Learning Compositional Representations for Few-Shot Recognition Supplementary Material

Pavel Tokmakov Yu-Xiong Wang Martial Hebert
Robotics Institute, Carnegie Mellon University
{ptokmako,yuxiongw,hebert}@cs.cmu.edu

This supplementary material provides additional experimental results and details. We begin by evaluating the effect of the number of attributes used for training on the model's few-shot performance in Section 1. Next, we explore the effect of plugging our compositional representations into existing few-shot learning methods, such as Prototypical Networks [4] and Matching Networks [5], as well as the effect of data augmentation on these methods in Section 2. We then provide a qualitative analysis of the learned representation using Network Dissection [7] in Section 3. In Section 4 we demonstrate how to apply our proposed regularization to deeper network architectures. Finally, we visualize the attributes used in our experiments on ImageNet together with their hierarchical structure in Sections 5, and provide additional implementation details in Section 6.

1. Effect of the Number of Attributes

One of the main concerns with our proposed approach is that obtaining attribute supervision can be expensive in practice. To partially mitigate it, we study the effect of the number of attributes used in training on the model's few-shot learning performance. To this end, we sample 75/50/25/15/5% of the attributes on CUB and SUN at random and train our models using these subsets with the soft constraint and orthogonality regularization. The results on the novel categories are presented in Figures 1 and 2, respectively. Encouragingly, the performance decreases only slightly with the number of attributes. In particular, with only 25% of the original attributes on CUB and 50% on SUN (which correspond to 34 and 45 attributes, respectively), the performance hardly changes. Moreover, we achieve noticeable improvements over the 0-attribute baseline by using as few as 5 attributes on SUN. This result strengthens our claim that the proposed approach can improve the performance of few-shot learning methods with only a small annotation overhead.

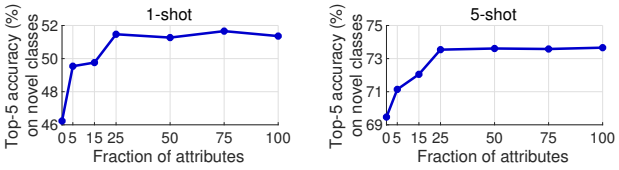

Figure 1. Evaluation of the effect of the number of attributes on the model's performance on the novel categories of the CUB dataset. Our approach provides significant improvements over the baseline even with as few as a quarter of the attributes.

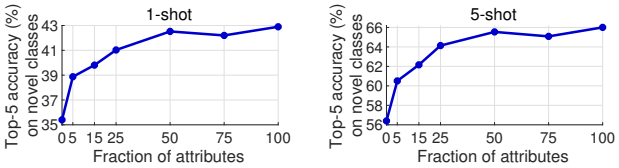

Figure 2. Evaluation of the effect of the number of attributes on the model's performance on the novel categories of the SUN dataset. In this case, half of the original number of attributes is enough to achieve comparable improvements over the baseline.

2. Additional Analysis of Existing Few-Shot Learning Methods

In the main paper, we have demonstrated that a simple cosine classifier learned on top of a frozen CNN, which was trained with our compositionality regularization, leads to state-of-the-art results on three datasets, outperforming more complex existing few-shot classification models, such as Protoypical Networks [4] and Matching Network [5]. It is natural to ask whether training these models with our compositional representation would lead to superior results. To answer this question, we train the CNN backbone on the base categories with a linear classifier and the compositionality regularization. On top of the compositional feature, we learn these few-shot models as described in the main paper. We report the results on the novel categories of the CUB-200-2011 dataset in Table 1.

We observe that using compositional representations indeed leads to improved performance for both Prototypical Networks and Matching Networks in almost all the settings. The improvements for Prototypical Networks are marginal.

	Novel		
	1-shot	2-shot	5-shot
PN	43.2	54.3	67.8
PN + data aug	44.0	54.8	68.1
PN w/ comp	41.5	55.0	68.4
PN w/ comp + data aug	42.2	55.7	68.7
MN	48.5	57.3	69.2
MN + data aug	49.3	57.9	69.7
MN w/ comp	50.9	59.5	71.6
MN w/ comp + data aug	51.6	59.9	72.0
Linear w/ comp	47.0	60.0	74.0
Cos w/ comp	52.5	63.6	73.8
Cos w/ comp + data aug	**53.6**	**64.8**	**74.6**

Table 1. Incorporating our compositional representations into existing few-shot classification models : top-5 accuracy on the novel categories of the CUB dataset. 'PN': Prototypical Networks, 'PN w/ comp': Prototypical Networks with our compositional representation, 'MN': Matching Networks, 'MN w/ comp': Matching Networks with our compositional representation, 'Linear w/ comp': our compositional representation with a linear classifier, 'Cos w/ comp': our compositional representation with a cosine classifier. The variants trained with data augmentation are marked with '+ data aug'.

The effect of compositional representations for Matching Networks is more pronounced, allowing them to outperform the linear classifier in the 1-shot evaluation setting. However, our cosine classifier remains superior to the few-shot learning methods. In addition, compared with our approach, data augmentation has a less effect on the performance improvement of traditional few-shot learning methods. Our approach again consistently outperforms the baselines with data augmentation. These experiments not only confirm the surprising effectiveness of the cosine classifier observed in the main paper, but also show that the proposed compositional representations can generalize to other scenarios and classification models.

3. Qualitative Analysis of Compositional Representations

We now qualitatively and quantitatively analyze the learned representations using Network Dissection: a framework for studying the interpretability of CNNs proposed by Zhou *et al*. [7]. They first collect a large dataset of images with pixel-level annotations, where the set of labels spans a diverse vocabulary of concepts from low-level (*e.g*., textures) to high-level (*e.g*., object and scene categories) concepts. They then probe each unit in a pre-trained CNN by treating it as a classifier for each of these concepts. If a unit achieves a score higher than a threshold for one of the concepts, it is assumed to capture the concept. The number of internal units that capture some interpretable concepts is then used as a measure of the interpretability of the network.

We compute this measure for the last layer of our networks (before the classification layer) for both the cosine classifier baseline and the cosine classifier with our compositionality regularization trained on SUN397. We observe that the baseline has 169 interpretable units out of 512, capturing 92 unique concepts. For our proposed compositional model, the number of interpretable units increases to 333 and the number of unique concepts increases to 119. Clearly, the proposed regularization results in learning a much more interpretable representation. To further analyze its properties, we present the distribution of the interpretable units for the baseline in Figure 3 and that for the proposed model in Figure 4, grouped by the concept type. We observe that our improvement in novel concepts mainly comes from the scene and object categories.

Another interesting observation is that most of the new interpretable units seem to be duplicates of the units that already existed in the baseline model. This is due to a limitation of the Network Dissection approach. Although the vocabulary of concepts which this evaluation can identify is relatively broad, it is still limited. Several different real-world concepts thus end up being mapped to a single label in the vocabulary. To illustrate this observation and further analyze our approach, we visualize the maximally activating images for several units that are mapped by Network Dissection to the category `house` in Figure 5. The figure also shows attention maps of the units within each image. The first two units, which are shared by the baseline and the proposed model, seem to capture the general concepts of a `wooden house` and a `stone house`. However, the other three units, which are only found in the model trained with the compositionality regularization, seem to capture parts of the house, such as `roof`, `window`, and `porch` (see attention maps). This observation further validates that our proposed approach leads to learning representations that capture the compositional structure of the concepts.

4. Effect of the Network Depth

Up till now we used a ResNet-10 backbone for all the experiments. In this section, we study the generalizability of the proposed compositionallity regularization to deeper network architectures. We conduct these experiments on the SUN397 dataset due to its large size and high quality of the attribute annotations. In Table 2 we compare the ResNet-10 model with cosine classifier to ResNet-18 and ResNet-34. First of all, we notice that the improvements with respect to the baseline due to compositionallity regularization are diminishing as the network depth increases. Moreover, the shallow 'ResNet-10, Cos w/ comp' model outperforms the deeper variants. We analyze this behavior and observe that the deeper models are able to learn attribute classifiers without significantly modifying their representation. This can be explained by the fact that the feature space of the last layer of the deep networks has a higher representation capability. We thus propose to adapt our regularization by applying it not only to the last, but also to the intermediate layers of the

	Novel			All		
	1-shot	2-shot	5-shot	1-shot	2-shot	5-shot
ResNet-10, Cos	35.4	45.6	56.4	52.1	56.7	61.9
ResNet-10, Cos w/ comp	43.4	54.5	**65.9**	54.9	60.4	66.3
ResNet-18, Cos	37.7	47.5	58.8	53.7	58.0	63.3
ResNet-18, Cos w/ comp	41.2	51.6	63.0	55.5	60.6	66.3
ResNet-18, Cos w/ deep comp	**43.9**	**54.7**	65.7	56.5	62.0	67.6
ResNet-34, Cos	38.5	48.8	60.2	54.3	58.8	64.4
ResNet-34, Cos w/ comp	41.0	51.5	62.5	56.1	60.9	66.3
ResNet-34, Cos w/ deep comp	43.1	53.7	65.7	**57.0**	**62.1**	**68.2**

Table 2. Evaluation of deeper architectures: top-5 accuracy on the novel and all (*i.e.*, $novel + base$) categories of the SUN dataset. 'Cos': the baseline with a cosine classifier, 'Cos w/ comp': our proposed compositional representation with a cosine classifier, 'Cos w/ deep comp': our proposed compositional representation with regularization applied to intermediate layers of the network.

network. In practice, we apply it to the outputs of all the ResNet blocks starting from the block 9. This new variant, denoted as 'Cos w/ deep comp", achieves improvements over the baseline for ResNet-18 and ResNet-34, which is comparable to those of 'Cos w/ comp' for ResNet-10. Such results confirm that our proposed approach is indeed applicable to deeper networks. The improvement is somewhat smaller for the novel classes though; all the three models perform approximately the same in this setting.

5. ImageNet Attributes

In Figure 6, we visualize the hierarchical structure of the attributes which we defined for the 389 base categories in the subset of ImageNet used in our experiments. Each node (including non-leaf nodes) represents a binary attribute and edges capture the parent-child relationships between the attributes. These relationships are used in the annotation process to prune irrelevant attributes (such as number of wheels for a living thing) and thus save the annotator's time. Note that our annotated attributes might not be the perfect set of attributes for ImageNet. Nevertheless, even with these imperfect attributes, our compositionality regularization approach allowed us to achieve the state-of-the-art result.

6. Additional Implementation Details

Training Schedules. On ImageNet we use the setting proposed in [3, 6], with a batch size of 256 and 90 training epochs. The learning rate is decreased by a factor of 10 every 30 epochs. On SUN397 we use the same batch size and total number of epochs, but decrease the learning rate after the first 60 epochs, and then again after 15 more epochs. On CUB-200-2011, which is a much smaller dataset, we use a batch size of 16 and train for 170 epochs. The learning rate is first decreased by a factor of 10 after 130 epochs, and then again after 20 more epochs. These schedules are selected on the validation set.

Compositionality Regularization. On ImageNet the trade-off hyper-parameter of the compositionality regularizer λ is set to 8, on SUN397 to 25, and on CUB-200-2011 to 15. The hyper-parameter of the orthogonality constraint β is set to 0.001, 0.0025, and 0.00035 on the three datasets, respectively. To select these values, we split the base categories in half, using one half as validation. The attribute annotations are sparse, with around 10% of them being labeled as positive for any given image on average. Due to this highly imbalanced distribution of training labels, all the attribute classifiers learn to predict the negative labels. To address it, we randomly sample a subset of the negative attributes for every example in every batch to balance the number of positive and negative examples.

Few-Shot Training. On ImageNet we train for 100 iterations for both cosine and linear classifiers. On SUN397 we train for 200 iterations for linear and 100 for cosine classifier. On CUB-200-2011 we train for 100 iterations for linear and 40 for cosine classifier. To select these values, we use the same split of the base categories discussed above, and train until the top-5 performance on the validation categories stops increasing. We use the same setting to select the optimal hyper-parameters for other methods.

References

[1] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[2] S. Gidaris and N. Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018.

[3] B. Hariharan and R. B. Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, 2017. 3

[4] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017. 1

[5] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *NIPS*, 2016. 1

[6] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018. 3

[7] B. Zhou, D. Bau, A. Oliva, and A. Torralba. Interpreting deep visual representations via network dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1, 2, 4

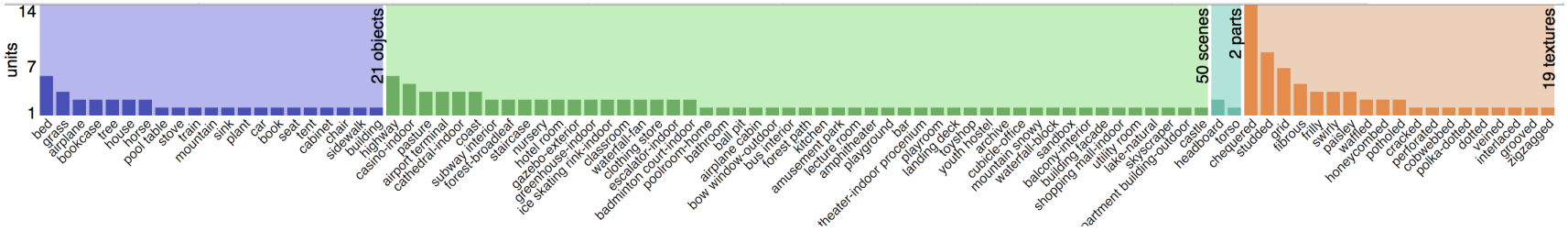

Figure 3. Distributions of interpretable units in the last layer of the baseline model trained with a cosine classifier on SUN397 according to Network Dissection [7]. The units are grouped by the type of the concepts they represent (*i.e.*, object, scene, part, or texture). Overall, this layer has **169** interpretable units, capturing **92** unique concepts.

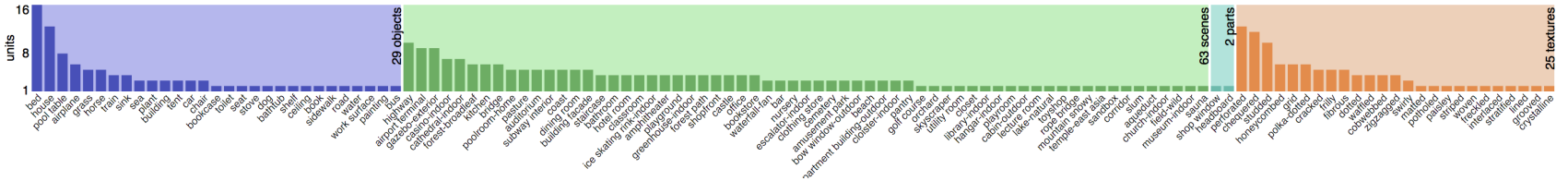

Figure 4. Distributions of interpretable units in the last layer of the model trained with a cosine classifier and our compositionality regularization on SUN397 according to Network Dissection [7]. The units are grouped by the type of the concepts they represent (*i.e.*, object, scene, part, or texture). Overall, this layer has **333** interpretable units, capturing **119** unique concepts.

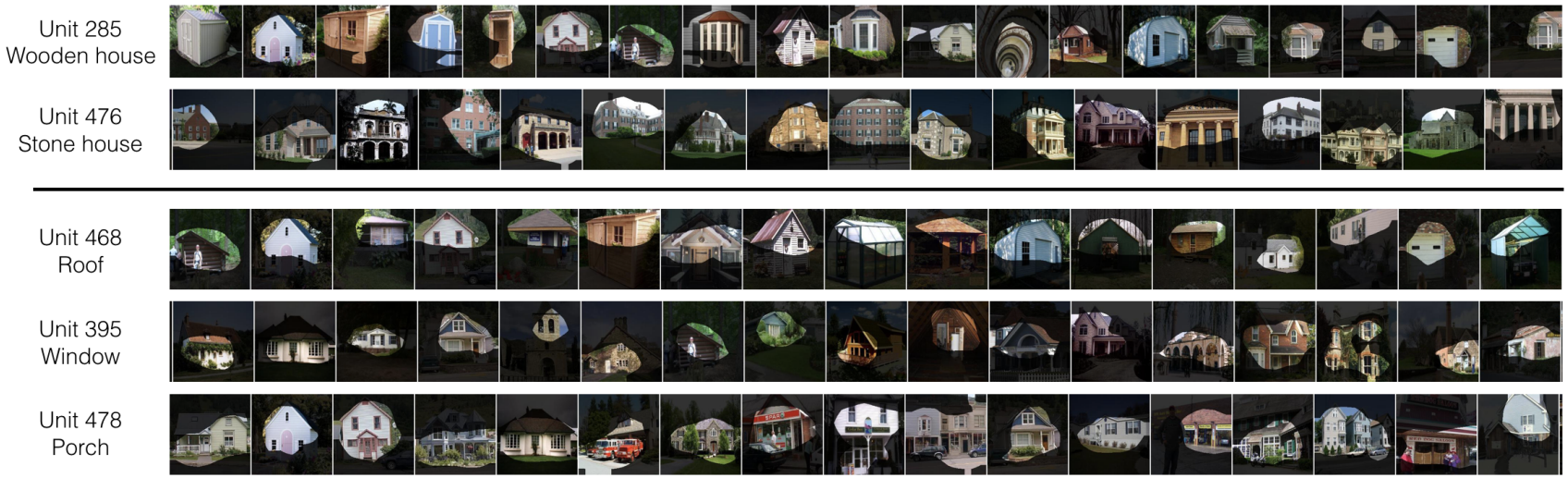

Figure 5. Top activating images for several units in the last layer of the network that are mapped to the concept `house` by Network Dissection, together with the units' attention maps. The first two units are found both in the baseline model and in the model trained with our compositionality regularization, and capture generic concepts: `wooden house` and `stone house`. The next three units are only found in the proposed model and capture parts of the house, such as `roof`, `window`, and `porch` (see attention maps).

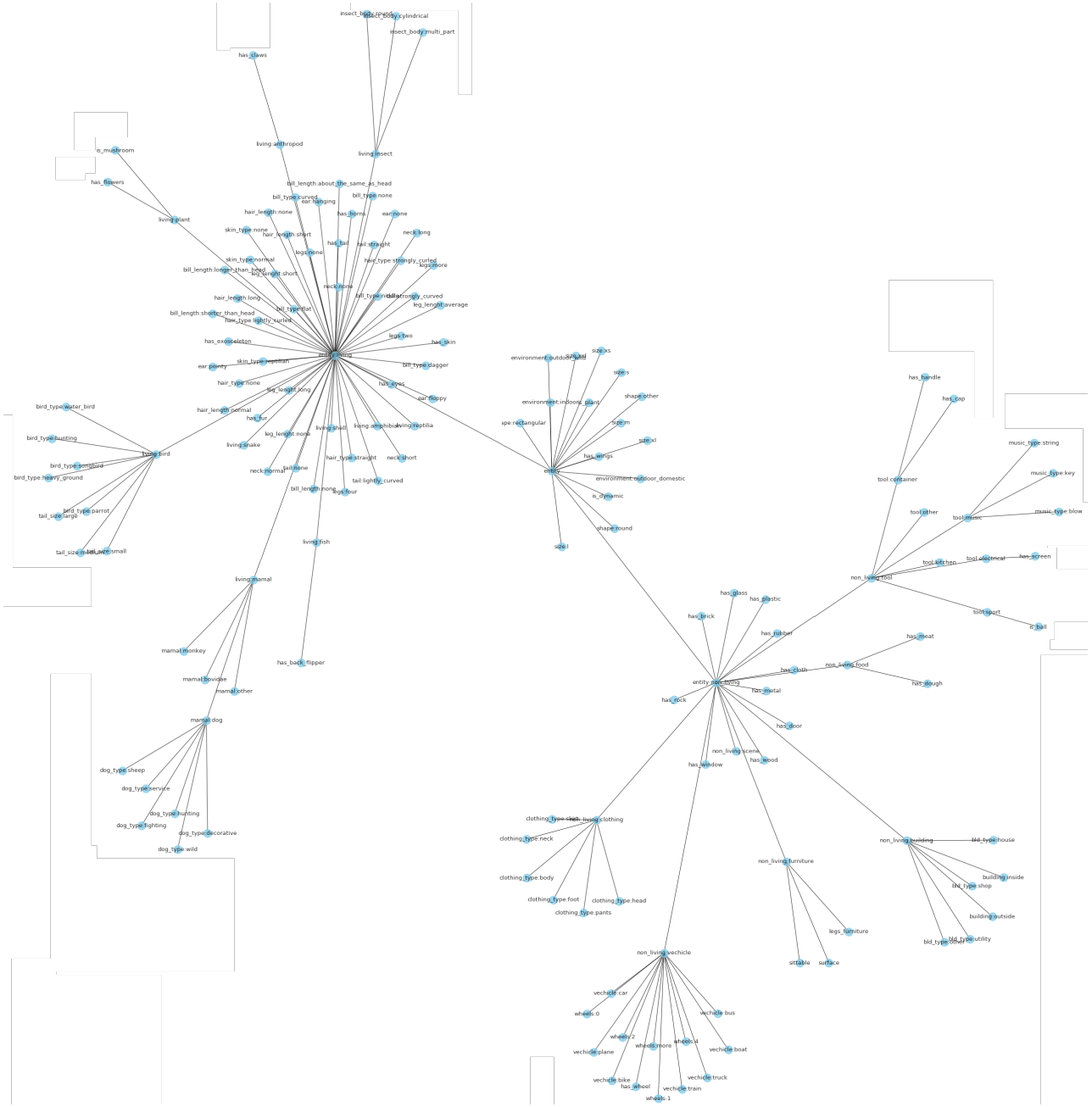

Figure 6. Attributes used in our ImageNet experiments together with their hierarchical structure. Each node represents a binary attribute and edges capture the parent-child relationships between the attributes.