Shaping Visual Representations with Language for Few-shot Classification

Jesse Mu
Stanford University
muj@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

Noah Goodman
Stanford University
ngoodman@stanford.edu

Abstract

Language is designed to convey useful information about the world, thus serving as a scaffold for efficient human learning. How can we let language guide representation learning in machine learning models? We explore this question in the setting of few-shot visual classification, proposing models which learn to perform visual classification while jointly predicting natural language task descriptions at train time. At test time, with no language available, we find that these language-influenced visual representations are more generalizable, compared to meta-learning baselines and approaches that explicitly use language as a bottleneck for classification.

1 Introduction

Humans are powerful and data-efficient learners partially due to the ability to *learn from language* [6, 30]: for instance, we can learn about *robins* not by seeing thousands of examples, but by being told that *a robin is a bird with a red belly and brown feathers*. This language not only helps us learn about robins, but shapes the way we view the world, constraining the hypotheses we form for other concepts [12]: given a new bird like *seagulls*, even without language we know to attend to salient features including belly and feather color.

In this paper, we propose to use language as a guide for representation learning, building few-shot classification models that learn visual representations while jointly predicting task-specific language during training. Crucially, our models can operate *without language at test time*: a more practical setting, since it is often unrealistic to assume that linguistic supervision is available for unseen classes encountered in the wild. Compared to meta-learning baselines and recent approaches which use language supervision as a more fundamental bottleneck in a model [1], we find this simple auxiliary training objective results in learned representations that generalize better to new concepts.

2 Related Work

Language has been shown to assist visual classification in various settings, including traditional visual classification with no transfer [16] and with language available at test time in the form of class labels or descriptions for zero- [10, 11, 27] or few-shot [24, 33] learning. Unlike this work, we study a setting where we have *no language at test time* and test tasks are unseen, so language from training can no longer be used as additional class information [cf. 16] or weak supervision for labeling additional in-domain data [cf. 15]. Our work can thus be seen as an instance of the *learning using privileged information* (LUPI) [31] framework, where richer supervision augments a model during training only.

In this framework, learning with attributes and other domain-specific rationales has been tackled extensively [8, 9, 29], but language remains relatively unexplored. [13] use METEOR scores between captions as a similarity metric for specializing embeddings for image retrieval, but do not directly

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

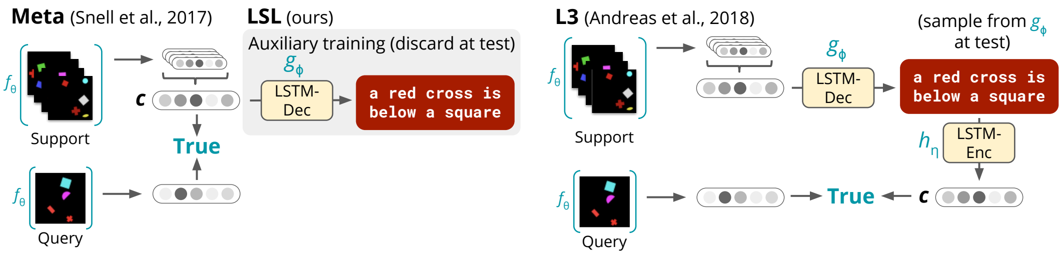

Meta (Snell et al., 2017)
f_θ
Support
f_θ
Query
c
True
LSL (ours)
Auxiliary training (discard at test)
g_ϕ
LSTM-Dec
a red cross is below a square
L3 (Andreas et al., 2018)
f_θ
Support
g_ϕ
LSTM-Dec
(sample from g_ϕ at test)
a red cross is below a square
h_n
LSTM-Enc
f_θ
Query
True
c

Figure 1: Building on prototype networks [26], we propose few-shot classification models whose learned representations are constrained to predict natural language descriptions of the task during training, in contrast to models [1] which explicitly use language as a bottleneck for classification.

ground language explanations. [28] explore a supervision setting similar to ours, except in highly structured text and symbolic domains where descriptions can be easily converted to executable forms via semantic parsing. Another line of work studies models which generate natural language explanations of decisions for interpretability for both textual (e.g. natural language inference; [3]) and visual [17, 18] tasks, but here we examine whether this act of predicting language can actually improve downstream task performance; similar ideas have been explored in text [22] and reinforcement learning [2, 14] domains. Our work is most similar to [1], which we describe and compare to later.

3 Language-shaped learning

We are interested in N-way, K-shot learning, where a task t consists of N *support* classes $\{\mathcal{S}_1^{(t)}, \ldots, \mathcal{S}_N^{(t)}\}$ with K examples each: $\mathcal{S}_n^{(t)} = \{\mathbf{x}_{n,1}^{(t)}, \ldots, \mathbf{x}_{n,K}^{(t)}\}$. Each task has M *query* examples $\mathcal{Q}^{(t)} = \{(\mathbf{x}_1^{(t)}, y_1^{(t)}), \ldots, (\mathbf{x}_M^{(t)}, y_M^{(t)})\}$. The goal is to predict each $y_m^{(t)} \in \{1, \ldots, N\}$ corresponding to the class of the m-th query example.

The approach we propose is applicable to any meta-learning framework that learns an embedding of its input. Here we use prototype networks [26], which have a simple but powerful inductive bias for few-shot learning. Prototype networks learn an embedding function f_θ for exemplars; the embeddings of all examples of a class n are then averaged to form a class "prototype" (omitting task $^{(t)}$ for clarity):

$$\mathbf{c}_n = \frac{1}{K} \sum_{k=1}^{K} f_\theta(\mathbf{x}_{n,k}) \tag{1}$$

Given a query point $(\mathbf{x}_m, y_m)$, we predict class n with probability proportional to some similarity function s between $\mathbf{c}_n$ and $f_\theta(\mathbf{x}_m)$:

$$p_\theta(\hat{y}_m = n \mid \mathbf{x}_m) \propto \exp\left(s\left(\mathbf{c}_n, f_\theta\left(\mathbf{x}_m\right)\right)\right). \tag{2}$$

The *classification loss* for a single task is

$$\mathcal{L}_{\text{CLS}}(\theta) = -\sum_{m=1}^{M} \log p_\theta\left(\hat{y}_m = y_m \mid \mathbf{x}_m\right). \tag{3}$$

Adding language. Now assume that during training we have for each class $\mathcal{S}_n$ a set of J_n associated natural language descriptions $\mathcal{W}_n = \{\mathbf{w}_1, \ldots, \mathbf{w}_{J_n}\}$, where $\mathbf{w}_j$ is a sequence of words $\mathbf{w}_j = (w_{j,1}, \ldots, w_{j,|\mathbf{w}_j|})$. Each $\mathbf{w}_j$ should explain features of $\mathcal{S}_n$ and need not be associated with individual examples.[1] In Figure 1, we have one description $\mathbf{w}_1 = (\texttt{A}, \texttt{red}, \ldots, \texttt{square})$. Our approach is simple: we constrain f_θ to learn prototypes that can also decode the class language descriptions. Let $\tilde{\mathbf{c}}_n$ be the prototype formed by averaging the support *and* query examples of class n. Then define a

[1] If we have language associated with individual examples, we can regularize at the instance-level, essentially learning image captioning along with visual classification. We did not observe major gains with instance-level supervision (vs class-level) in the tasks explored here, in which case class-level language, being much easier to collect, is preferable; however, there likely exist tasks where instance-level supervision is superior.

language model g_ϕ (e.g. a recurrent neural network), which conditioned on $\tilde{\mathbf{c}}_n$ provides a probability distribution over descriptions $g_\phi(\hat{\mathbf{w}}_j \mid \tilde{\mathbf{c}}_n)$ with a corresponding *natural language loss*:

$$\mathcal{L}_{\text{NL}}(\theta, \phi) = -\sum_{n=1}^{N} \sum_{j=1}^{J_n} \log g_\phi(\mathbf{w}_j \mid \tilde{\mathbf{c}}_n), \tag{4}$$

i.e. the total negative log-likelihood of the class descriptions across all classes in the task. Now we jointly minimize both losses:

$$\operatorname*{arg\,min}_{\theta, \phi} \left[\mathcal{L}_{\text{CLS}}(\theta) + \lambda_{\text{NL}} \mathcal{L}_{\text{NL}}(\theta, \phi) \right], \tag{5}$$

where λ_{NL} is a tunable parameter controlling the weight of the natural language loss. At test, we simply discard g_ϕ and use f_θ to classify. With this component, we call our approach *language-shaped learning* (LSL) (Figure 1).

Relation to L3. LSL is similar to another recent model for this setting: *Learning with Latent Language* (L3) [1], which proposes to use language not only as a supervision source, but as a bottleneck for classification (Figure 1). L3 has the same basic architecture of LSL, but the concepts $\mathbf{c}_n$ are the language descriptions themselves, embedded with an additional recurrent neural network (RNN) encoder h_η: $\mathbf{c}_n = h_\eta(\mathbf{w}_n)$. During training, the ground-truth description is used for classification, while g_ϕ is trained to produce the description; at test, L3 *samples* descriptions $\hat{\mathbf{w}}_n$ from g_ϕ, keeping the description most similar to the support according to the similarity function s.

While L3 has been shown to outperform meta-learning baselines, there are two potential sources of this benefit: is it the linguistic bottleneck itself, or the regularization imposed by training f_θ to predict language? Our evaluation aims to disentangle these effects: LSL isolates the regularization component, and thus is simpler than L3 since it (1) does not require the additional embedding module h_η and (2) does not need the test-time language sampling procedure.[2]

4 Experiments

Here we describe our two tasks and models. For full training details and code, see Appendix A.

ShapeWorld. First, we use the ShapeWorld [20] dataset devised by [1], which consists of 9000 training, 1000 validation, and 4000 test tasks (Figure 2).[3] Each task contains a single support set of $K = 4$ images representing a visual concept with an associated (artificial) English language description, generated with a minimal recursion semantics representation of the concept [7]. Each concept is a spatial relation between two objects, optionally qualified by color and/or shape; 2-3 distractor shapes are also present in each image. The task is to predict whether a single query image $\mathbf{x}$ belongs to the concept.

Model details are identical to [1] for easy comparison. f_θ is the final convolutional layer of a fixed ImageNet-pretrained VGG-16 [25] fed through two fully-connected layers:

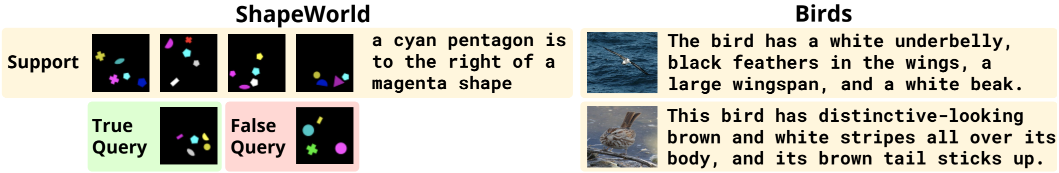
$$f_\theta(\mathbf{x}) = \text{FC}(\text{ReLU}(\text{FC}(\text{VGG-16}(\mathbf{x})))). \tag{6}$$

Since this is a binary classification task with only 1 (positive) support class $\mathcal{S}$ and prototype $\mathbf{c}$, we define the similarity function $s(a, b) = \sigma(a \cdot b)$ and the prediction $P(\hat{y} = 1 \mid \mathbf{x}) = s\left(f_\theta(\mathbf{x}), \mathbf{c}\right)$. g_ϕ is a gated recurrent unit (GRU) RNN [5] with hidden size $h = 512$, trained with teacher forcing. Using a grid search on the validation set, we set $\lambda_{\text{NL}} = 20$.

Birds. To see if LSL can scale to more realistic scenarios, we use the Caltech-UCSD Birds dataset [32], which contains 200 bird species, each with 40–60 images, split into 100 train, 50 validation,

[2] Indeed, LSL is nearly identical to the "Meta+Joint" baseline of [1], who observed no improvement with this method. However, they use separate encoders for the support and query examples, with only the support encoder trained to predict language, resulting in overfitting of the query encoder.

[3] This is a larger version with 4x as many test tasks for more stable confidence intervals (see Appendix A).



a cyan pentagon is to the right of a magenta shape

The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.

This bird has distinctive-looking brown and white stripes all over its body, and its brown tail sticks up.

Figure 2: Example language and query examples for ShapeWorld and Birds (best viewed in color).

Table 1: Model test accuracies (± 95% CI) across 1000 (ShapeWorld) and 600 (Birds) tasks.

	ShapeWorld	Birds
Meta	60.59 ± 1.07	57.97 ± 0.96
L3	66.60 ± 1.18	53.96 ± 1.06
LSL	$\mathbf{67.29} \pm \mathbf{1.03}$	$\mathbf{61.24} \pm \mathbf{0.96}$

and 50 test classes. We use the language descriptions collected by [23], where AMT crowdworkers were asked to describe images of birds in detail, without reference to the species (Figure 2).

While 10 English descriptions per image are available in [23], we assume a more realistic scenario where we have *much less* language available only at the class level: removing one-to-one associations between images and their descriptions, we aggregate a total of D descriptions for each class, and for each k-shot training episode we sample k descriptions from each class n to use as descriptions $\mathcal{W}_n$. In practice, we found good results with as little as $D = 20$ descriptions per class (2000 total) which we report here; for results varying this number, see Appendix B.

We evaluate on the 5-way, 1-shot setting, and as f_θ use the 4-layer convolutional backbone used in much of the few-shot literature [4]. Here we use a learned bilinear similarity function, $s(a, b) = a^\top \mathbf{W} b$, where $\mathbf{W}$ is learned jointly with the model. g_ϕ is a GRU with hidden size $h = 200$, and with another grid search we set $\lambda_{\text{NL}} = 3$.

5 Results

Results are located in Table 1. For ShapeWorld, LSL outperforms its meta-learning baseline (Meta) by 6.7 points, and does as well as L3. For Birds, we observe a smaller but still significant 3.3 point increase over Meta, while L3's performance drops below baseline. We thus conclude that any benefit of L3 is mostly due to the regularizing effect that language has on its image representations, rather than the linguistic bottleneck. Isolating the regularization, we find that LSL is the superior yet conceptually simpler model, and L3's discrete bottleneck can actually hurt in some settings.

To identify which aspects of language are most helpful for the model, we examine LSL performance under ablated language supervision: (1) keeping only a list of common color words, (2) filtering out color words, (3) shuffling the words in each caption, and (4) shuffling the captions across tasks (Figure 3). We find that while the benefits of color/no-color language varies for ShapeWorld and Birds, neither component is completely sufficient for the full benefit of language supervision, demonstrating that LSL is able to leverage both colors and other attributes (e.g. size, shape) exposed through language. Word order is important for Birds but surprisingly unimportant for ShapeWorld. Finally,

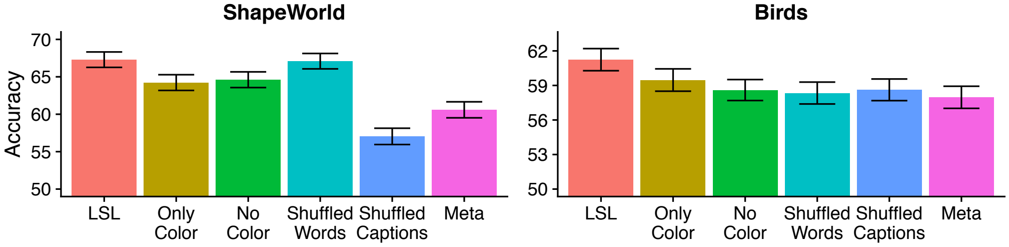


Figure 3: LSL language ablation studies, accuracies reported as in Table 1.

when the captions are shuffled and thus the linguistic signal is random, LSL for Birds suffers no performance loss compared to Meta, while LSL for ShapeWorld drops significantly, likely because the language descriptions are more central to the task.

6 Discussion

We presented a method for regularizing a few-shot visual recognition model by forcing the model to predict natural language descriptions during training. Across two tasks, the language-influenced representations learned with such models improved generalization over those without linguistic supervision. By comparing to L3, we found that if a model has been trained to learn representations which expose the features and abstractions in language, a linguistic bottleneck on top of this already language-shaped representation is unnecessary, at least for the kinds of visual tasks explored here.

The line between language and sufficiently rich attributes and rationales is blurry, and as recent work has shown [29], the performance gains in this work can likely be observed by regularizing with attributes. However, unlike attributes and annotator rationales, language is (1) a more natural medium for annotators, (2) does not require preconceived restrictions on the kinds of features relevant to the task, and (3) is abundant in unsupervised forms. This last point suggests we can shape representations with language from external resources (e.g. the Web), a promising future direction of work.

Acknowledgments

We thank Pang Wei Koh for helpful discussions, and Mike Wu and Jacob Andreas for discussions and code. This work was supported by an NSF Graduate Research Fellowship for JM, a SAIL-Toyota Research Award, and the Office of Naval Research grant ONR MURI N00014-16-1-2007. Toyota Research Institute (TRI) provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

References

[1] J. Andreas, D. Klein, and S. Levine. Learning with latent language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2166–2179, 2018.

[2] D. Bahdanau, F. Hill, J. Leike, E. Hughes, A. Hosseini, P. Kohli, and E. Grefenstette. Learning to understand goal specifications by modelling reward. In *International Conference on Learning Representations (ICLR)*, 2019.

[3] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. e-snli: natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9539–9549, 2018.

[4] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019.

[5] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[6] S. Chopra, M. H. Tessler, and N. D. Goodman. The first crank of the cultural ratchet: Learning and transmitting concepts through language. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 226–232, 2019.

[7] A. A. Copestake, G. Emerson, M. W. Goodman, M. Horvat, A. Kuhnle, and E. Muszynska. Resources for building applications with dependency minimal recursion semantics. In *International Conference on Language Resources and Evaluation (LREC)*, 2016.

[8] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2013.

[9] J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1395–1402, 2011.

[10] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2584–2591, 2013.

[11] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, R. Marc'Aurelio, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2121–2129, 2013.

[12] N. Goodman. *Fact, fiction, and forecast*. Harvard University Press, Cambridge, MA, 1955.

[13] A. Gordo and D. Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6589–6598, 2017.

[14] P. Goyal, S. Niekum, and R. J. Mooney. Using natural language for reward shaping in reinforcement learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2385–2391, 7 2019.

[15] B. Hancock, P. Varma, S. Wang, M. Bringmann, P. Liang, and C. Ré. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, 2018.

[16] X. He and Y. Peng. Fine-grained image classification via combining vision and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5994–6002, 2017.

[17] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2016.

[18] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279, 2018.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[20] A. Kuhnle and A. Copestake. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*, 2017.

[21] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[22] N. F. Rajani, B. McCann, C. Xiong, and R. Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4932–4942, Florence, Italy, July 2019.

[23] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–58, 2016.

[24] E. Schwartz, L. Karlinsky, R. Feris, R. Giryes, and A. M. Bronstein. Baby steps towards few-shot learning with multiple semantics. *arXiv preprint arXiv:1906.01905*, 2019.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4077–4087, 2017.

[27] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 935–943, 2013.

[28] S. Srivastava, I. Labutov, and T. Mitchell. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1527–1536, 2017.

[29] P. Tokmakov, Y.-X. Wang, and M. Hebert. Learning compositional representations for few-shot recognition. *arXiv preprint arXiv:1812.09213*, 2018.

[30] M. Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA, 1999.

[31] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011.

[33] C. Xing, N. Rostamzadeh, B. N. Oreshkin, and P. O. Pinheiro. Adaptive cross-modal few-shot learning. *arXiv preprint arXiv:1902.07104*, 2019.

A Task/model training details

Our code is publicly available at `https://github.com/jayelm/lsl`.

A.1 ShapeWorld

f_θ. Like [1], f_θ starts with features extracted from the last convolutional layer of a fixed ImageNet-pretrained VGG-19 network [25]. These 4608-d embeddings are then fed into two fully connected layers $\in \mathbb{R}^{4608 \times 512}, \mathbb{R}^{512 \times 512}$ with one ReLU nonlinearity in between.

LSL. For LSL, the 512-d embedding from f_θ directly initializes the 512-d hidden state of the GRU g_ϕ. We use 300-d word embeddings initialized randomly (initializing with GloVe made no significant difference).

L3. f_θ and g_ϕ are the same as in LSL and Meta. h_η is a unidirectional 1-layer GRU with hidden size 512 sharing the same word embeddings as g_ϕ. The output of the last hidden state is taken as the embedding of the description $\mathbf{w}^{(t)}$. Like [1], a total of 10 descriptions per task are sampled at test time.

Training. We train for 50 epochs, each epoch consisting of 100 batches with 100 tasks in each batch, with the Adam optimizer [19] and a learning rate of 0.001. We selected the model with highest epoch validation accuracy during training. This differs slightly from [1], who use different numbers of epochs per model and did not specify how they were chosen; otherwise, the training and evaluation process is the same.

Data. We recreated the ShapeWorld dataset using the same code as [1], except generating 4x as many test tasks (4000 vs 1000) for more stable confidence intervals.

Note that results for both L3 *and* Baseline (Meta) are 3–4 points lower than the scores of the corresponding implementations in [1]. This is likely due to (1) differences in model initialization due to our PyTorch reimplementation, (2) recreation of the dataset, and (3) our use of early stopping.

A.2 Birds

f_θ. The 4-layer convolutional backbone f_θ is the same as the one used in much of the few-shot literature [4, 26]. The model has 4 convolutional blocks, each consisting of a 64-filter 3x3 convolution, batch normalization, ReLU nonlinearity, and 2x2 max-pooling layer. With an input image size of 84×84 this results in 1600-d image embeddings. Finally, the similarity metric matrix $\mathbf{W}$ has dimension 1600×1600.

LSL. The resulting 1600-d image embeddings are fed into a single linear layer $\in \mathbb{R}^{1600 \times 200}$ which initializes the 200-d hidden state of the GRU. We initialize embeddings with GloVe [21]. We did not observe significant gains from increasing the size of the decoder g_ϕ.

L3. f_θ and g_ϕ are the same. h_η is a unidirectional GRU with hidden size 200 sharing the same embeddings as g_ϕ. The last hidden state is taken as the concept $\mathbf{c}_n$. 10 descriptions per class are sampled at test time. We did not observe significant gains from increasing the size of the decoder g_ϕ or encoder h_η, nor increasing the number of descriptions sampled per class at test.

Training. For ease of comparison to the few-shot literature we use the same training and evaluation process as [4]. Models were trained for 60000 episodes, each episode consisting of one randomly sampled task with 16 query images per class. Models were trained end-to-end with the Adam optimizer [19] and a learning rate of 0.001. We select the model with the highest validation accuracy after training.

Data. Like [4], we use standard data preprocessing and training augmentation: ImageNet mean pixel normalization, random cropping, horizontal flipping, and color jittering.

B Amount of language supervision

See Figure 4. With enough language supervision (one caption for each image), L3's performance approaches baseline. Meanwhile, LSL shows limited gains as the amount of language supervision increases past 10 captions per class.

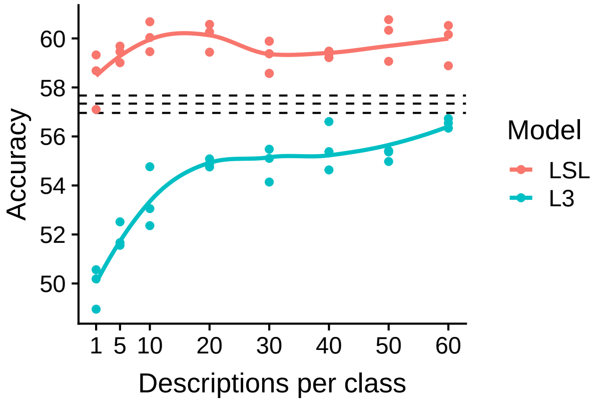


Figure 4: Results varying the amount of language descriptions available per class. At 60, 1 caption per image is available. Each dot is a separate independently trained model, and the dashed lines represent independently trained baselines (Meta).