models in computer vision, human classification is considerably more flexible and efficient. In fact, humans can learn new categories with just a few examples (i.g., zero-shot or few-shot learning) and this category knowledge can be transferred to new exemplars (Ashby & Maddox, 2005; Ashby & Ell, 2001). However, this job is not easy at all for most classification models because they are usually biased towards learning basic or subordinate-level features which are hardly generalized to new higher-level categories. Understanding human category learning and obtaining human-like visual representation is therefore important task for both behavioral and computer vision.

What can be so different about category learning between humans and machines? One possible difference is language. Human learning goes beyond the one-to-one correspondence of perceptual stimulus and cue; Human uses language and the semantic information it conveys, and by doing so they could actively seek and identify the relationship between various objects in the world (Hays, 2000; Levinson, 1997; Lupyan & Lewis, 2017). In computer vision, of course, especially under the Zero-shot and Few-shot learning task, many attempts have been made to learn complex semantic relationships between objects using relational information (Sung et al., 2018; Annadani & Biswas, 2018), attribute labels (Lampert et al., 2013; Akata et al., 2015; Chen et al., 2018), and word vectors (Frome et al., 2013) to increase the generalizability of the model's performance.

However, few studies have systematically studied how different patterns of labels influence what models exposed to the same visual inputs learn (but see Peterson et al., 2018) In this study, we trained the equivalently designed CNNs with different types of labels and explored how the visual representations learnt by these models are distributed – how comprehensive and separable each category cluster is. We also collected human similarity judgements in the Odd-one-out task where the person had to select which of three images is most different. With this dataset and using categorical representations extracted from our trained models, we could predict human similarity decisions fairly well with the highest accuracy of 74% and understand which labeling schemes produce the most human-like representation.

2 MODEL TRAINING

The goal of this study is to examine how linguistic label changes the learnt visual representation in Convolutional neural network(CNN). In order to achieve this, we trained the equivalently designed CNNs for classification, but each time with the different linguistic labels as groundtruth. In addition, we trained Convolutional autoencoder (Conv AE), which also encodes the images using the the same Convolutional structure as the other models do but instead of being supervised to predict the class of image, the aim of this model is to generate the same output image at the input. This Conv AE represents in a sense the model not trained with any linguistic label at all, compared to the other models given some types of linguistic labels. The description of each model and labels used for training are provided below.

- **Convolutional Autoencoder (CAE)**: Autoencoder with Convolutional encoder and decoder trained to output the same image as input
- **Basic CNN (Basic)**: CNN model trained with one-hot encoding of basic-level categories, n=30
- **Superordinate CNN (Super)**: CNN model trained with one-hot encoding of superordinate-level categories, n=10
- **Combined basic and superordinate CNN (Combined)**: CNN model trained with two-hot encoding of both basic and superordinate-level categories, n=40(10+30)
- **Basic-Superordinate CNN (Basic-Super)**: CNN model trained with one-hot encoding of basic-level categories first (n=30), and then finetuned with one-hot encoding of superordinate categories (n=10)
- **Word vector CNN (Wordvec)**: CNN model trained with basic-level word vectors extracted from Fasttext word embedding model (Bojanowski et al., 2017), dimension=300

Across the different labeling conditions, the architecture of CNNs remained exactly the same, except for the output layer and its activation function. The general pipeline used for CNNs is described in the Figure 1. Our CNN models consist of five blocks of two Convolutional layers followed by Max pooling and Batch normalization layers. Through all Convolutional and Max pooling operations, the zero padding was employed to produce the output feature maps with the same size of the input. The output of Convolutional layer, the ”bottleneck” feature which later was extracted and analyzed for studying model’s visual representation (dim=1568), was then fed into one fully connected dense layer. Rectified linear units (ReLU) was used as activation function after each convolution. The output activation function differs depending on which linguistic labels are used: softmax function for Basic, Super, and Basic-Super CNN, sigmoid function for combined Combined CNN, and linear function for Wordvec CNN. For CAE, the same Convolutional architecture was employed for encoder and decoder part, with the hidden layer in the model (dim=1568) serving as bottleneck feature for analysis. For output function in CAE, linear function was used.

All models are trained and validated on the images of 30 categories from IMAGENET 2012 dataset (Deng et al., 2009), and tested on the images of the same 30 categories from THINGS dataset (Hebart et al., 2019). These 30 basic-level categories can be grouped into 10 higher-level categories – superordinate-level, including ’mammal’, ’bird’, ’inset’, ’fruit’, ’vegetable’, ’vehicle’, ’container’, ’kitchen appliance’, ’musical instrument’, and ’tool’. A full list of 30 categories with their superordinate-level categories are provided in the Appendix. All input images were converted from RGB to BGR and then zero-centered each channel with respect to the ImageNet dataset. Different loss function was used for training each model: Categorical Crossentropy loss for Basic, Super, and Basic-Super CNN, Binary Crossentropy loss for Combined CNN, and Mean Squared Error loss for both Wordvec CNN and CAE model. All models were trained using a version of optimization algorithm Adam (Kingma & Ba, 2014), using the mini-batch size of 64. During training, early stopping was implemented and the model with the lowest validation loss was used for the following analysis.

3 BEHAVIORAL DATA

To compare visual representation of our trained models with that of human, we also collected human similarity judgements in Odd-one-out task, as done in Zheng et al. (2019). In Odd-one-out task, the

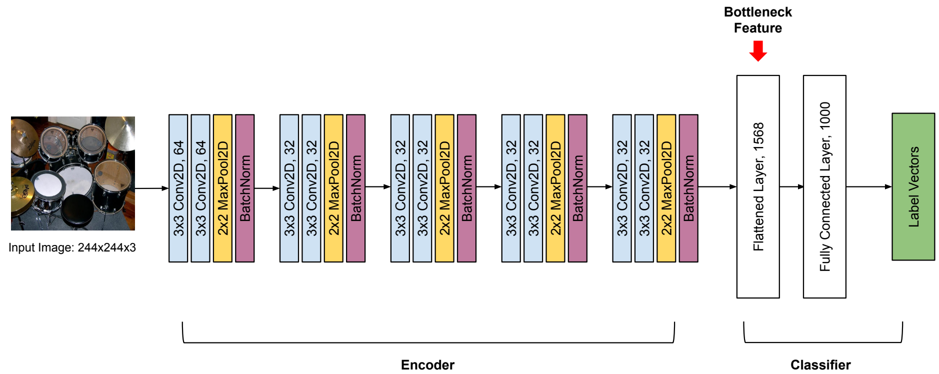


Figure 1: General pipeline for CNNs used for the study. Rectified linear units (ReLU) was used as activation function after each convolution. For final classification, we used softmax function for basic and superordinate category classification, sigmoid function for combined basic and superordinate category classification, and linear function for word vector prediction. The other architecture remained the same across tasks

participant was presented three images, triplet, and was asked to choose which image is the most different from the other two. The triplet consisted of three exemplar images from 30 categories used for our model training. Almost all exemplar images used for the data collection came from Zheng et al. (2019), but for 'crate', 'hammer', 'harmonica', and 'screwdriver' images were replaced with new one to increase image quality and category representativeness. There are 4060 possible triplets in total that can be generated from all 30 categories, but we collected human data on a subset of them to reduce time and cost of data collection. This subset includes 1) all ten triplets where three images came from the same superordinate category e.g., 'orangutan', 'lion', 'gazelle' 2) all 435 triplets where two images came from the same superordinate category e.g., 'orangutan', 'lion', 'minivan', and 3) 1375 triplets where all images came from different categories e.g., 'orangutan', 'minivan', 'lemon', making 1820 unique triplets in total. 51 Amazon Mechanical Turk(AMT) workers participated in this task, each making responses on $\sim$200 triplets. After removing the responses with RT below 500ms, we collected 9697 similarity judgements data with each triplet viewed by 5.6 workers on average (min =4 , max=51).

4 Experiments

4.1 Evaluating model performance

Although our goal is not to beat the state of art vision model in classification, we evaluated classification accuracy so as to confirm the validity of learnt visual representations of our trained models i.e., to check if models successfully gained categorical knowledge to the extent that it could show the actual effects of different labels on learning. For evaluating classification accuracy, we reported results on several metrics: 1) top@k – the percentage of accurate classification on test images where the true class needs to be in the top K prediction for it to be counted as accurate, 2) average precision and 3) average recall over all categories. All metrics are computed over on the test dataset (THINGS; Hebart et al., 2019). Since Wordvec CNN is predicting word vector, not class, its classification performance was approximated by calculating cosine similarity between predicted and true word vectors and choosing the corresponding class from top@k similarities. The classification results from CAE was not reported, because it aims to generate the input-like images, not to predict the class of image. A few examples of image generated from CAE are attached in the Appendix. As can be seen in Table 1, all trained models performed classification fairly well (all models top@5 acc $>.82$), although there's still room from improvement in classification for Wordvec CNN.

Table 1: **Classification accuracy results from trained models.** Exact match accuracy is the same as top@2 accuracy from Combined CNN and the same as top@1 accuaracy for the other models. Precision and recall reported here were sample-wise averaged for Combined CNN and micro-averaged for the other models.

Model type	**# class**	**Accuracy**			**Average Precision**	**Average Recall**
		exact match	**top@3**	**top@5**		
Basic CNN	30	0.90	0.98	**0.99**	0.90	0.90
Super CNN	10	**0.95**	**0.99**	**0.99**	0.94	0.94
Combined CNN	40	0.91	0.95	0.97	0.91	0.91
Basic-Super CNN	10	**0.95**	**0.99**	**0.99**	**0.95**	**0.95**
Wordvec CNN	30	0.52	0.74	0.82	0.52	0.52

4.2 EXPLORING VISUAL REPRESENTATIONS

To explore how the model's visual representations change as different linguistic labeling schemes are deployed, we extracted and analyzed on the bottleneck features from each model i.e., the final output of Convolutional layer with the feature vector dimension equal to 1568 (see Figure 1). For analysis, we first measured representational similarity of all images in the training dataset (IMAGENET 2012; Deng et al., 2009) between/within category. These representational distributions were visualized using t-SNE (Maaten & Hinton, 2008) which are attached in Appendix. We also analyzed the similarity between categorical representations by plotting similarity matrix. To create categorical representations, we simply averaged the obtained bottleneck features from all training images per category, creating in a sense "prototypical" representation for each class.

Representational similarity between/within category

To investigate how distinct semantic labeling tighten or loosen the cluster of visual representations of models, we computed the cosine distance of all images between/within category and its ratio. As shown in Table 2, the ratio of between to within category distance is higher in overall when computed using basic-level taxonomy, compared to superordinate-level. This result implies that the cluster of visual representations in basic-level category is more dense and tightened in general, which resonates with previous psychological findings ascribing the frequent usage of basic-level taxonomy to utility maximization behavior, i.e., basic-level category has relatively good discriminability while remaining abstract enough to be generalized to multiple exemplars (Corter & Gluck, 1992).

If comparing the results between our trained models, the categorical representation of Wordvec CNN was observed to be the most tightly clustered as evidenced by its highest value of between/within ratio, with Basic-Super CNN and Super CNN achieving the next highest numbers. Interestingly, when the model is trained with both basic and superordinate labels at the same time, its categorical representation became more scattered and less distinguishable to each other, compared to other models trained with linguistic labels. Lastly, CAE produced the lowest between/within ratio value, suggesting that even if CAE had successfully learnt visual features that are enough to generate input-like images, these visual representations are poorly discriminable in both basic and superordinate levels.

Visualization of categorical representations:

To examine whether the hierarchical semantic structure of 30 categories (e.g., every category belongs to one of ten superordinate categories) are reflected in the visual representations learnt by models, we visualized categorical representations using the cosine similarity matrix in Figure 2. For more complete comparison, we also added the results using visual features extracted from FastText word vectors (Bojanowski et al., 2017) and early VGG16 layer (Simonyan & Zisserman, 2014) i.e., the output from the first max pooling layer. A clear difference in categorical representations was observed depending on whether the models trained with linguistic labels or not; while no hierarchical pattern was observed for both early Vgg16 and CAE features, various semantic structures were observed in the others, e.g., dark squares recurrently emerged in different hierarchies dividing 1) nature vs non-nature, 2) edible vs non-edible and 3) superordinate categories. Interestingly, despite

Table 2: **Between/within category distance and its ratio.** Cosine distance (1-cosine angle of two feature vectors) was used for distance metric. As the value gets larger, the visual representations of images becomes less similar between/within category

Model type	By superordinate category			By basic category		
	between	**within**	**between/within**	**between**	**within**	**between/within**
CAE	0.02	0.19	0.11	0.03	0.19	0.15
Basic CNN	0.36	**0.55**	0.64	0.43	0.52	0.84
Super CNN	0.33	0.47	0.71	0.36	0.46	0.80
Combined CNN	0.29	0.48	0.61	0.35	0.45	0.78
Basic-Super CNN	**0.40**	0.53	0.76	**0.46**	**0.51**	0.90
Wordvec CNN	0.36	0.37	**0.95**	0.40	0.35	**1.14**

Wordvec CNN in Figure 2h being trained on the same FastText word vectors described in Figure 2a, their representations are very different. Having visual information as well as linguistic supervision, Wordvec CNN demonstrated more semantically structured visual representations.

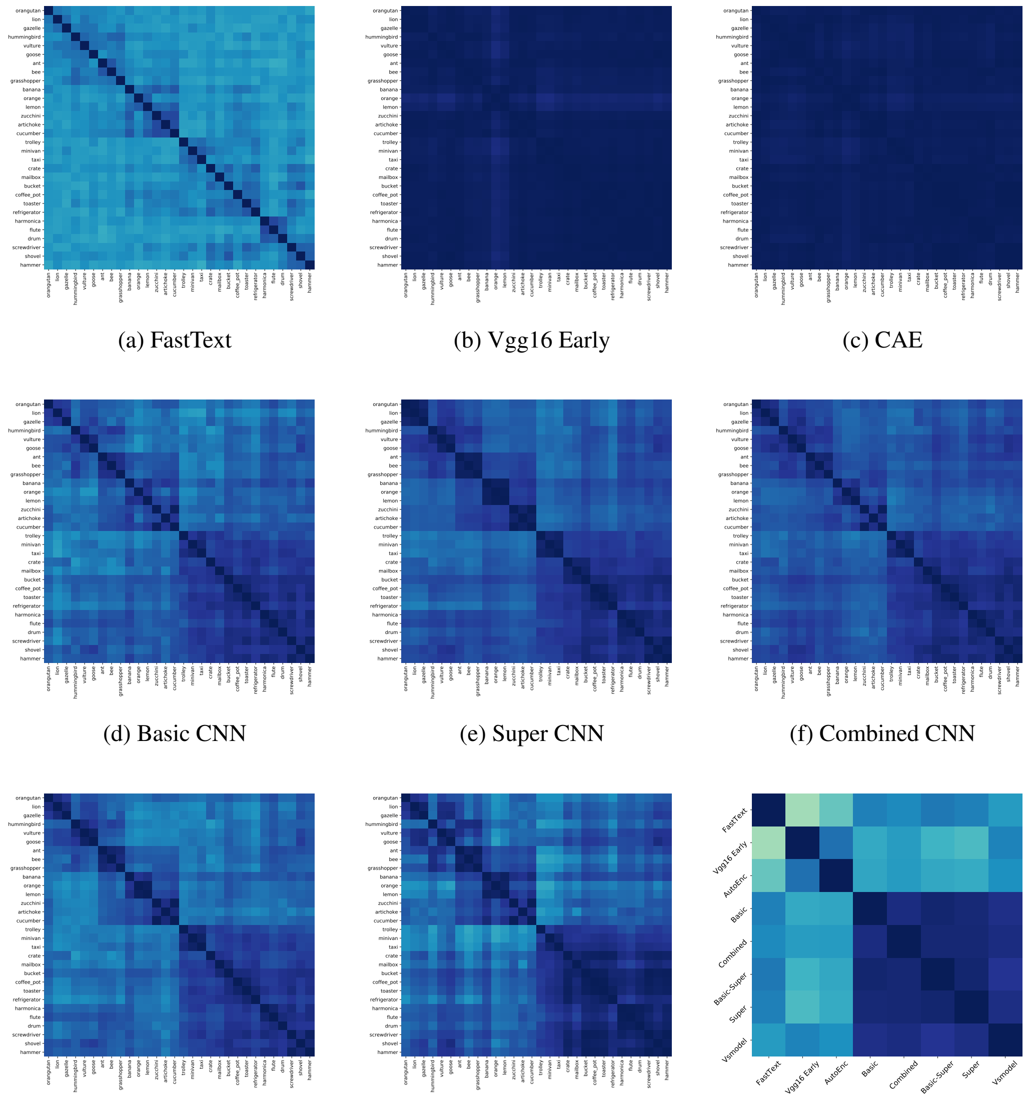
4.3 PREDICTING HUMAN VISUAL BEHAVIOR

Finally, we compared the visual representations learnt by our models with human representation by evaluating how well they can predict human similarity judgements in the Odd-one-out task (See Section 3). The response from models was generated by comparing cosine similarities between three visual representations given a triplet of three images and selecting the most dissimilar one from the others. For comparison, three kinds of visual representations are computed 1) IMAGENET categorical representations, where features were averaged over $\sim$1000 images per category from IMAGENET training dataset (Deng et al., 2009) THINGS categorical representations, where features were averaged over $\sim$10 images per category from THINGS dataset (Hebart et al., 2019), and 3) Single Exemplar representation, where only one feature per category was generated using 30 exemplar images used for behavioral data collection. Together with the results from FastText (Bojanowski et al., 2017) and Vgg16 Early Layer (Simonyan & Zisserman, 2014), upper and lower bound and baseline results were reported as below.

- **Null Acc**: Accuracy that could be achieved by predicting every sample as the one most frequent class in the dataset, lower bound results, 35%.
- **Bayes Acc**: Accuracy that could be achieved by predicting the sample as the most frequent class in each unique triplet set, upper bound results, 84%.
- **SPoSE Acc**: Accuracy that could be achieved by using the SPoSE model (Zheng et al., 2019), a probabilistic model that is directly trained on human responses on all triplets from 1854 THINGS objects, 80%.

As shown in the Figure 3, triplet prediction accuracy of all models was highest when IMAGENET categorical representations were used and lowest when single exemplar representations were used. Comparing the model performance on human triplet data, our trained model performed fairly well: highest accuracy (74%) was achieved by Super CNN. This performance is even more promising when considering that these models were not trained on human data itself as was the SPoSE model whose performance was around 80%. Overall, the CNNs trained with superordinate categories like Super CNN or Basic-Super CNN achieved higher accuracy, while CAE and Vgg16 Early did not. The results together suggest that the representations that humans use in a visual task are highly semantic in fact, leveraging whole categorical information especially in a superordinate level.

To further investigate the influence of semantics and superordinate-level information on model performance, we broke down the triplet data into six conditions: (1) by the number of superordinate categories that a triplet belongs to (NSUPER), e.g., For a triplet of ’orangutan’, ’lion’, ’gazelle’, NSUPER equals to 1 (’mammal’), for a triplet of ’orangutan’, ’lion’, ’lemon’, NSUPER equals to 2 (’mammal’ and ’fruit’), and for a triplet of ’orangutan’, ’lemon’, ’minivan, NSUPER equals to 3 (’mammal’,’fruit’,’vehicle), and (2) by the accuracy of FastText predictions (FastText Correct). As



(a) FastText (b) Vgg16 Early (c) CAE

(d) Basic CNN (e) Super CNN (f) Combined CNN

(g) Basic-Super CNN (h) Wordvec CNN (i) Model Comparison

Figure 2: **Visualizations of cosine similarity matrix between 30 categorical representations.** When drawing similarity matrix, the categories from the same superordinate group are placed near together with the order of 'mammal', 'bird', 'insect', 'fruit', 'vegetable', 'vehicle', 'container', 'kitchen appliance', 'musical instrument', and 'tool' (from left to right on x-axis and from top to bottom on y-axix). Darker color denotes higher similarity.

reported in the Table 3, when a triple came from all different three superordinate categories, the best accuracy was achieved by SPoSE model. However, when the response was made on a triplet with one unique superordinate category, the response cannot be explained by semantic similarity by FastText predictions, the performance of our supervised models was actually better, especially if using visual representations from CAE or Super CNN. When there were two unique superordinate categories in a triplet and only one of the image came from the different category, the human responses were best predicted by the methods using the which superordinate class each image belongs to.

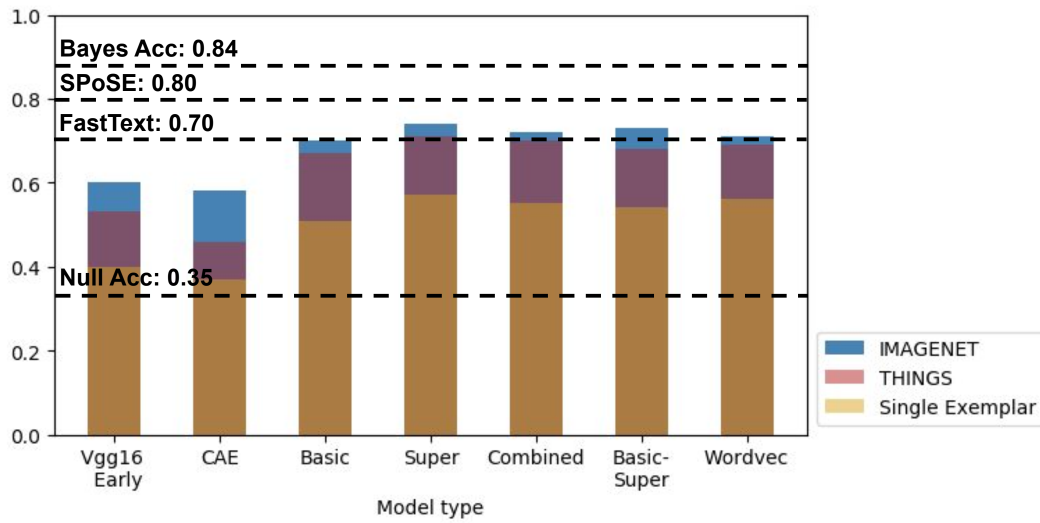


Figure 3: **Comparison of triplet prediction accuracy**. IMAGENET: using categorical representations averaged over IMAGENET training dataset ($\sim$ 1000 images per category). THINGS: using categorical representation averaged over THINGS dataset ($\sim$10 images per category). Single Exemplar: using visual representation of single image used for behavioral data collection. Other baseline accuracy are drawn in dashed lines.

Table 3: **Triplet prediction accuracy**. NSUPER: the number of superordinate categories that a triplet belongs to. FastText Correct: accuracy of Fasttext predictions. Odd by Super: accuracy of predictions by the odd superordinate category

NSUPER	FastText Correct	Odd by Super	Vgg16 Early	CAE	Basic	Super	Combined	Basic-Super	Wordvec	SPoSE	# data
1	False	0	0.58	**0.60**	0.50	**0.60**	0.58	0.47	0.59	0.32	222
	True	0	0.33	0.56	0.60	0.59	0.70	0.75	0.71	**0.80**	285
2	False	**0.31**	0.22	0.24	0.24	0.30	0.26	0.28	0.20	0.30	496
	True	**0.99**	0.88	0.84	0.95	0.99	0.98	0.98	0.93	0.98	3612
3	False	0	0.38	0.37	0.43	0.46	0.43	0.44	0.42	**0.57**	2231
	True	0	0.52	0.47	0.71	0.76	0.71	0.73	0.76	**0.89**	2851

5 CONCLUSION

We examined the visual representations learnt by CNNs when supervised by different types of labels and compared them with human similarity judgements. The representations learned by the models are shaped enormously by the kinds of supervision the models get suggesting that much of the categorical structure is not present in the visual input, but requires top-down guidance in the form of category labels. Surprisingly, the kind of supervised input that proved most effective in matching human performance on an triplet odd-one-out task was training with superordinate labels (vehicle, tool, etc.). Such labels allow the networks to perform better now only when the odd-one-out comes from a different superordinate category – this is not surprising – but also when all three images come from different superordinate categories (e.g., when choosing between a banana, a bee, and a screwdriver). Our ongoing work is examining exactly how the different types of labels shape visual representations and how labeling schemes modeled on specific languages (e.g., English vs. Mandarin) may translate to differential human and CNN classificacation performance.

Acknowledgments

Research was supported by an NSF-PAC award. Details will be added post- review.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438, 2015.

Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7603–7612, 2018.

F Gregory Ashby and Shawn W Ell. The neurobiology of human category learning. *Trends in cognitive sciences*, 5(5):204–210, 2001.

F Gregory Ashby and W Todd Maddox. Human category learning. *Annu. Rev. Psychol.*, 56:149–178, 2005.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1043–1052, 2018.

James E Corter and Mark A Gluck. Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, 111(2):291, 1992.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

Paul R Hays. From the jurassic dark: Linguistic relativity as evolutionary necessity. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pp. 159–172, 2000.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *bioRxiv*, pp. 545954, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2013.

Stephen C Levinson. From outer to inner space: linguistic categories and non-linguistic thinking. *Language and conceptualization*, pp. 13–45, 1997.

Gary Lupyan and Molly Lewis. From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Language, Cognition and Neuroscience*, pp. 1–19, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Joshua C Peterson, Paul Soulos, Aida Nematzadeh, and Thomas L Griffiths. Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. *arXiv preprint arXiv:1805.07647*, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Charles Y. Zheng, Francisco Pereira, Chris I. Baker, and Martin N. Hebart. Revealing interpretable object representations from human behavior. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryxSrhC9KX.

A APPENDIX

A.1 LIST OF 30 CATEGORIES

Superordinate-level Category	Basic-level Category	Wordnet ID
Mammal	Orangutan	n02480495
	Gazelle	n02423022
	Lion	n02129165
Insect	Ant	n02219486
	Bee	n02206856
	Grasshopper	n02226429
Bird	Hummingbird	n01833805
	Goose	n01855672
	Vulture	n01616318
Vegetable	Artichoke	n07718747
	Cucumber	n07718472
	Zucchini	n07716358
Fruit	Orange	n07747607
	Lemon	n07749582
	Banna	n07753592
Tool	Hammer	n03481172
	Screwdriver	n04154565
	Shovel	n04208210
Vehicle	Minivan	n03770679
	Trolley	n04335435
	Taxi	n02930766
Musical Instrument	Drum	n03249569
	Flute	n03372029
	Harmonica	n03494278
Kitchen Appliance	Refrigerator	n04070727
	Toaster	n04442312
	Coffee pot	n03063689
Container	Bucket	n02909870
	Mailbox	n03710193
	Crate	n03127925

A.2 CONV AUTOENCODER RESULTS