Spontaneous Decomposition from Grouped Network Pathways

Satchel Grant [1] Matheus Dias [1] Sahil Kulkarni[1]

[1]Stanford University

Introduction

- Self-Supervised Learning (SSL) is a machine learning method for learning from unlabeled data. This is desirable due to the high cost to create labelled datasets and the computational costs of training networks from scratch.
- Recent SSL methods such as BYOL and DINO show that it is possible to create robust image representations by using only positive samples.
- We propose a SSL class of models that is reminiscent of an ensemble and show that it can lead to learning robust and interpretable representations, while reducing the number of parameters needed for effective performance.

Methods

1. Sample an image $x \sim \mathcal{X}$
2. Apply transformation $t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$ resulting in two images $v := t(x)$ and $v' := t'(x)$
3. For each student leaf $f_j \in \{f_1, ..., f_m\}$ input v to generate its corresponding output representation $z_j := f_j(v)$
4. Apply the aggregation function over the students' outputs to generate an aggregate student output representation $Z := A(z_1, ..., z_m)$
5. Follow the same procedure for the teacher leaves, resulting in a final teacher output $Z' := A(z'_1, ..., z'_m)$
6. Backpropagate only on the student encoder using the loss function $\mathcal{L}$
7. For each teacher leaf $g_j \in \{g_1, ..., g_m\}$ update its weights as an exponential moving average of the corresponding the student leaf's weights, i.e. $g_{jW} \leftarrow \tau g_{jW} + (1 - \tau) f_{jW}$

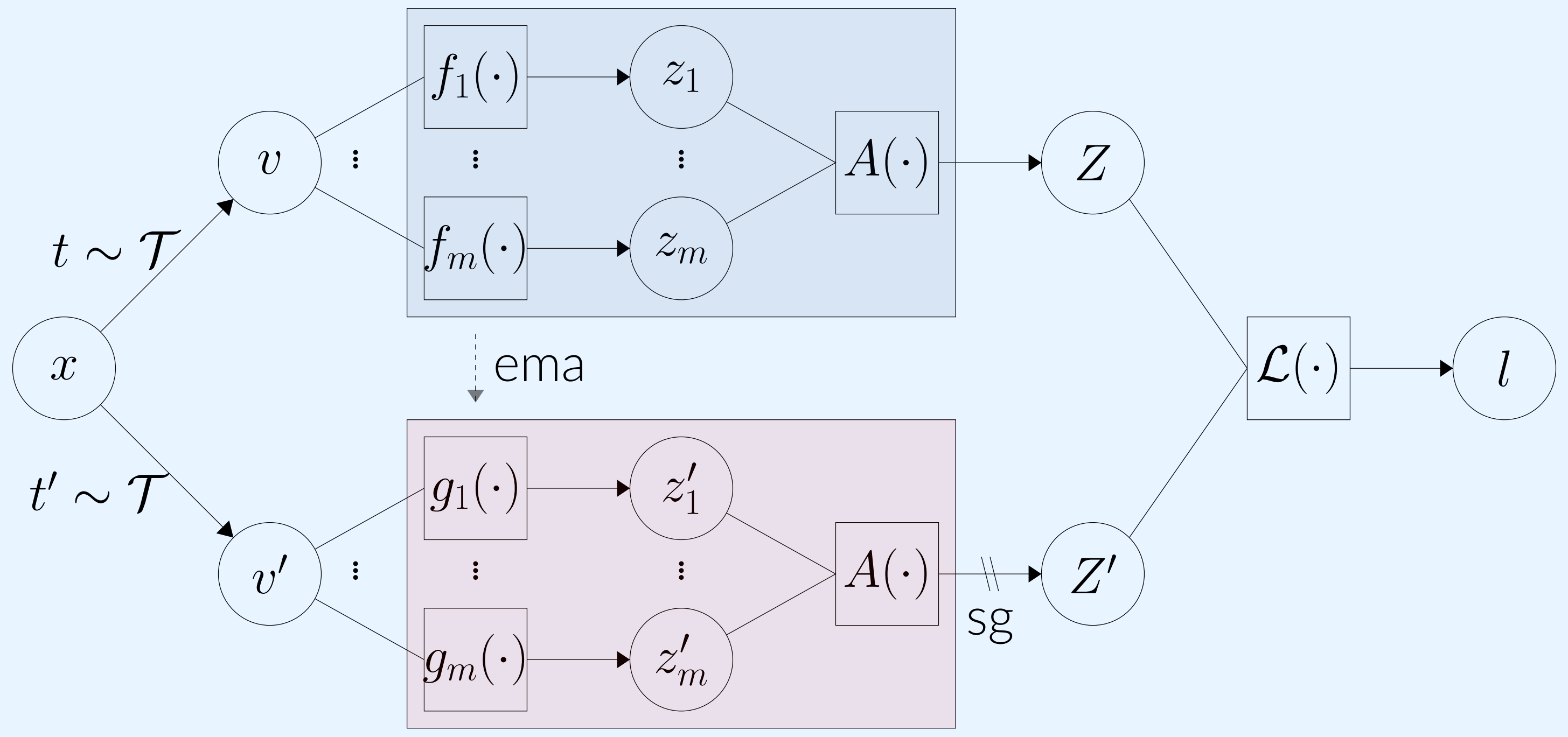


Figure 1. TreeNet Architecture. Teacher leaves, g_j's, are updated using an exponential moving average (ema) and sg means stop-gradient. The student and teacher encoders are shown in blue and red, respectively.

- At test time we pick the teacher encoder for downstream tasks. Here, it receives an unaugmented image as input and outputs a single encoded representation.

Experiments & Analysis

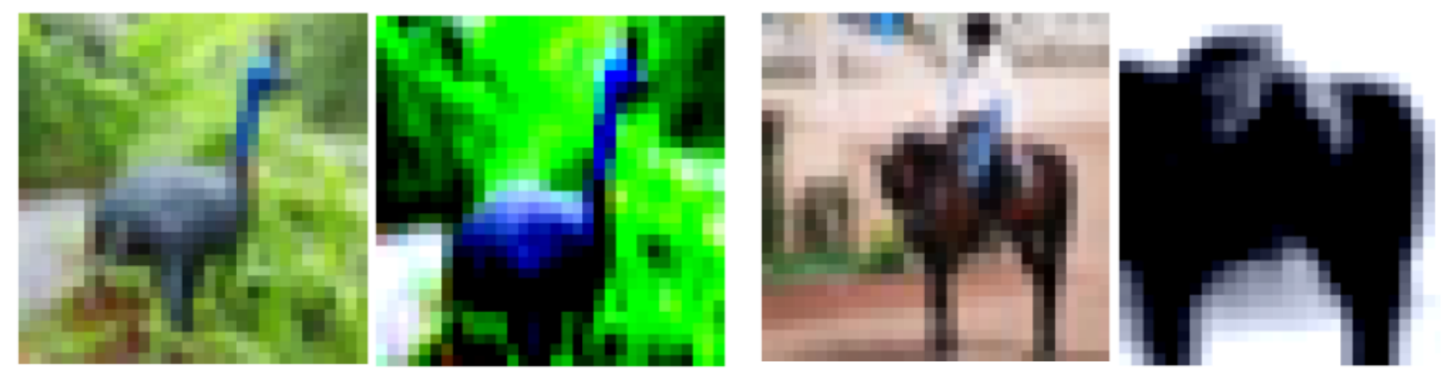
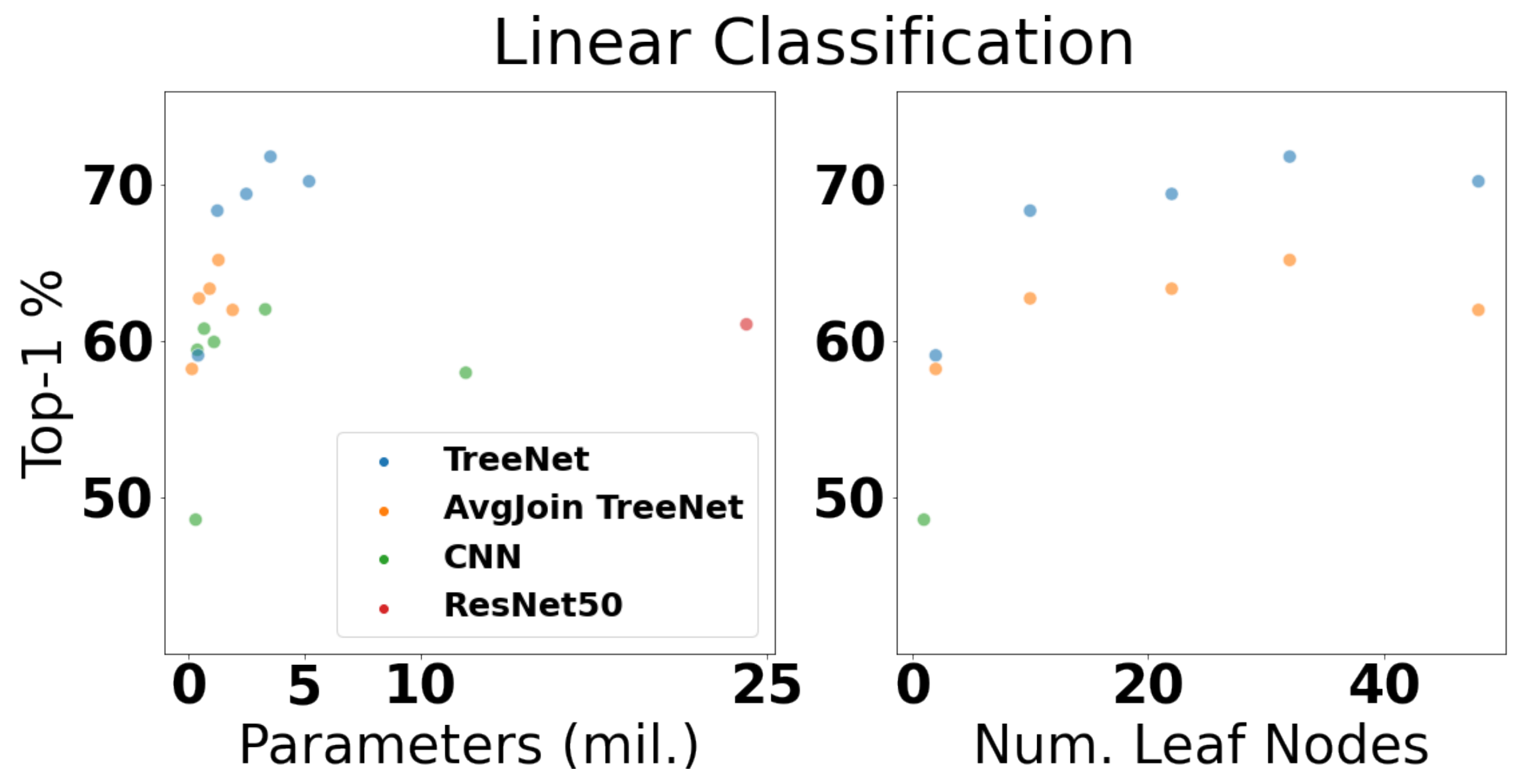


Figure 2. We use the CIFAR-10 dataset as our baseline upon which the transformations are applied. We apply flip and color jitter, normalization, and global and local cropping as transformations to create our augmentations.



To evaluate the quality of our learned representations, we train a linear mapping from the encoder outputs to their labels. The figure above shows TreeNet models perform best overall while significantly reducing the number of parameters.

Type	Channels	N Leaf	Accuracy
DenseJoin	8 - 16 - 32 - 48 - 64	32	**71.8**
AvgJoin	8 - 16 - 32 - 48 - 64	32	**65.2**
CNN	32-128-256-512-1028	1	62.0
ResNet50			61.1

Table 1. Linear evaluation accuracies of different model types and their convolutional channel depths.

Next, we use saliency maps to corroborate our hypothesis that grouped network pathways lead to learning compositional image representations:

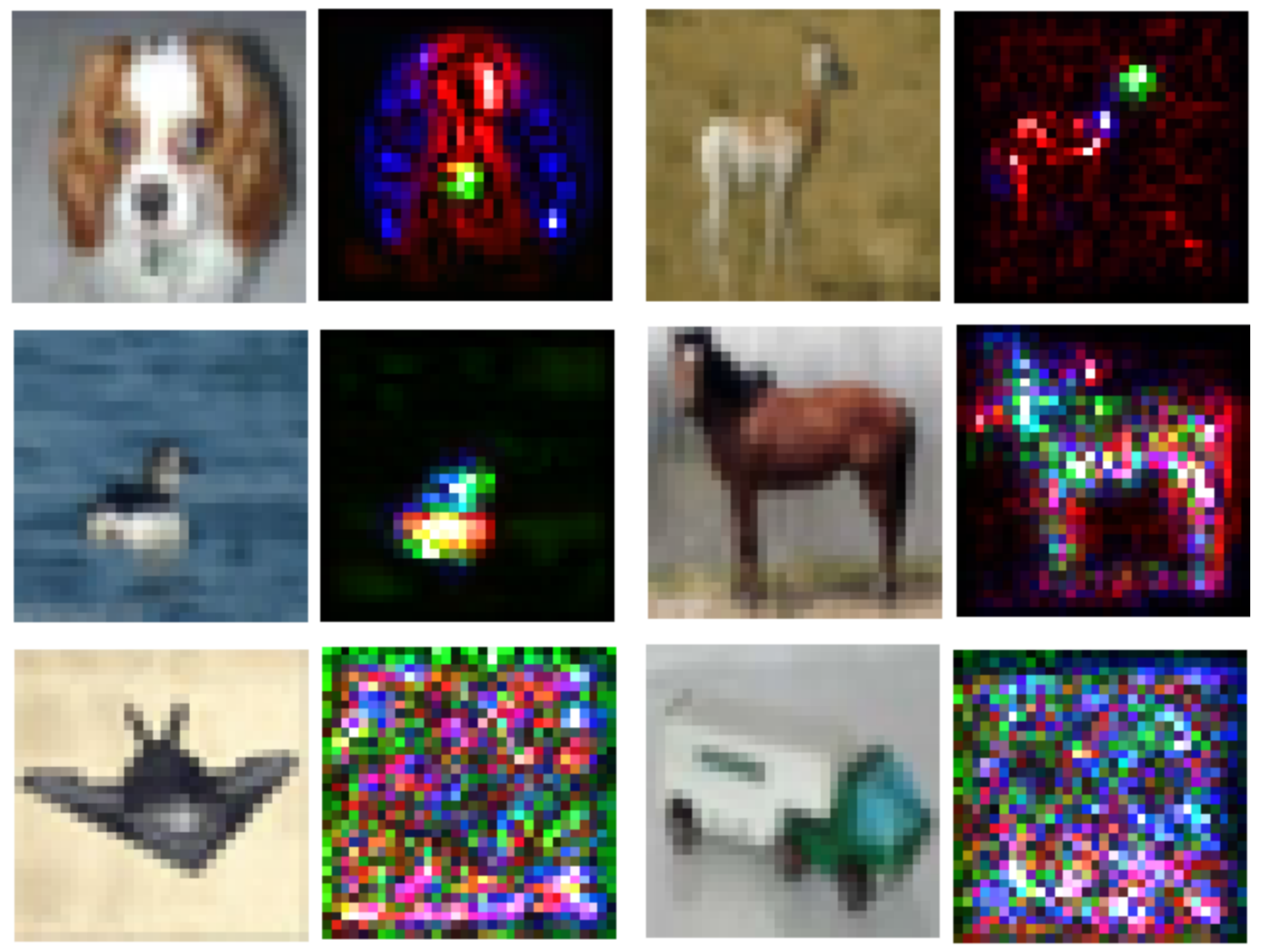


Figure 3. Overlayed saliency maps, each containing three unique leaf nodes - loosely categorized as specialists, generalists, and failed learners.

Analysis Cont.

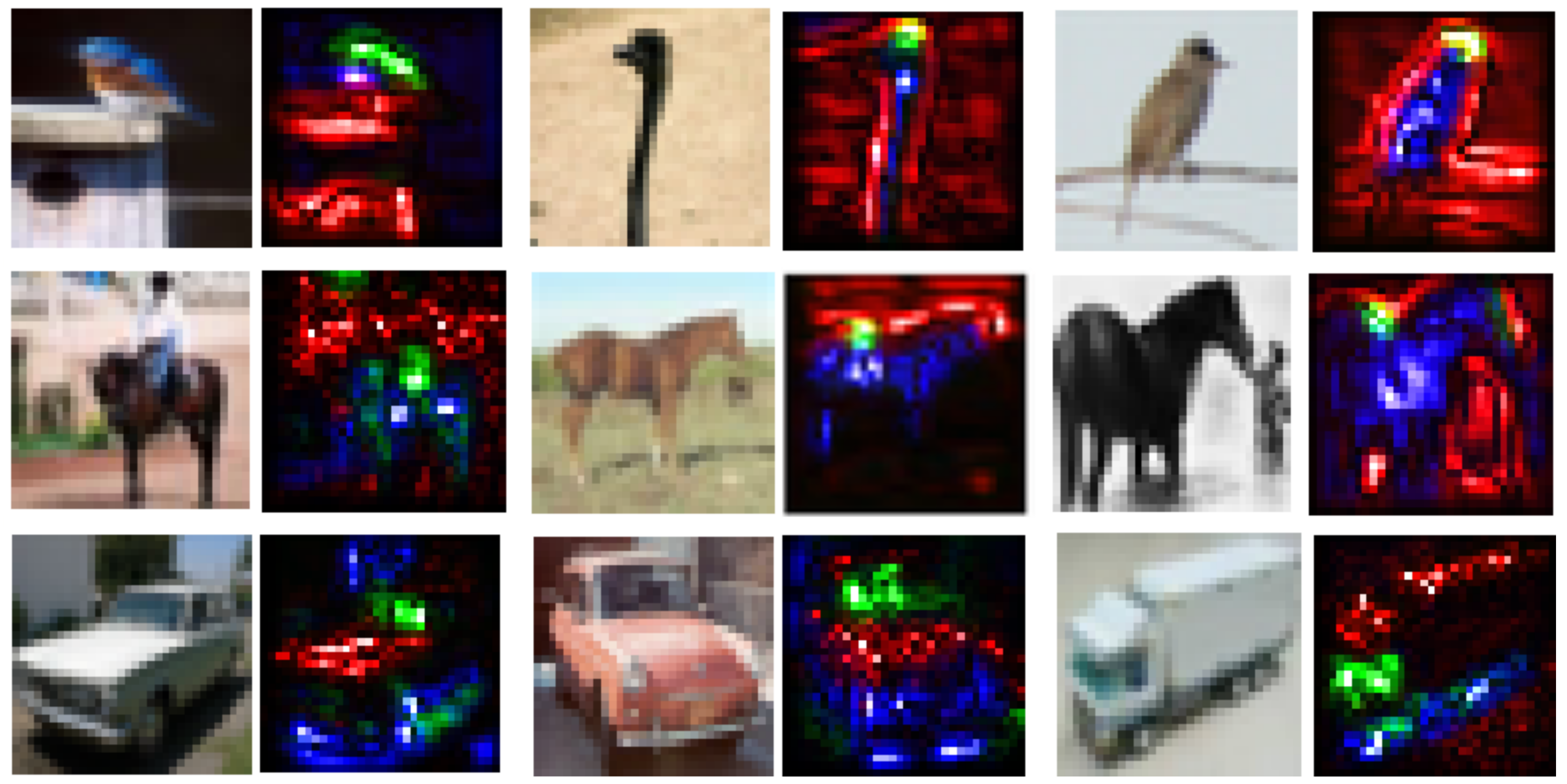


Figure 4. Saliency maps of consistent specialist nodes.

- Leaf nodes appear to fall into three broad categories: "specialists", "generalists", and "failed learners", seen in Figure 3.
- Specialists often fall into two subcategories: "inconsistent" and "consistent".

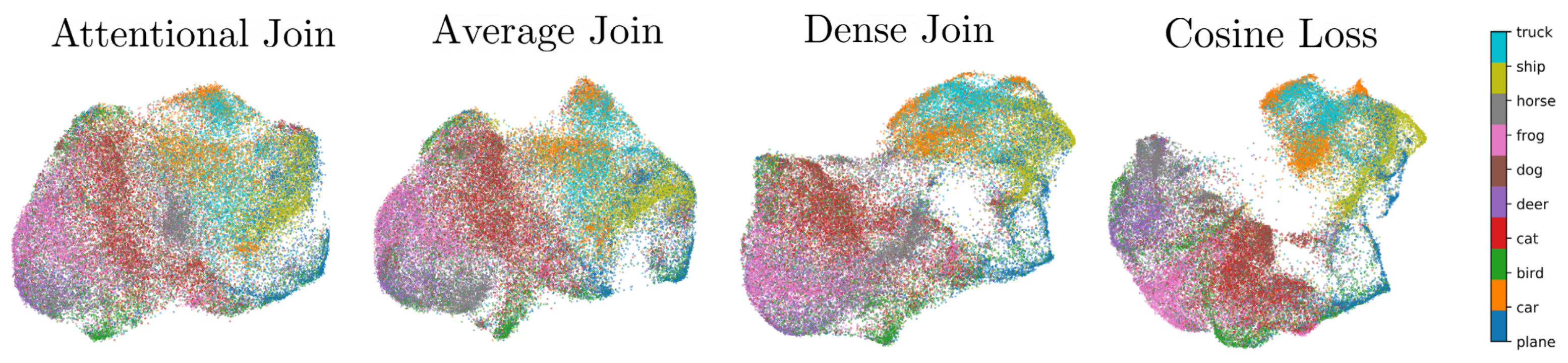


Figure 5. UMAP visualizations of image representations by architecture.

- UMAP visualization of learned representations indicates that using a dense join aggregation leads better separated features.

Conclusions

- Results indicate superior Top-1 Accuracy for CIFAR-10 when compared to reasonable baselines.
- We show that compositional image representations arise simply from imposing an inductive bias of disconnected pathways within the model.
- We believe there is space for further architectural improvement: our model lacks translational granularity over the input data, and our TreeNet architecture could have a more smooth learning process by using ResNet leaf nodes.