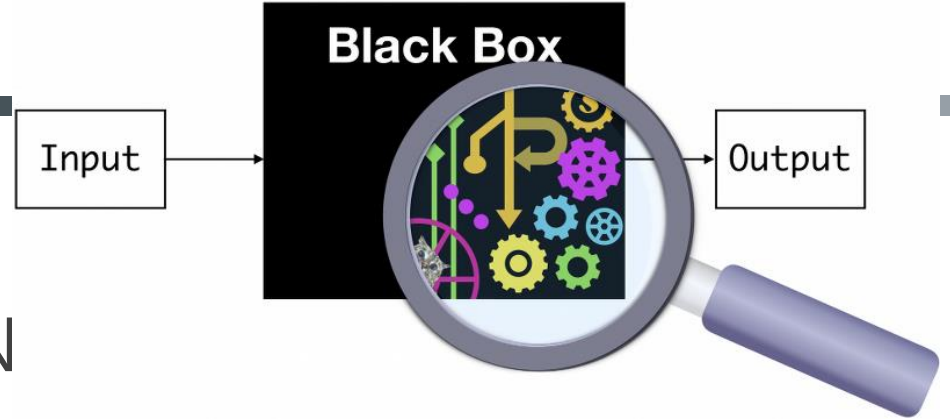
NBDT: NEURAL-BACKED DECISION TREE

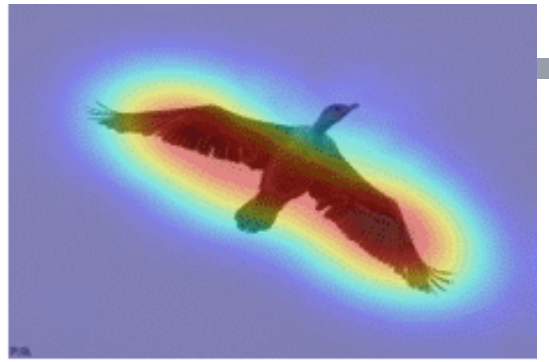
Alvin Wan[1], **Lisa Dunlap**[1]*, **Daniel Ho**[1]*, **Jihan Yin**[1], **Scott Lee**[1], **Suzanne Petryk**[1], **Sarah Adel Bargal**[2], **Joseph E. Gonzalez**[1]
UC Berkeley[1], Boston University[2]
{alvinwan,ldunlap,danielho,jihan_yin,scott.lee.3898,spetryk,jegonzal}@berkeley.edu
sbargal@bu.edu

2021. 08. 27.

Hyunsoo, Yu

INDEX

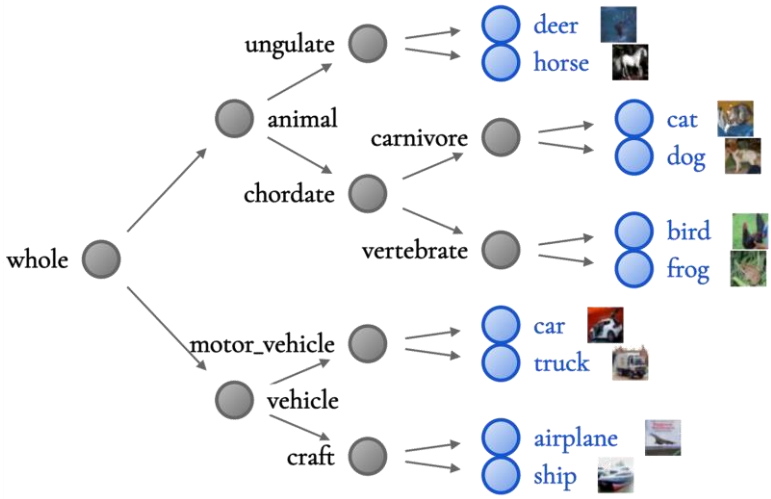




Saliency map

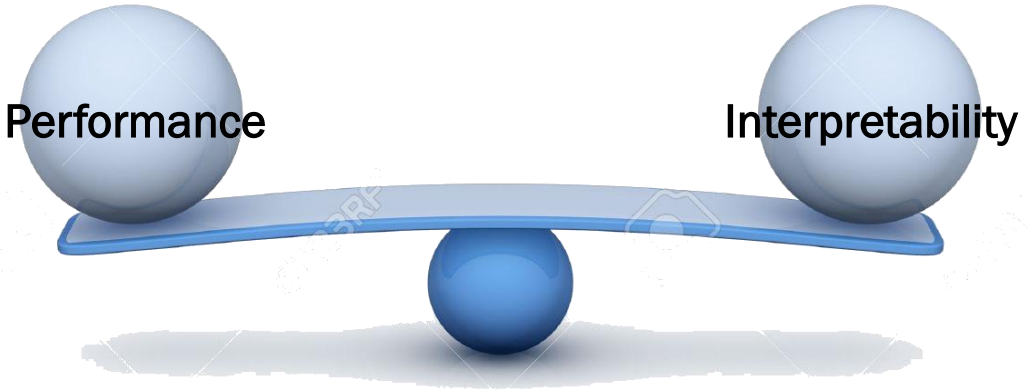
1. INTRODUCTION

- Many computer vision application required to show **the model's decision process**
  - Sophisticate deep learning are regarded as black box traditionally
- Efforts in explainable computer vision
  - **Saliency map**
  - **Sequential decision process**
- Saliency map can't capture the model's decision-making process
- But **Rule-based model(e.g., decision tree)** can dismantle predictions into a sequence of smaller semantically meaningful decision process



Sequential Decision process

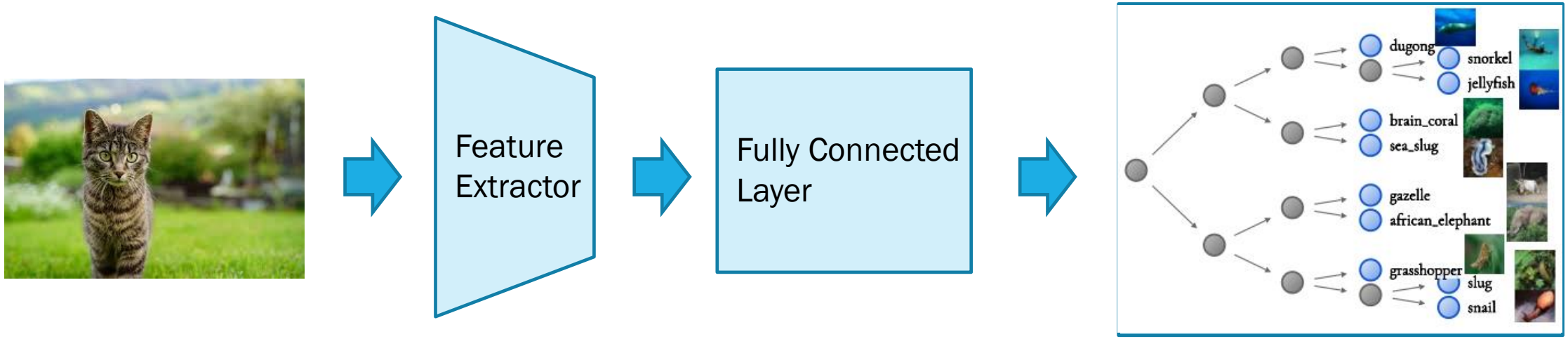
1. INTRODUCTION



- Many efforts to combine deep learning & decision trees suffer from
  - Significant accuracy loss (a)
  - Reduced interpretability due to accuracy optimization (b)
  - Tree structures that offer limited insights into the model credibility (c)
- So **Neural-Backed Decision Trees (NBDTs)** is proposed to improve both accuracy (a) & interpretability (b) with preserving properties like sequential and discrete decision (c).

2. METHOD

- NBDTs replace a network's final linear layer with a decision tree
  - NBDTs use **path probabilities for inference** to tolerate highly uncertain intermediate decision (2.1)
  - NBDTs build **a hierarchy from pre-trained model weights** to lessen overfitting (2.2 & 2.3)
  - NBDTs train with **a hierarchical loss** to significantly better learn high-level decisions (2.4)

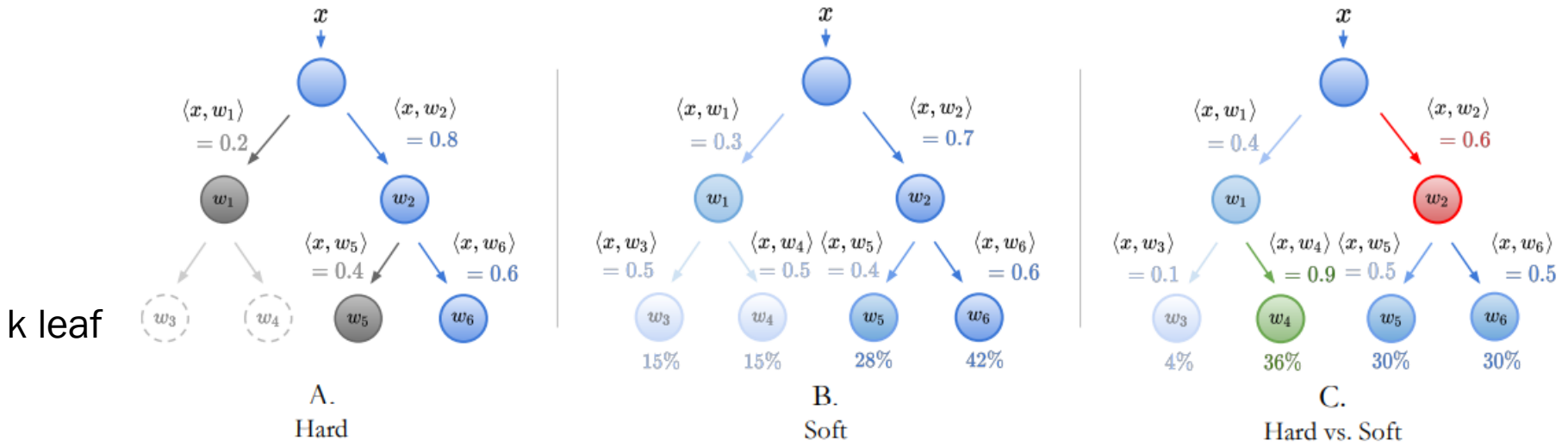


Induced Hierarchies

2. METHOD
2.1 **INFERENCE**

path probabilities for inference

- Soft Decision tree vs Hard Decision tree
  - Soft decision tree can recover from a root's mistake



k leaf

- Probability of leaf K

$$p(k) = \Pi_{i \in P_k} p(C_k(i)|i)$$

- Final Prediction

$$\hat{k} = \text{argmax}_k p(k) = \text{argmax}_k \Pi_{i \in P_k} p(C_k(i)|i)$$

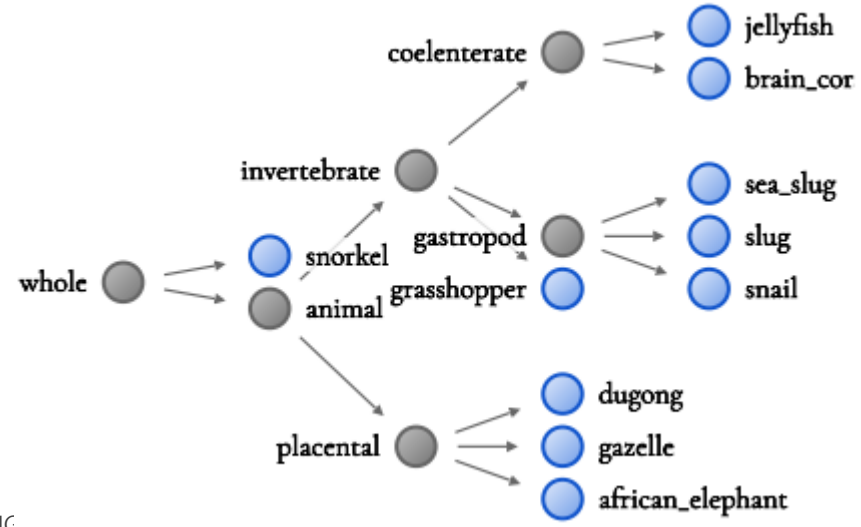
C_k : {child}

P_k : {path}

a hierarchy from pre-trained model weights

2. METHOD
2.3 **BUILDING INDUCED HIERARCHIES**

- Conventional decision-tree-based methods use
  - Hierarchies built with data-dependent heuristics(e.g., Information Gain)
    - Induce overfitting due to visual similarity
  - Existing hierarchies like WordNet



sky

vs

bird

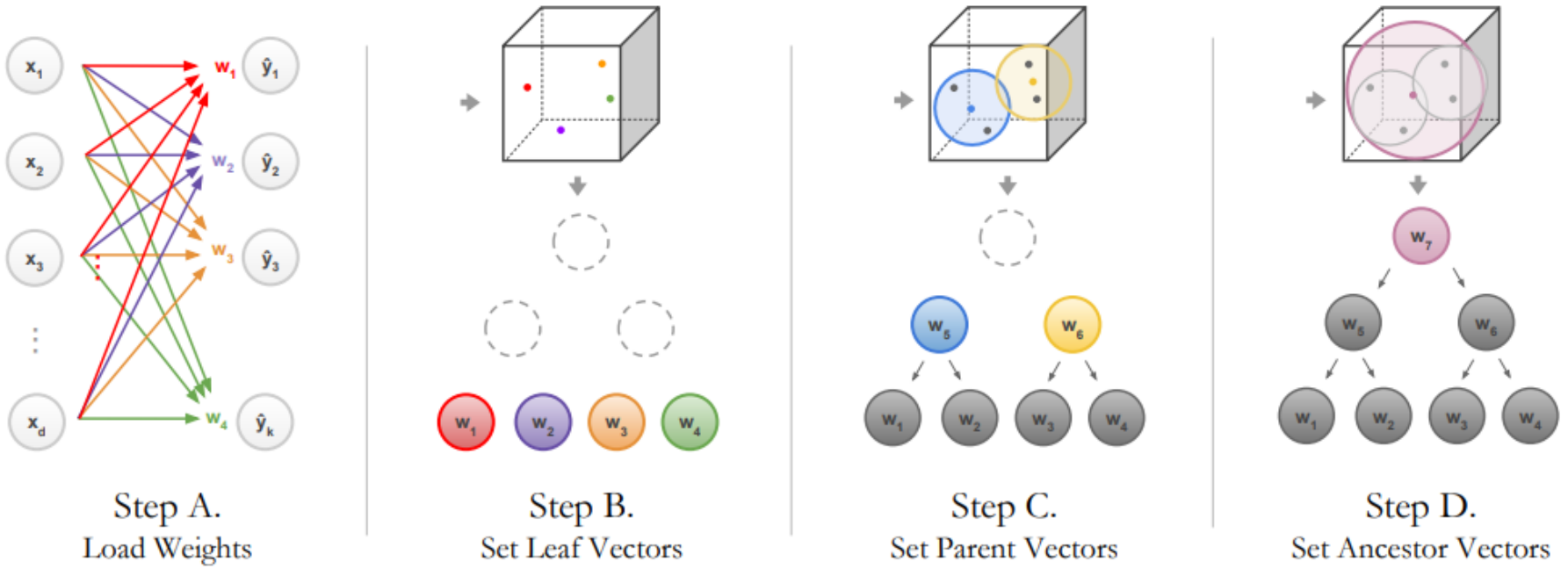
cat

vs

bird

a hierarchy from pre-trained model weights

2. METHOD
2.3 **BUILDING INDUCED HIERARCHIES**

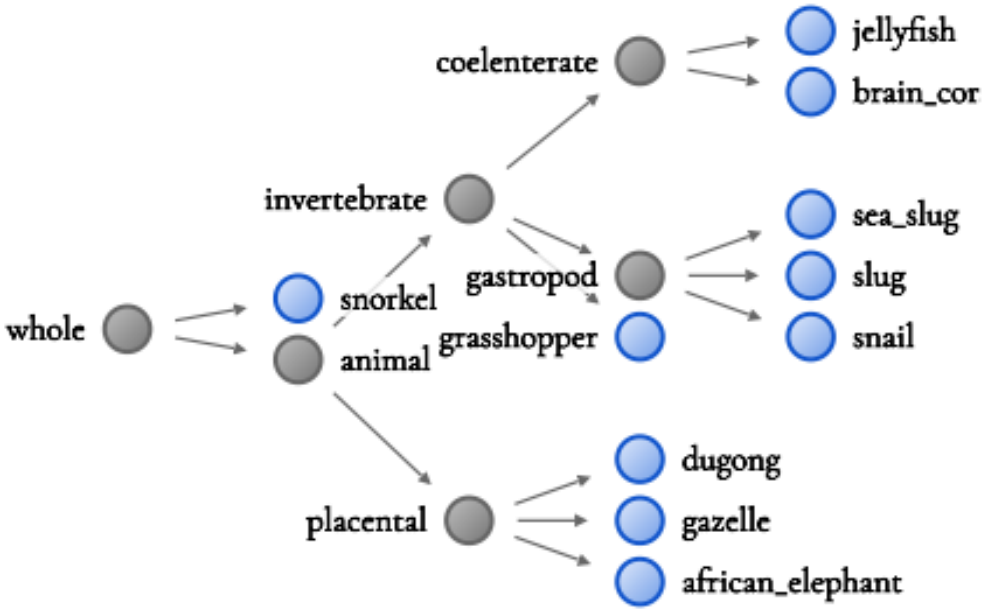


- A. From pre-trained model, the fully-connected layer weights are loaded
- B~D. run hierarchical agglomerative clustering from leaf vectors to the ancestor vectors.

a hierarchy from pre-trained model weights

2. METHOD
2.3 **LABELING DECISION NODES WITH WORDNET**

- WordNet used to assign meaning to nodes
  - Hierarchy of nouns



(a) WordNet Hierarchy

a hierarchical loss

2. METHOD
2.4 **FINE-TUNING WITH TREE SUPERVISION LOSS**

$$\mathcal{L} = \beta_t \underbrace{\textsc{CrossEntropy}(\mathcal{D}_{\text{pred}}, \mathcal{D}_{\text{label}})}_{\mathcal{L}_{\text{original}}} + \omega_t \underbrace{\textsc{CrossEntropy}(\mathcal{D}_{\text{nbdt}}, \mathcal{D}_{\text{label}})}_{\mathcal{L}_{\text{soft}}}$$

- $\mathcal{D}_{\text{nbdt}} = \{p(k)\}_{k=1}^{K}$
- ω_t, β_t are time-varying weights
- ω_t grows linearly from $\omega_0 = 0$ to $\omega_T = 0.5$
- β_t decays linearly

3. EXPERIMENT

- NBDTs obtain state-of-the-art results for interpretable models on image classification
- Results are reported on different models(ResNet, WideResNet, EfficientNet) and datasets(CIFAR10, CIFAR100, TInyImageNet, ImageNet).

3. EXPERIMENT
3.1 RESULTS

- Small-scale datasets

Method	Backbone	Expl?	CIFAR10	CIFAR100	TinyImageNet
NN	WideResNet28x10	✗	*97.62%*	*82.09%*	*67.65%*
ANT-A*	*n/a*	✓	93.28%	*n/a*	*n/a*
DDN	NiN	✗	90.32%	68.35%	*n/a*
DCDJ	NiN	✗	*n/a*	69.0%	*n/a*
NofE	ResNet56-4x	✗	*n/a*	76.24%	*n/a*
CNN-RNN	WideResNet28x10	✓	*n/a*	76.23%	*n/a*
NBDT-S (Ours)	WideResNet28x10	✓	**97.55%**	**82.97%**	**67.72%**
NN	ResNet18	✗	*94.97%*	*75.92%*	*64.13%*
DNDF	ResNet18	✗	94.32%	67.18%	44.56%
XOC	ResNet18	✓	93.12%	*n/a*	*n/a*
DT	ResNet18	✓	93.97%	64.45%	52.09%
NBDT-S (Ours)	ResNet18	✓	**94.82%**	**77.09%**	**64.23%**

3. EXPERIMENT
3.2 ANALYSIS

- Comparison of Hierarchies

Table 2: Comparisons of Hierarchies. We demonstrate that our weight-space hierarchy bests taxonomy and data-dependent hierarchies. In particular, the induced hierarchy achieves better performance than (a) the WordNet hierarchy, (b) a classic decision tree's information gain hierarchy, built over neural features ("Info Gain"), and (c) an oblique decision tree built over neural features ("OC1").

Dataset	Backbone	Original	Induced	Info Gain	WordNet	OC1
CIFAR10	ResNet18	*94.97%*	**94.82%**	93.97%	94.37%	94.33%
CIFAR100	ResNet18	*75.92%*	**77.09%**	64.45%	74.08%	38.67%
TinyImageNet200	ResNet18	*64.13%*	**64.23%**	52.09%	60.26%	15.63%

3. EXPERIMENT
3.2 ANALYSIS

- Comparison of Losses

Table 3: Comparisons of Losses. Training the NBDT using tree supervision loss with a linearly increasing weight ("TreeSup(t)") is superior to training (a) with a constant-weight tree supervision loss ("TreeSup"), (b) with a hierarchical softmax ("HrchSmax") and (c) without extra loss terms. ("None"). Δ is the accuracy difference between our soft loss and hierarchical softmax.

Dataset	Backbone	Original	TreeSup(t)	TreeSup	None	HrchSmax
CIFAR10	ResNet18	*94.97%*	**94.82%**	94.76%	94.38%	93.97%
CIFAR100	ResNet18	*75.92%*	**77.09%**	74.92%	61.93%	74.09%
TinyImageNet200	ResNet18	*64.13%*	**64.23%**	62.74%	45.51%	61.12%

3. EXPERIMENT
3.2 ANALYSIS

- Applying proposed Loss on Original Neural Network

Table 5: Original Neural Network. We compare the model's accuracy before and after the tree supervision loss, using ResNet18, WideResNet on CIFAR100, TinyImageNet. Our loss increases the original network accuracy consistently by $\sim .8 - 2.4\%$. NN-S is the network trained with the tree supervision loss.

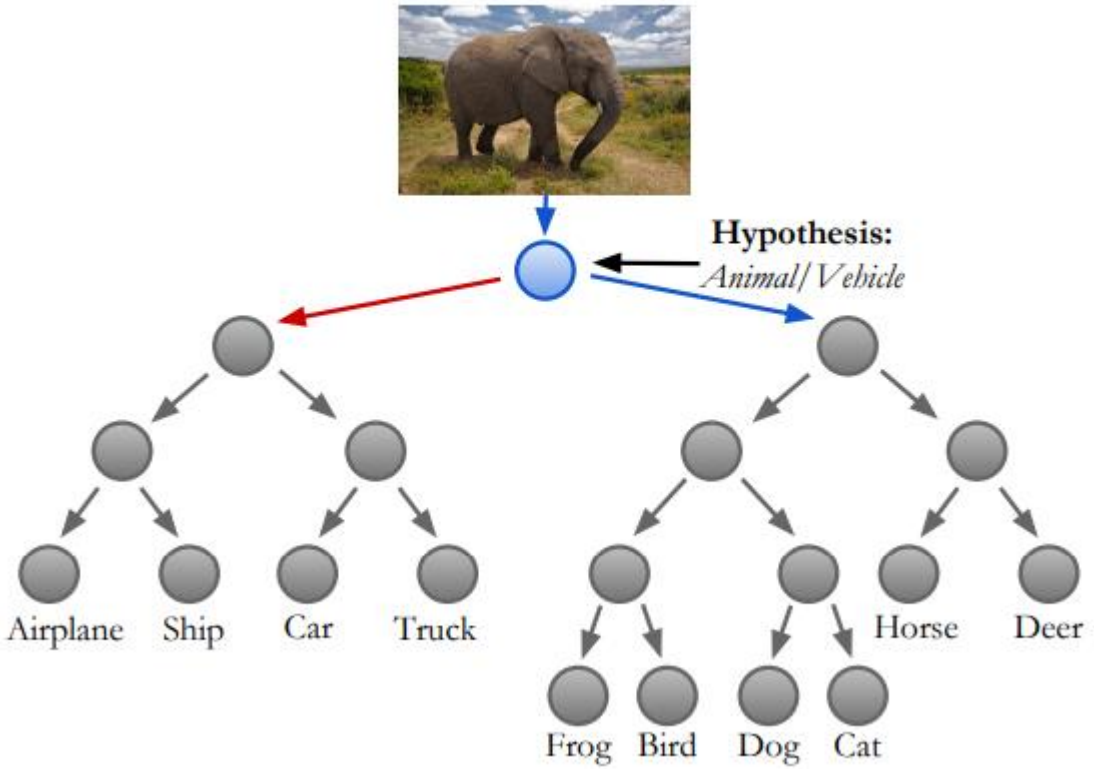
Dataset	Backbone	NN	NN-S
C100	R18	75.92%	**76.96%**
T200	R18	64.13%	**66.55%**
C100	WRN28	82.09%	**82.87%**
T200	WRN28	67.65%	**68.51%**

3. EXPERIMENT
3.2 ANALYSIS

- Zero-Shot Superclass Generalization

Table 6: Zero-Shot Superclass Generalization. We evaluate a CIFAR10-trained NBDT (ResNet18 backbone) inner node's ability to generalize beyond seen classes. We label TinyImageNet with superclass labels (e.g. label *Dog* with *Animal*) and evaluate nodes distinguishing between said superclasses. We compare to the baseline ResNet18: check if the prediction is within the right superclass.

n_{class}	Superclasses	R18	NBDT-S
71	Animal *vs.* Vehicle	66.08%	**74.79%**
36	Placental *vs.* Vertebrate	45.50%	**54.89%**
19	Carnivore *vs.* Ungulate	51.37%	**67.78%**
9	Motor Vehicle *vs.* Craft	69.33%	**77.78%**



4. INTERPRETABILITY

- To measure interpretability, an interpretability definition by Poursabzi-Sangdeh et al[1]. is adopted.
  - : A model is interpretable if a human can validate its prediction, determining when the model has made a sizable mistake.

[1] F Poursabzi-Sangdeh, D Goldstein, J Hofman, J Vaughan, and H Wallach. Manipulating and measuring model interpretability. In MLConf, 2018.

4. INTERPRETABILITY
4.1 SURVEY: IDENTIFYING FAULTY MODEL PREDICTIONS

”How well can someone detect when the model has made a sizable mistake?”

- Humans can identify misclassifications with NBDT explanations more accurately than with saliency

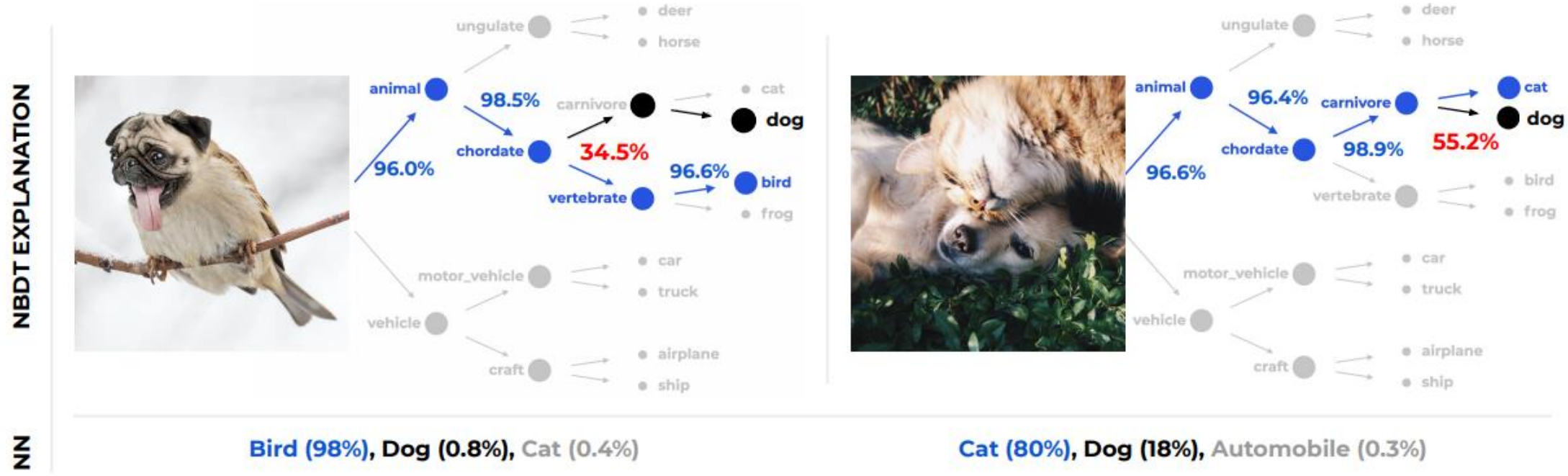
- Survey
  - Given 2 correctly classified images and 1 mis-classified images, users must predict which image was incorrectly classified.
  - Only the model explanations are given.
- Results
  - Saliency map>
    - **87** predictions were correctly identified as wrong.
  - NBDT>
    - **237** predictions were correctly identified as wrong.

4. INTERPRETABILITY
4.1 SURVEY: EXPLANATION-GUIDED IMAGE CLASSIFICATION

"To what extent do people follow a model's predictions when it is beneficial to do so?"

- Survey
  - 1. User is asked to classify a blurred image => 163/600 responses are correct.(27.2% acc)
  - 2. User is asked to classify a blurred image with model's guide.
    - => **312/600** responses agreed with NBDT & **167/600** responses agreed with saliency map
    - Even NBDT & saliency map's accuracy are 30%

4. INTERPRETABILITY
4.3 ANALYSIS : IDENTIFYING FAULTY DATASET LABELS



- We can identify ambiguous labels by finding samples with high “path entropy”

5. CONCLUSION

- Neural-Backed Decision Tree
  - Improved accuracy
  - Improved interpretability by drawing unique insights from the proposed hierarchy