ICML 2022

ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder

Sangwon Kim

Jaeyeal Nam

Byoung Chul Ko

eddiesangwonkim@gmail.com
https://cvpr.kmu.ac.kr

Interpretation of Transformers

- Multi-head self attention - Allocate pair-wise attention values between all patches

Interpretation of Transformers

- Multi-head self attention - Allocate pair-wise attention values between all patches

- Transformer Visualization

  ✓ Use attention values as relevancy scores to discover important patches

  ✓ Average the relevancy scores of multiple layers

Interpretation of Transformers

- Multi-head self attention - Allocate pair-wise attention values between all patches

- Transformer Visualization

✓ Use attention values as relevancy scores to discover important patches

✓ Average the relevancy scores of multiple layers

✗ Each layer focuses on a different patch

✗ A simple Avg. can not properly consider the role of each layer

✗ Important signals can be obscured

[IJCAI19] The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations

Interpretable Methods

- Post-hoc / Model-agnostic explanation
  - Difficult to achieve both accuracy and interpretability concurrently; heatmap generation is expensive; flexible but unreliable

Saliency-based methods

Activation Maximization

LRP

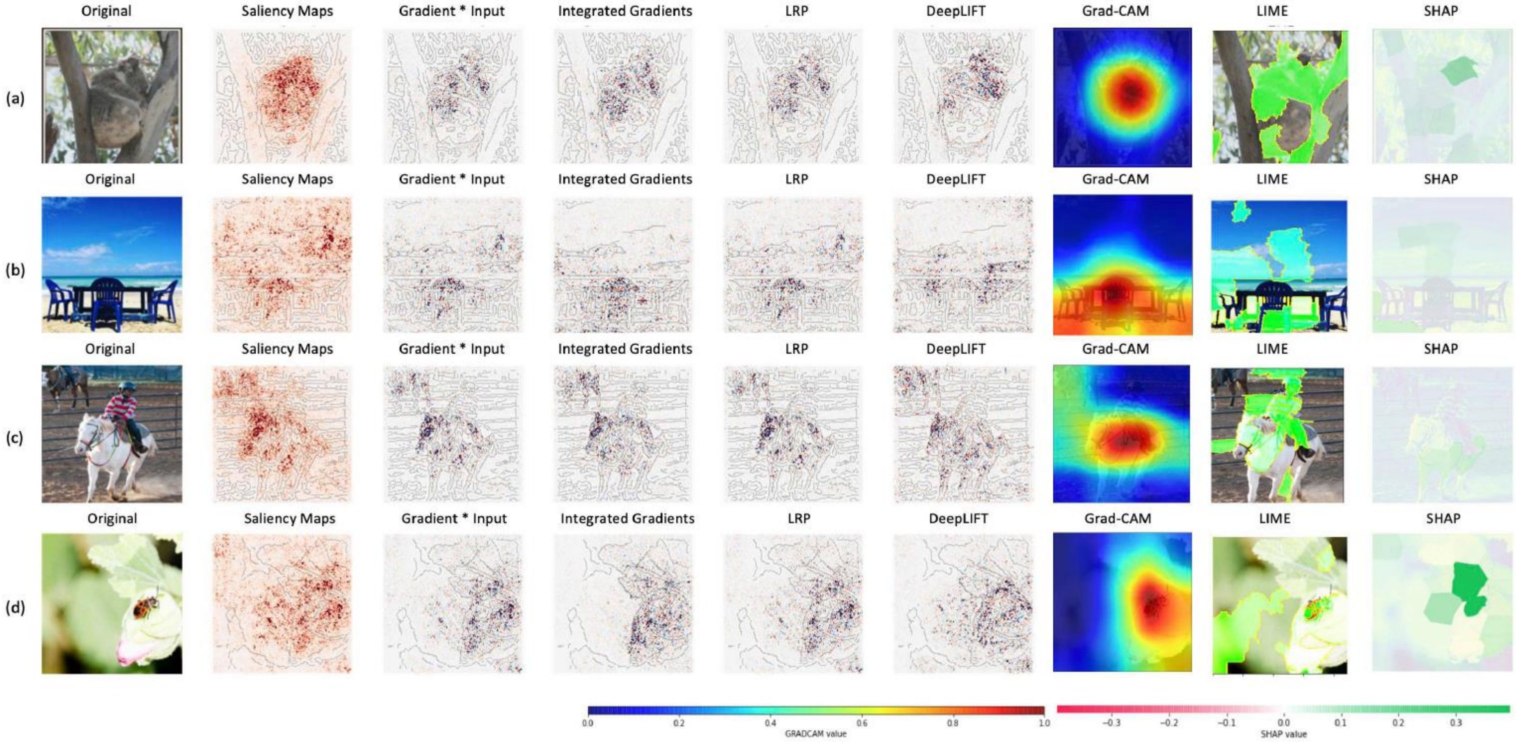
DeepLIFT

GradCAM

LIME

DeepSHAP

…



ViT-NeT

- Intrinsic interpretation

  - Integrate an interpretable model directly into existing model structures

ViT-NeT

- Intrinsic interpretation

  - Integrate an interpretable model directly into existing model structures

- Show the overall behavior of the classification model faithfully

- Provide a hierarchical description of the decision-making process

ViT-NeT

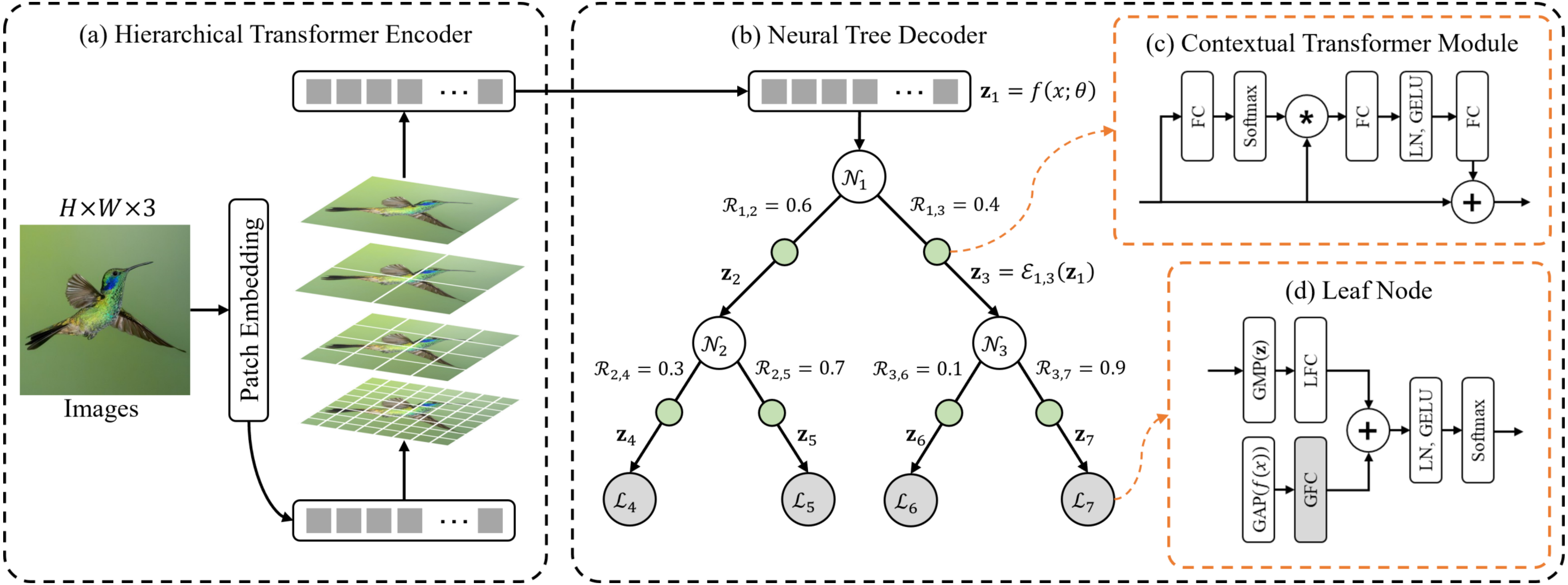
- Intrinsic interpretation

  - Integrate an interpretable model directly into existing model structures

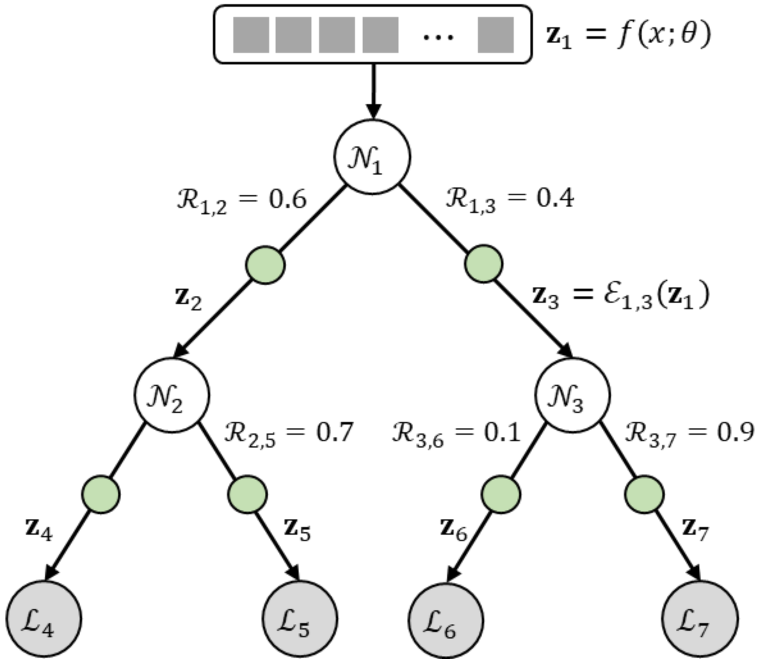
- Show the overall behavior of the classification model faithfully
- Provide a hierarchical description of the decision-making process
- Achieve both high fine-grained accuracy and Interpretability

“Interpretation by design”

ViT-NeT


(a) Hierarchical Transformer Encoder
$H \times W \times 3$
Images
Patch Embedding
(b) Neural Tree Decoder
$\mathbf{z}_1 = f(x;\theta)$
$\mathcal{N}_1$
$\mathcal{R}_{1,2} = 0.6$
$\mathcal{R}_{1,3} = 0.4$
$\mathbf{z}_2$
$\mathbf{z}_3 = \mathcal{E}_{1,3}(\mathbf{z}_1)$
$\mathcal{N}_2$
$\mathcal{N}_3$
$\mathcal{R}_{2,4} = 0.3$
$\mathcal{R}_{2,5} = 0.7$
$\mathcal{R}_{3,6} = 0.1$
$\mathcal{R}_{3,7} = 0.9$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{z}_7$
$\mathcal{L}_4$
$\mathcal{L}_5$
$\mathcal{L}_6$
$\mathcal{L}_7$
(c) Contextual Transformer Module
FC
Softmax
*
FC
LN, GELU
FC
+
(d) Leaf Node
GMP(z)
LFC
GAP(f(x))
GFC
+
LN, GELU
Softmax

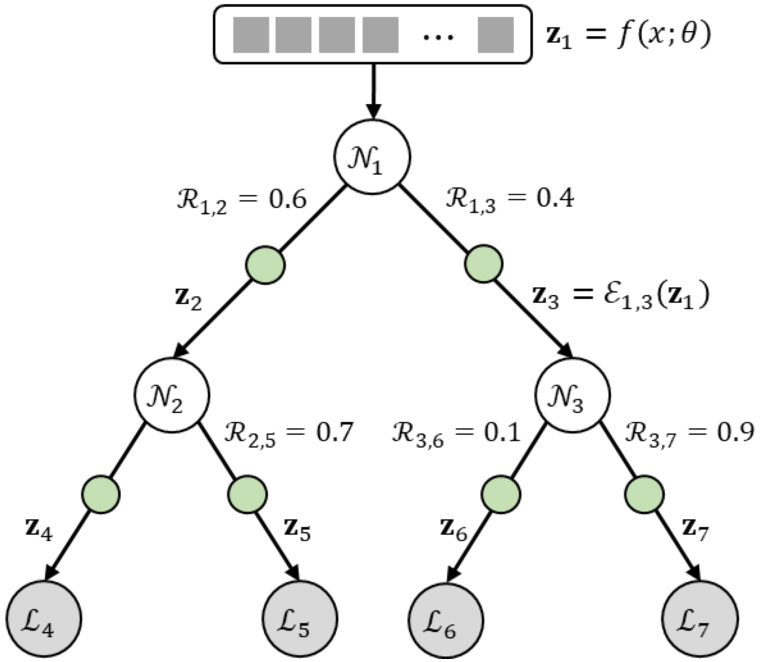
ViT-NeT



$\mathbf{z}_1 = f(x;\theta)$
$\mathcal{N}_1$
$\mathcal{R}_{1,2} = 0.6$
$\mathcal{R}_{1,3} = 0.4$
$\mathbf{z}_2$
$\mathbf{z}_3 = \mathcal{E}_{1,3}(\mathbf{z}_1)$
$\mathcal{N}_2$
$\mathcal{N}_3$
$\mathcal{R}_{2,5} = 0.7$
$\mathcal{R}_{3,6} = 0.1$
$\mathcal{R}_{3,7} = 0.9$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{z}_7$
$\mathcal{L}_4$
$\mathcal{L}_5$
$\mathcal{L}_6$
$\mathcal{L}_7$

- **Node ($\mathcal{N}_i$)** – Routing score ($\mathcal{R}$) calculation using prototype vector ($\mathcal{P}$)

ViT-NeT



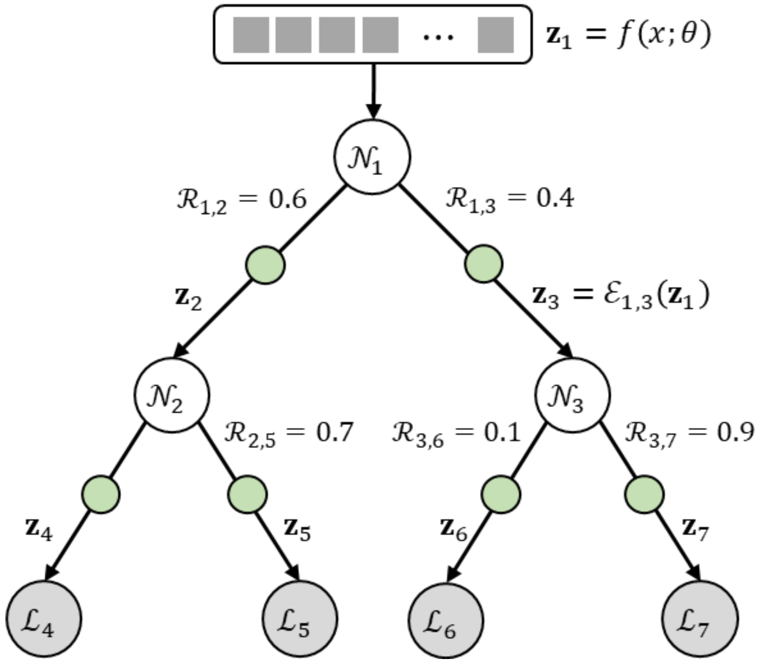
$\mathbf{z}_1 = f(x;\theta)$
$\mathcal{N}_1$
$\mathcal{R}_{1,2} = 0.6$
$\mathcal{R}_{1,3} = 0.4$
$\mathbf{z}_2$
$\mathbf{z}_3 = \mathcal{E}_{1,3}(\mathbf{z}_1)$
$\mathcal{N}_2$
$\mathcal{N}_3$
$\mathcal{R}_{2,5} = 0.7$
$\mathcal{R}_{3,6} = 0.1$
$\mathcal{R}_{3,7} = 0.9$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{z}_7$
$\mathcal{L}_4$
$\mathcal{L}_5$
$\mathcal{L}_6$
$\mathcal{L}_7$

- **Node ($\mathcal{N}_i$)** – Routing score ($\mathcal{R}$) calculation using prototype vector ($\mathcal{P}$)

1) Similarity ($\mathcal{P} \Leftrightarrow \mathbf{z}$)

$$\mathcal{N}(\mathbf{z}_i) = \max_{\tilde{\mathbf{z}} \in \mathrm{patches}(\mathbf{z}_i)} \log((\|\tilde{\mathbf{z}} - \mathcal{P}_i\|_2^2 + 1)/(\|\tilde{\mathbf{z}} - \mathcal{P}_i\|_2^2 + \epsilon))$$

ViT-NeT



$\mathbf{z}_1 = f(x;\theta)$
$\mathcal{N}_1$
$\mathcal{R}_{1,2} = 0.6$
$\mathcal{R}_{1,3} = 0.4$
$\mathbf{z}_2$
$\mathbf{z}_3 = \mathcal{E}_{1,3}(\mathbf{z}_1)$
$\mathcal{N}_2$
$\mathcal{N}_3$
$\mathcal{R}_{2,5} = 0.7$
$\mathcal{R}_{3,6} = 0.1$
$\mathcal{R}_{3,7} = 0.9$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{z}_7$
$\mathcal{L}_4$
$\mathcal{L}_5$
$\mathcal{L}_6$
$\mathcal{L}_7$

- **Node ($\mathcal{N}_i$)** – Routing score ($\mathcal{R}$) calculation using prototype vector ($\mathcal{P}$)

1) Similarity ($\mathcal{P} \Leftrightarrow \mathbf{z}$)

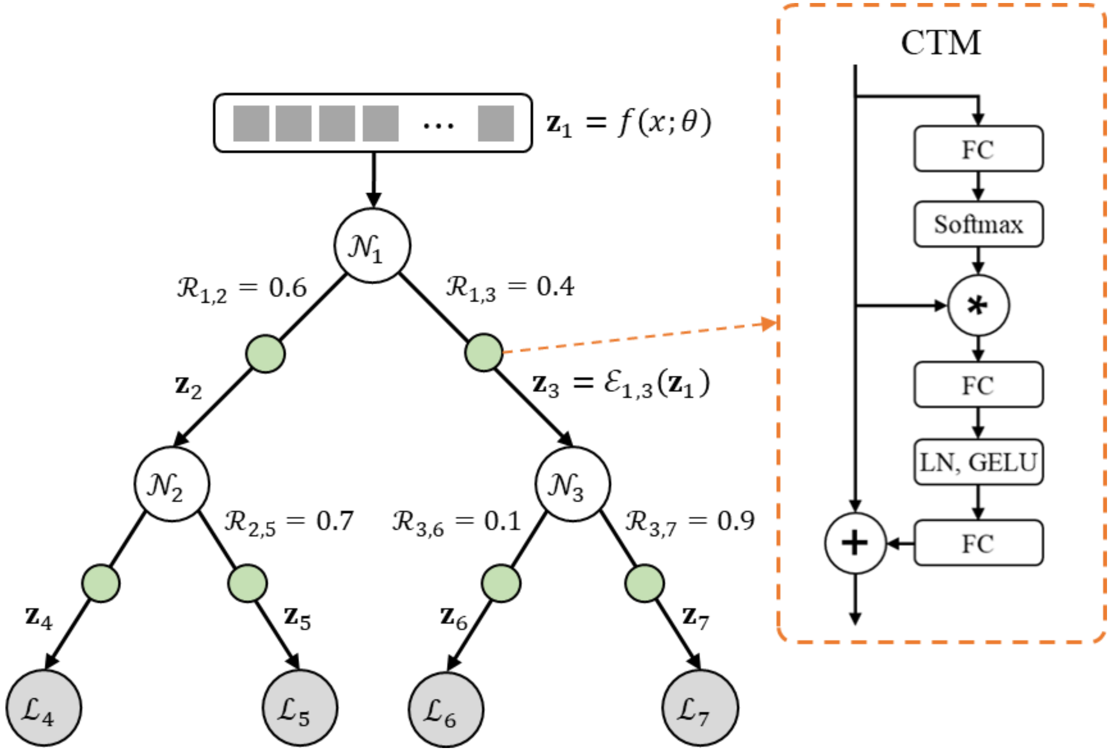
$$\mathcal{N}(\mathbf{z}_i) = \max_{\tilde{\mathbf{z}} \in \text{patches}(\mathbf{z}_i)} \log((\|\tilde{\mathbf{z}} - \mathcal{P}_i\|_2^2 + 1)/(\|\tilde{\mathbf{z}} - \mathcal{P}_i\|_2^2 + \epsilon))$$

2) Calculate Routing score based on similarity

$\mathcal{R}_{i,2\times i}(\mathbf{z}_i) = [\mathcal{N}(\mathbf{z}_i)]_0^1$ //The right child node

$\mathcal{R}_{i,2\times i+1}(\mathbf{z}_i) = 1 - [\mathcal{N}(\mathbf{z}_i)]_0^1$ //The left child node

ViT-NeT

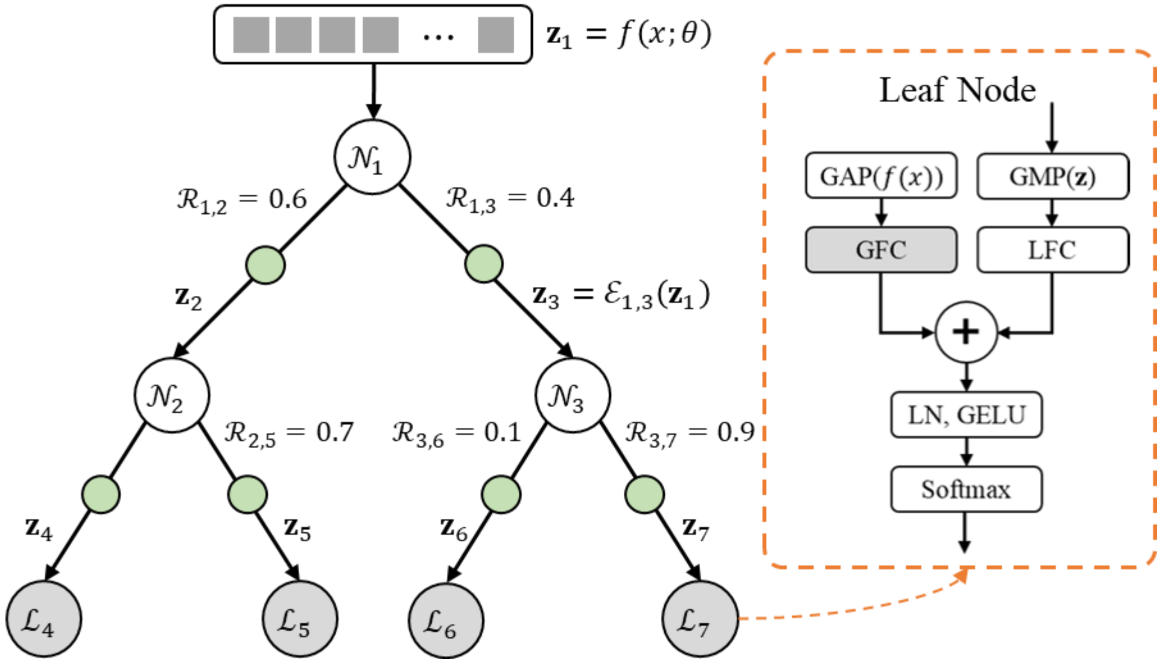


$\mathbf{z}_1 = f(x;\theta)$
$\mathcal{N}_1$
$\mathcal{R}_{1,2} = 0.6$
$\mathcal{R}_{1,3} = 0.4$
$\mathbf{z}_2$
$\mathbf{z}_3 = \mathcal{E}_{1,3}(\mathbf{z}_1)$
$\mathcal{N}_2$
$\mathcal{N}_3$
$\mathcal{R}_{2,5} = 0.7$
$\mathcal{R}_{3,6} = 0.1$
$\mathcal{R}_{3,7} = 0.9$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{z}_7$
$\mathcal{L}_4$
$\mathcal{L}_5$
$\mathcal{L}_6$
$\mathcal{L}_7$
CTM
FC
Softmax
*
FC
LN, GELU
FC
+

- **Edge($\mathcal{E}_{i,j}$)**

$$\mathbf{z}_j = \mathcal{E}_{i,j}(\mathbf{z}_i) = \mathrm{CTM}(\mathbf{z}_i)$$

ViT-NeT



- **Edge($\mathcal{E}_{i,j}$)**

$$\mathbf{z}_j = \mathcal{E}_{i,j}(\mathbf{z}_i) = \mathrm{CTM}(\mathbf{z}_i)$$

- **Leaf ($\mathcal{L}_i$)**

$$\rho(\mathbf{z}_i) = \prod_{(i,j)\in p} \mathcal{R}_{i,j}\left(\mathcal{E}_{i,j}(\mathbf{z}_i)\right)$$

$$\mathcal{L}(\mathbf{z}_l, x) = \mathrm{LN}\left(\mathrm{FC}\left(\mathrm{GMP}(\mathbf{z}_l)\right)\right) + \mathrm{FC}\left(\mathrm{GAP}\left(f(x;\theta)\right)\right)$$

$$\hat{y} = \sum_{l\in\mathbb{L}} \sigma\left(\mathcal{L}(z_l, x)\right) \cdot \rho_l(\mathbf{z}_1)$$

Experiments

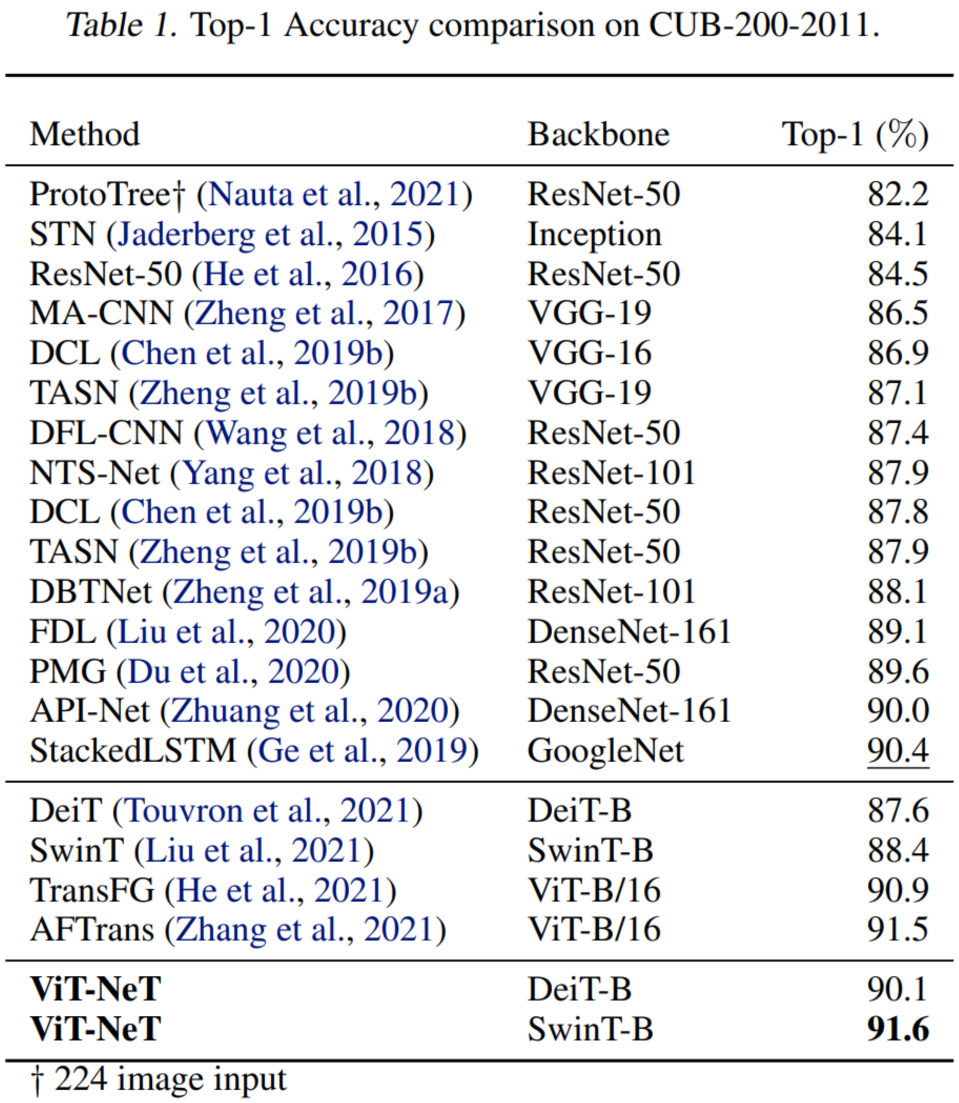


Table 1. Top-1 Accuracy comparison on CUB-200-2011.

Method	Backbone	Top-1 (%)
ProtoTree† (Nauta et al., 2021)	ResNet-50	82.2
STN (Jaderberg et al., 2015)	Inception	84.1
ResNet-50 (He et al., 2016)	ResNet-50	84.5
MA-CNN (Zheng et al., 2017)	VGG-19	86.5
DCL (Chen et al., 2019b)	VGG-16	86.9
TASN (Zheng et al., 2019b)	VGG-19	87.1
DFL-CNN (Wang et al., 2018)	ResNet-50	87.4
NTS-Net (Yang et al., 2018)	ResNet-101	87.9
DCL (Chen et al., 2019b)	ResNet-50	87.8
TASN (Zheng et al., 2019b)	ResNet-50	87.9
DBTNet (Zheng et al., 2019a)	ResNet-101	88.1
FDL (Liu et al., 2020)	DenseNet-161	89.1
PMG (Du et al., 2020)	ResNet-50	89.6
API-Net (Zhuang et al., 2020)	DenseNet-161	90.0
StackedLSTM (Ge et al., 2019)	GoogleNet	90.4
DeiT (Touvron et al., 2021)	DeiT-B	87.6
SwinT (Liu et al., 2021)	SwinT-B	88.4
TransFG (He et al., 2021)	ViT-B/16	90.9
AFTrans (Zhang et al., 2021)	ViT-B/16	91.5
ViT-NeT	DeiT-B	90.1
ViT-NeT	SwinT-B	**91.6**

† 224 image input

Experiments

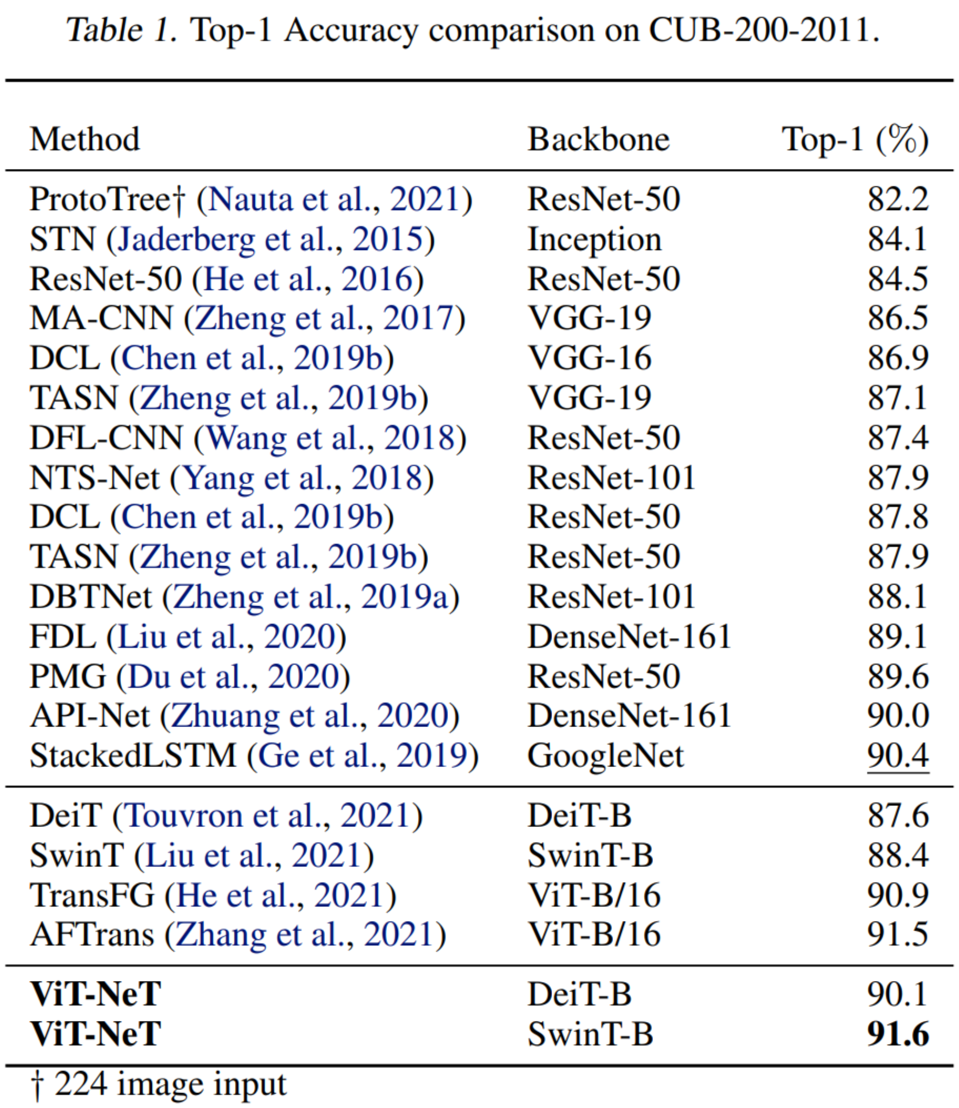


Table 1. Top-1 Accuracy comparison on CUB-200-2011.

Method	Backbone	Top-1 (%)
ProtoTree† (Nauta et al., 2021)	ResNet-50	82.2
STN (Jaderberg et al., 2015)	Inception	84.1
ResNet-50 (He et al., 2016)	ResNet-50	84.5
MA-CNN (Zheng et al., 2017)	VGG-19	86.5
DCL (Chen et al., 2019b)	VGG-16	86.9
TASN (Zheng et al., 2019b)	VGG-19	87.1
DFL-CNN (Wang et al., 2018)	ResNet-50	87.4
NTS-Net (Yang et al., 2018)	ResNet-101	87.9
DCL (Chen et al., 2019b)	ResNet-50	87.8
TASN (Zheng et al., 2019b)	ResNet-50	87.9
DBTNet (Zheng et al., 2019a)	ResNet-101	88.1
FDL (Liu et al., 2020)	DenseNet-161	89.1
PMG (Du et al., 2020)	ResNet-50	89.6
API-Net (Zhuang et al., 2020)	DenseNet-161	90.0
StackedLSTM (Ge et al., 2019)	GoogleNet	90.4
DeiT (Touvron et al., 2021)	DeiT-B	87.6
SwinT (Liu et al., 2021)	SwinT-B	88.4
TransFG (He et al., 2021)	ViT-B/16	90.9
AFTrans (Zhang et al., 2021)	ViT-B/16	91.5
ViT-NeT	DeiT-B	90.1
ViT-NeT	SwinT-B	**91.6**

† 224 image input

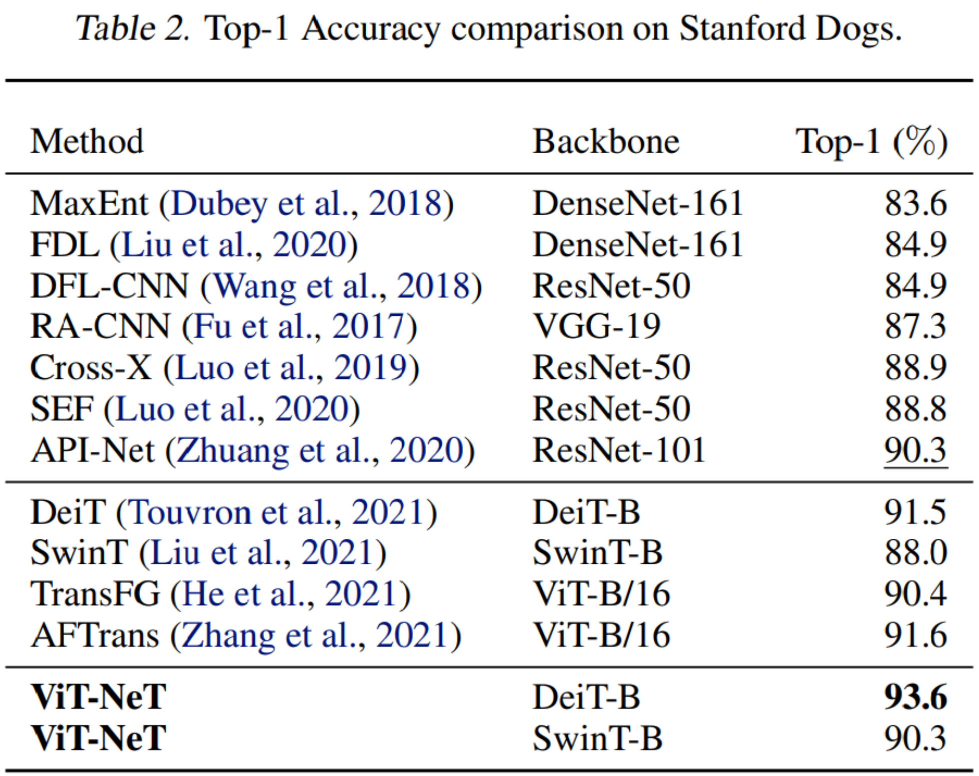


Table 2. Top-1 Accuracy comparison on Stanford Dogs.

Method	Backbone	Top-1 (%)
MaxEnt (Dubey et al., 2018)	DenseNet-161	83.6
FDL (Liu et al., 2020)	DenseNet-161	84.9
DFL-CNN (Wang et al., 2018)	ResNet-50	84.9
RA-CNN (Fu et al., 2017)	VGG-19	87.3
Cross-X (Luo et al., 2019)	ResNet-50	88.9
SEF (Luo et al., 2020)	ResNet-50	88.8
API-Net (Zhuang et al., 2020)	ResNet-101	90.3
DeiT (Touvron et al., 2021)	DeiT-B	91.5
SwinT (Liu et al., 2021)	SwinT-B	88.0
TransFG (He et al., 2021)	ViT-B/16	90.4
AFTrans (Zhang et al., 2021)	ViT-B/16	91.6
ViT-NeT	DeiT-B	**93.6**
ViT-NeT	SwinT-B	90.3

Experiments

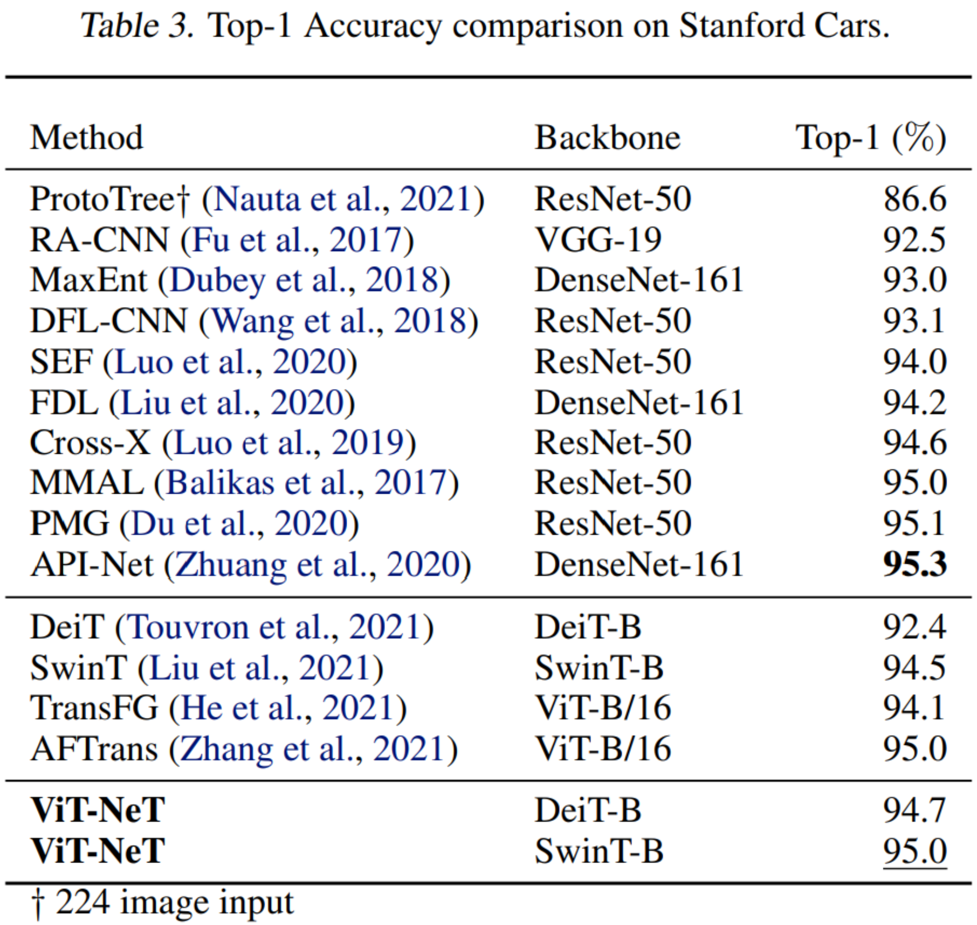


Table 3. Top-1 Accuracy comparison on Stanford Cars.

Method	Backbone	Top-1 (%)
ProtoTree† (Nauta et al., 2021)	ResNet-50	86.6
RA-CNN (Fu et al., 2017)	VGG-19	92.5
MaxEnt (Dubey et al., 2018)	DenseNet-161	93.0
DFL-CNN (Wang et al., 2018)	ResNet-50	93.1
SEF (Luo et al., 2020)	ResNet-50	94.0
FDL (Liu et al., 2020)	DenseNet-161	94.2
Cross-X (Luo et al., 2019)	ResNet-50	94.6
MMAL (Balikas et al., 2017)	ResNet-50	95.0
PMG (Du et al., 2020)	ResNet-50	95.1
API-Net (Zhuang et al., 2020)	DenseNet-161	**95.3**
DeiT (Touvron et al., 2021)	DeiT-B	92.4
SwinT (Liu et al., 2021)	SwinT-B	94.5
TransFG (He et al., 2021)	ViT-B/16	94.1
AFTrans (Zhang et al., 2021)	ViT-B/16	95.0
ViT-NeT	DeiT-B	94.7
ViT-NeT	SwinT-B	95.0

† 224 image input

Experiments

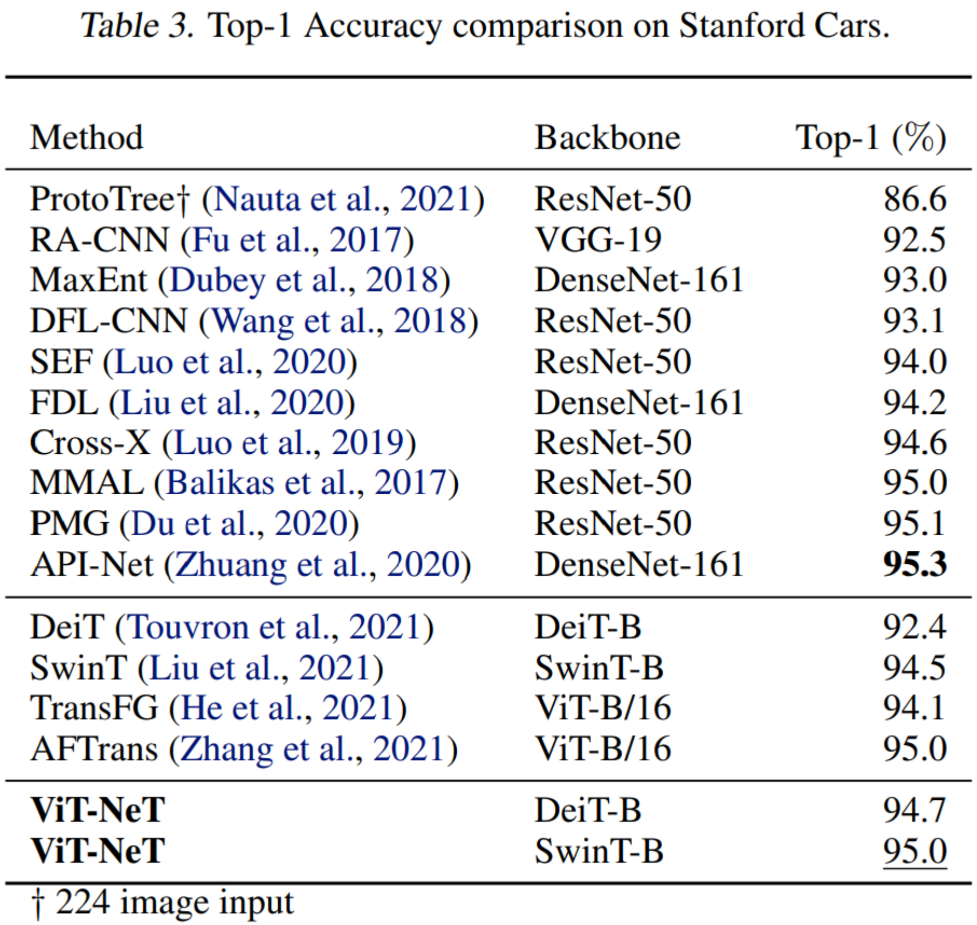


Table 3. Top-1 Accuracy comparison on Stanford Cars.

Method	Backbone	Top-1 (%)
ProtoTree† (Nauta et al., 2021)	ResNet-50	86.6
RA-CNN (Fu et al., 2017)	VGG-19	92.5
MaxEnt (Dubey et al., 2018)	DenseNet-161	93.0
DFL-CNN (Wang et al., 2018)	ResNet-50	93.1
SEF (Luo et al., 2020)	ResNet-50	94.0
FDL (Liu et al., 2020)	DenseNet-161	94.2
Cross-X (Luo et al., 2019)	ResNet-50	94.6
MMAL (Balikas et al., 2017)	ResNet-50	95.0
PMG (Du et al., 2020)	ResNet-50	95.1
API-Net (Zhuang et al., 2020)	DenseNet-161	**95.3**
DeiT (Touvron et al., 2021)	DeiT-B	92.4
SwinT (Liu et al., 2021)	SwinT-B	94.5
TransFG (He et al., 2021)	ViT-B/16	94.1
AFTrans (Zhang et al., 2021)	ViT-B/16	95.0
ViT-NeT	DeiT-B	94.7
ViT-NeT	SwinT-B	95.0

† 224 image input

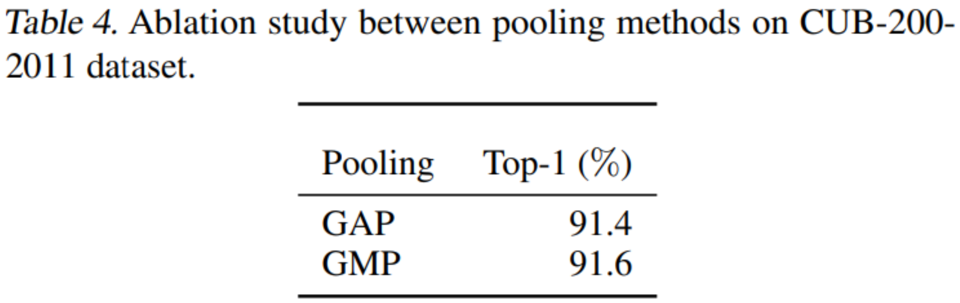


Table 4. Ablation study between pooling methods on CUB-200-2011 dataset.

Pooling	Top-1 (%)
GAP	91.4
GMP	91.6

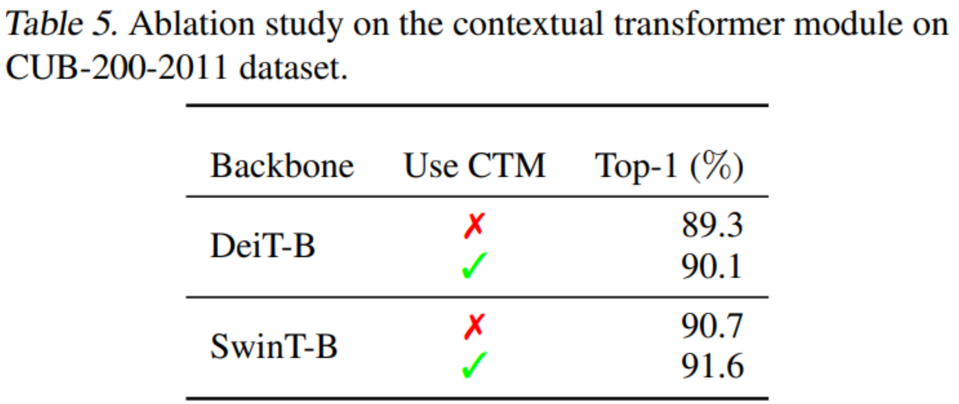


Table 5. Ablation study on the contextual transformer module on CUB-200-2011 dataset.

Backbone	Use CTM	Top-1 (%)
DeiT-B	✗	89.3
	✓	90.1
SwinT-B	✗	90.7
	✓	91.6

Experiments

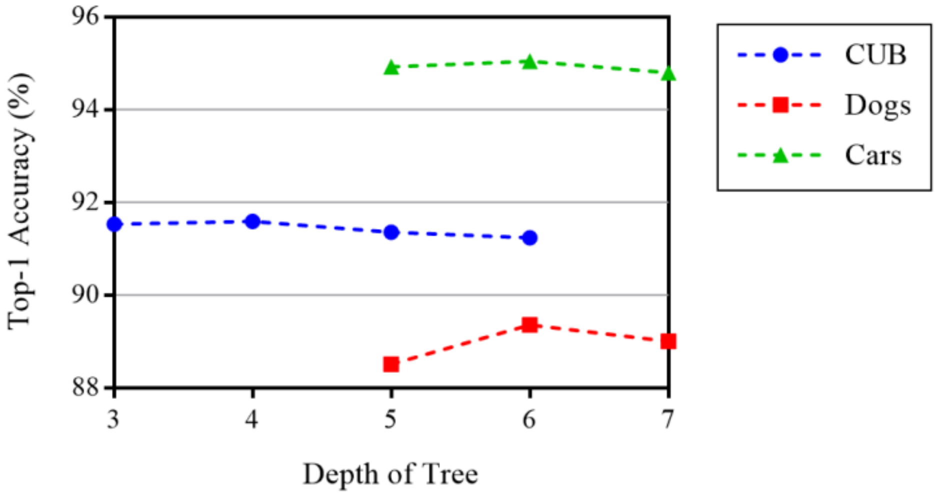


Figure 3. Top-1 Accuracy of a ViT-NeT with effect of the depth of the neural tree decoder on CUB-200-2011, Stanford Dogs and Stanford Cars.

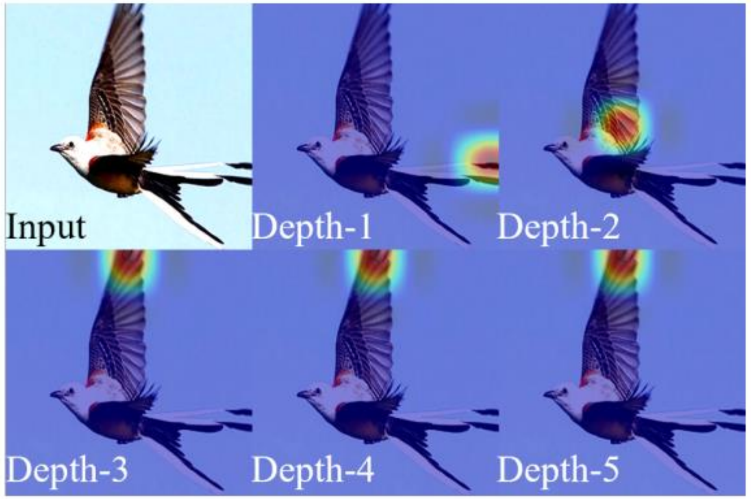


Figure 4. Prototype responses of a specific decision path in the NeT with the depth of 5. As the tree depth increases, model overfitting occurs, resulting in identical prototype responses at the tip of the bird's wing.

Experiments

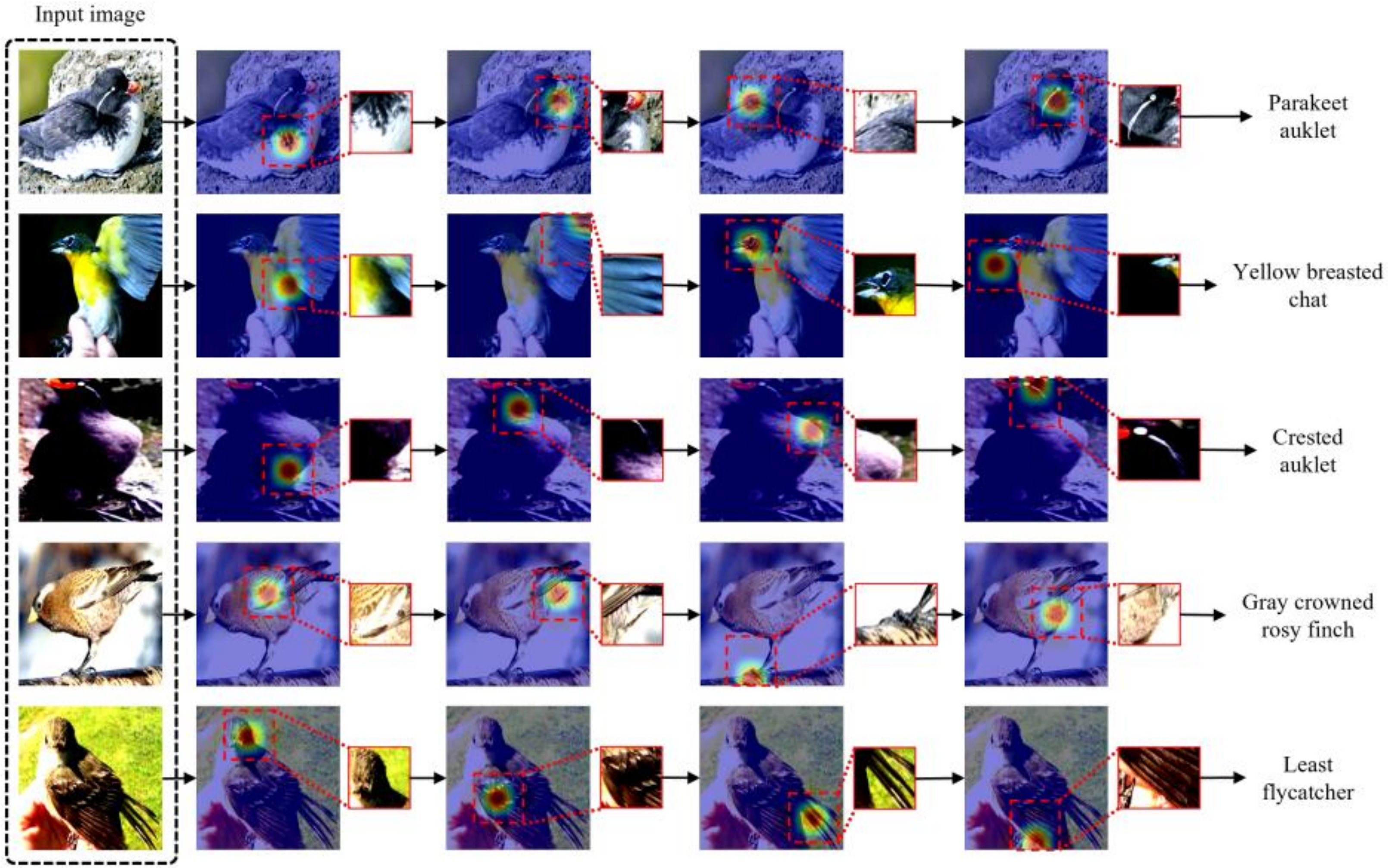


Figure 5. Visualized local interpretations showing sequential decision-making on randomly sampled images. The proposed NeT found tails, beaks, wings, feathers, claws, and eyes in the given images.

Conclusion

- The first proposal of **decision-making process interpretation** for ViT
- **Integrated interpretability** directly into the structure of classification models to provide **intrinsic interpretation**
- Provides **local interpretation** showing the routing of specific input images
- Achieved both high **accuracy** and **interpretability** of ViT

Thanks!

eddiesangwonkim@gmail.com

https://github.com/jumpsnack/ViT-NeT