tinyML® Research Symposium

Enabling Ultra-low Power Machine Learning at the Edge

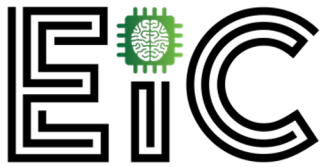
March 27, 2023

www.tinyML.org

AugViT: Improving Vision Transformer Training by Marrying Attention and Data Augmentation

Zhongzhi Yu, Yonggan Fu, Chaojian Li, and Yingyan (Celine) Lin

Georgia Institute of Technology



Efficient and Intelligent Computing Lab

Background: ViTs are Powerful

- **Vision Transformers (ViTs)** achieve SOTA accuracies across various vision tasks

Background: ViTs are Powerful

- **Vision Transformers (ViTs)** achieve SOTA accuracies across various vision tasks

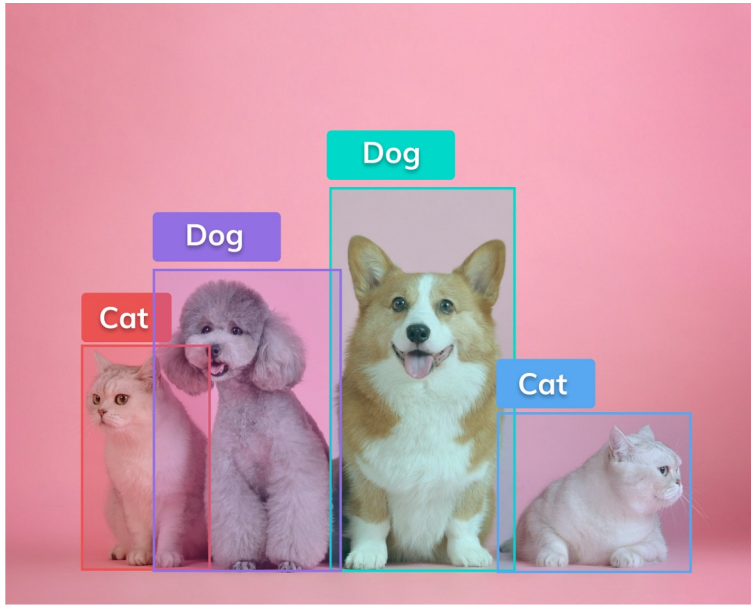
Task	CNN	ViT
Classification @ImageNet	**RegNetY: 82.9%** [I. Radosavovic, CVPR'20]	**Swin-Base: 83.5% (+0.6%)** [Z. Liu, ICML'21]

CAT

Background: ViTs are Powerful

- **Vision Transformers (ViTs)** achieve SOTA accuracies across various vision tasks

Task	CNN	ViT
Classification @ImageNet	**RegNetY: 82.9%** [I. Radosavovic, CVPR'20]	**Swin-Base: 83.5% (+0.6%)** [Z. Liu, ICML'21]
Detection @COCO	**Faster RCNN: 42.0%** [S. Ren, ICCV'15]	**DETR: 44.9% (+2.9%)** [N. Carion, ECCV'20]



Background: ViTs are Powerful

- **Vision Transformers (ViTs)** achieve SOTA accuracies across various vision tasks

Task	CNN	ViT
Classification @ImageNet	**RegNetY: 82.9%** [I. Radosavovic, CVPR'20]	**Swin-Base: 83.5% (+0.6%)** [Z. Liu, ICML'21]
Detection @COCO	**Faster RCNN: 42.0%** [S. Ren, ICCV'15]	**DETR: 44.9% (+2.9%)** [N. Carion, ECCV'20]
Segmentation @COCO	**K-Net: 54.6%** [W. Zhang, NeurIPS'21]	**Mask2Former: 57.8% (+3.2%)** [B. Cheng, CVPR'22]

Background: ViTs' Architecture

- ViTs adopt a **patch-based** image processing pipeline

Background: ViTs' Architecture

- ViTs adopt a **patch-based** image processing pipeline

[A. Dosovitskiy, ICLR'21]

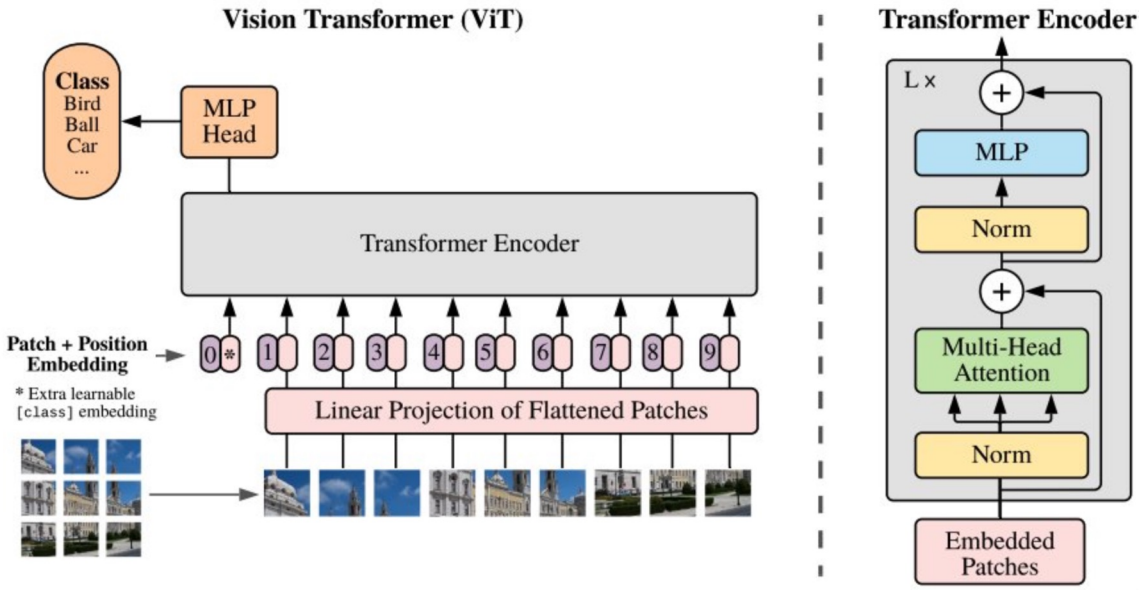
Background: ViTs' Architecture

- ViTs adopt a **patch-based** image processing pipeline

[A. Dosovitskiy, ICLR'21]

Background: ViTs' Architecture

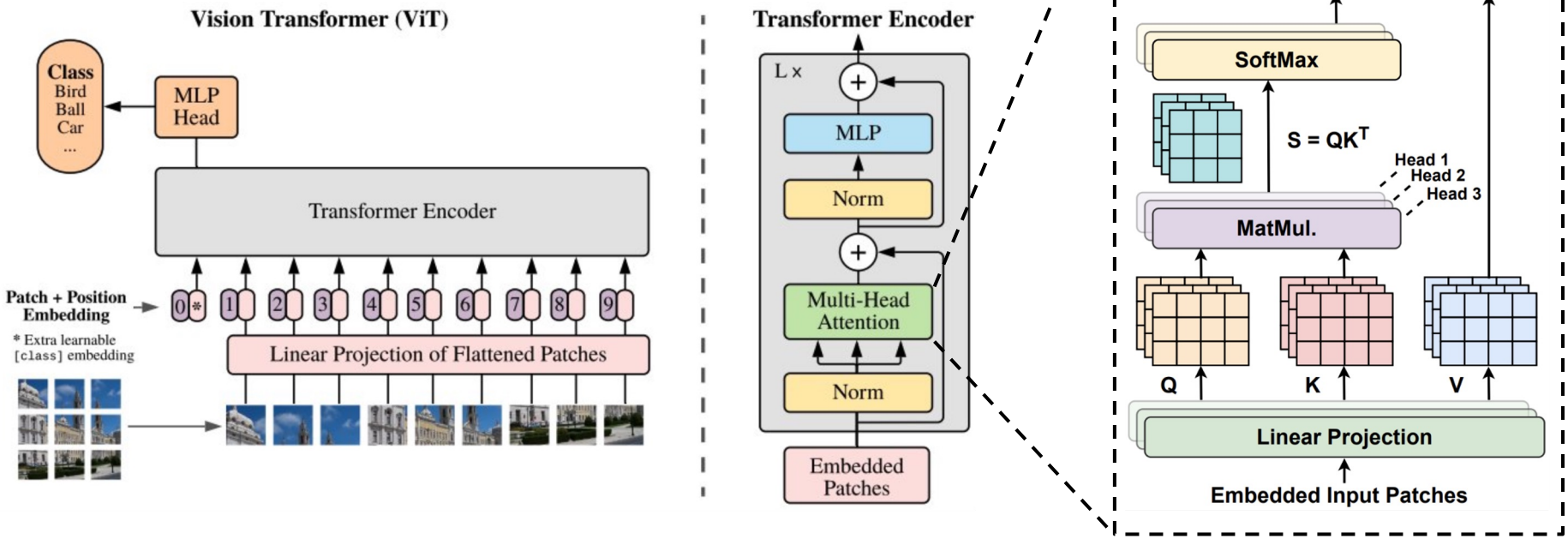
- ViTs adopt a **patch-based** image processing pipeline



[A. Dosovitskiy, ICLR'21]

Background: ViTs' Architecture

- ViTs adopt a **patch-based** image processing pipeline
  - **Self-attention**: Key component to extract **patch-wise correlations**

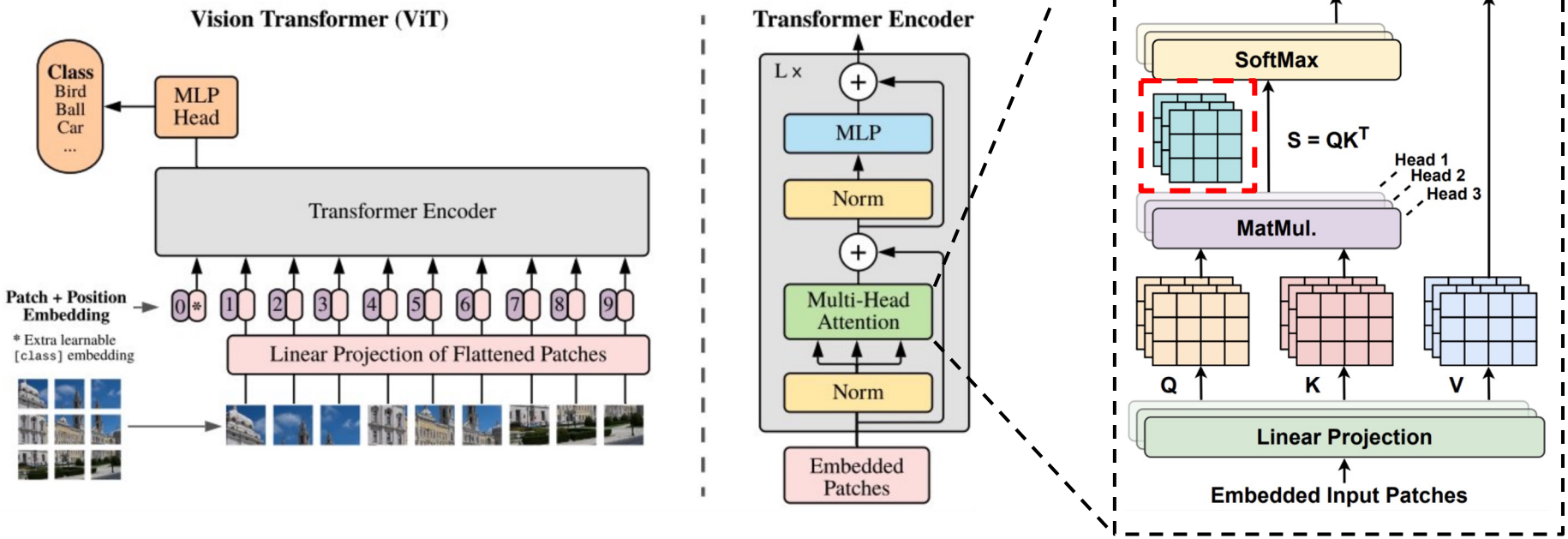


[A. Dosovitskiy, ICLR'21]

[H. You, HPCA'23]

Background: ViTs' Architecture

- ViTs adopt a **patch-based** image processing pipeline
  - **Self-attention**: Key component to extract **patch-wise correlations**



[A. Dosovitskiy, ICLR'21]

[H. You, HPCA'23]

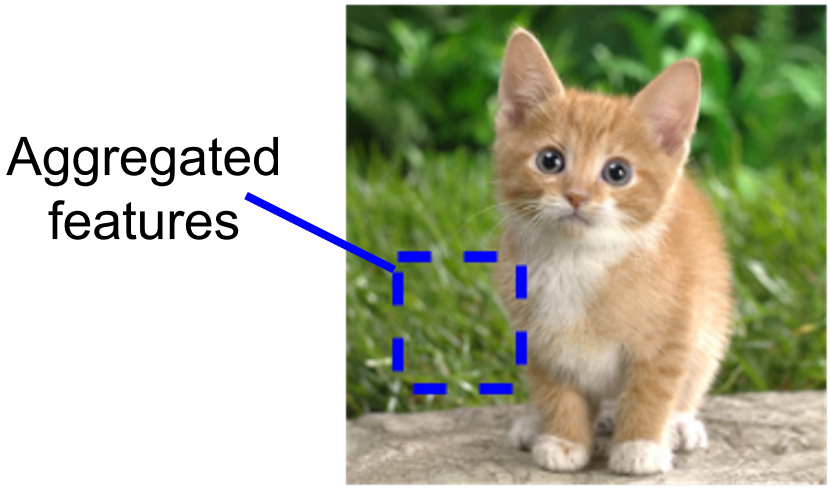
Challenge: ViTs Lack of Inductive Bias

Challenge: ViTs Lack of Inductive Bias

- **CNN**: Features are **locally aggregated**

  → Nearby features are generally more related

CNN aggregates nearby features



Challenge: ViTs Lack of Inductive Bias

- **CNN**: Features are **locally aggregated**

  → Nearby features are generally more related

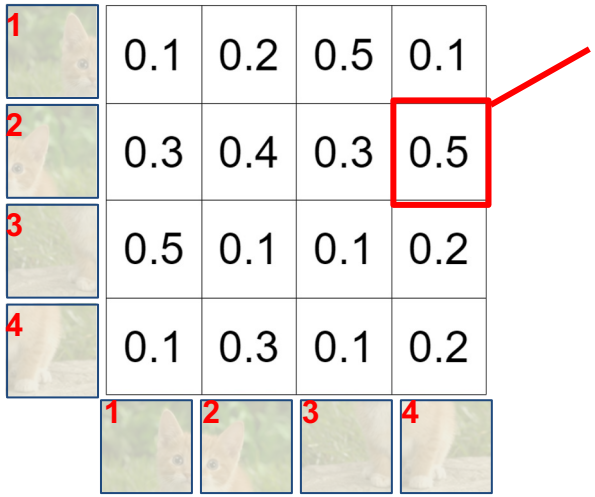
- **ViT**: Computes **patch-wise correlation**

  → No assumption on feature relationships

CNN aggregates nearby features

Aggregated features

ViT correlates all patches



Correlation between 2nd and 4th patches

Challenge: ViTs Lack of Inductive Bias

- **CNN**: Features are **locally aggregated**

  → Nearby features are generally more related

- **ViT**: Computes **patch-wise correlation**

  → No assumption on feature relationships

ViTs need dedicatedly designed

data augmentation

>10% accuracy change on ImageNet when using different data augmentations *[A. Steiner, TMLR'22]*

Limitation of Existing Data Augmentation Techniques

Original Image

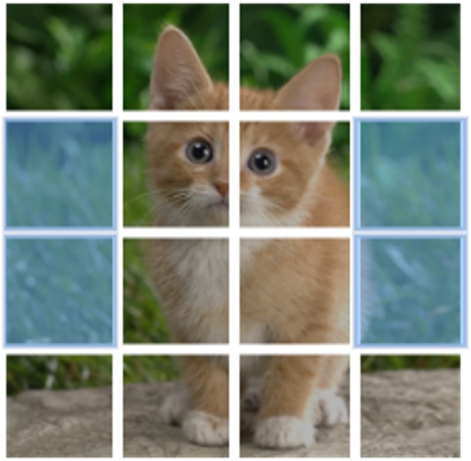
Image-wise
Augmentation

Region-wise
Augmentation

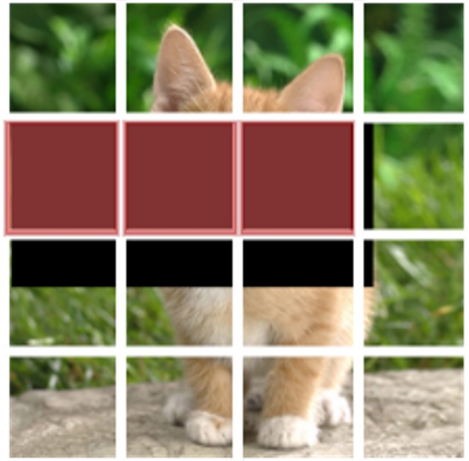
Limitation of Existing Data Augmentation Techniques

Original Image

Image-wise
Augmentation



Region-wise
Augmentation

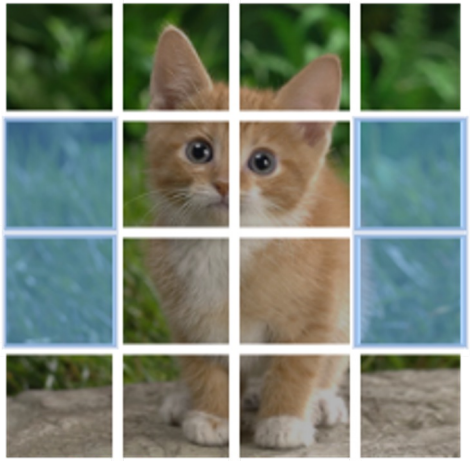


Limitation of Existing Data Augmentation Techniques

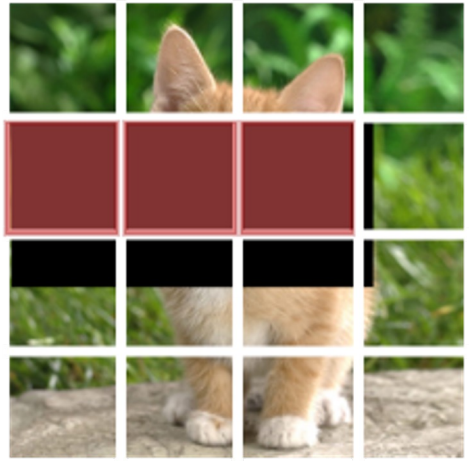
- **Limitations of CNN-based data augmentations**
  - Limited diversity between patches (**blue**)
  - Meaningless patches (**red**)

Original Image

Image-wise *Augmentation*



Region-wise *Augmentation*



Opportunities in ViT-based Data Augmentation

***Our Goal*: Boost ViT's accuracy via developing dedicated augmentation strategies**

- **Opportunities leveraged by our AugViT:**

  Differences between CNNs and ViTs

Opportunities in ViT-based Data Augmentation

***Our Goal*: Boost ViT's accuracy via developing dedicated augmentation strategies**

- **Opportunities leveraged by our AugViT:**

  Differences between CNNs and ViTs

  – *Diff 1*: Using patches as processing units

  – *Diff 2*: The adoption of self-attention modules

Opportunities in ViT-based Data Augmentation

***Our Goal*: Boost ViT's accuracy via developing dedicated augmentation strategies**

- **Opportunities leveraged by our AugViT:**

  Differences between CNNs and ViTs

  – *Diff 1*: Using patches as processing units

  ➡ Introduce patch-awareness into augmentation

  – *Diff 2*: The adoption of self-attention modules

  ➡ Guide augmentation intensities using self-attention scores

Our Contributions

- Propose AugViT, **an input-adaptive data augmentation** framework to boost ViTs' achievable task accuracy
- Integrate two enablers
  - A set of **patch-aware** augmentation techniques
  - An **attention-to-augmentation-intensity** mapper
- Consistently boost ViTs' accuracy-efficiency trade-off across **two tasks and ten representative ViT models**

Our Contributions

- **Propose AugViT, an input-adaptive data augmentation framework to boost ViTs' achievable task accuracy**

- Integrate two enablers
  - A set of **patch-aware** augmentation techniques
  - An **attention-to-augmentation-intensity** mapper

- Consistently boost ViTs' accuracy-efficiency trade-off across **two tasks and ten representative ViT models**

The Proposed AugViT Framework

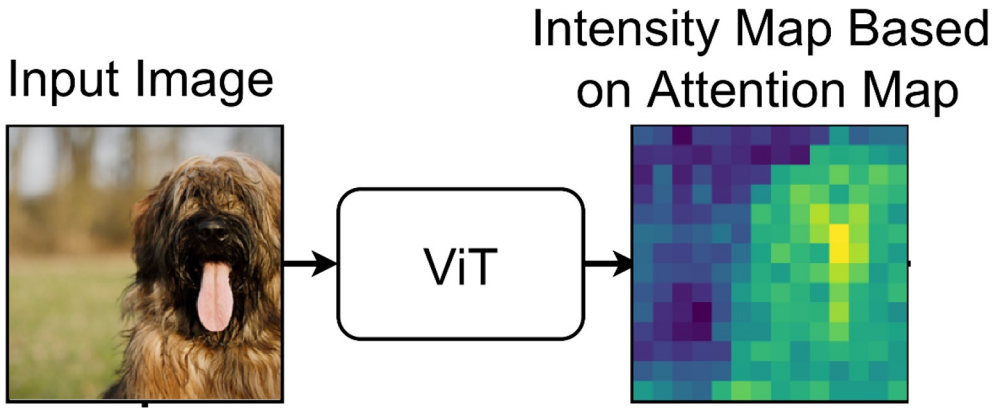
- **Core idea: Leverage attention map as an indicator to guide augmentation in each patch**

The Proposed AugViT Framework

- **Core idea: Leverage attention map as an indicator to guide augmentation in each patch**
  - Patches with **higher** attention will use **less** intense augmentation to preserve their features

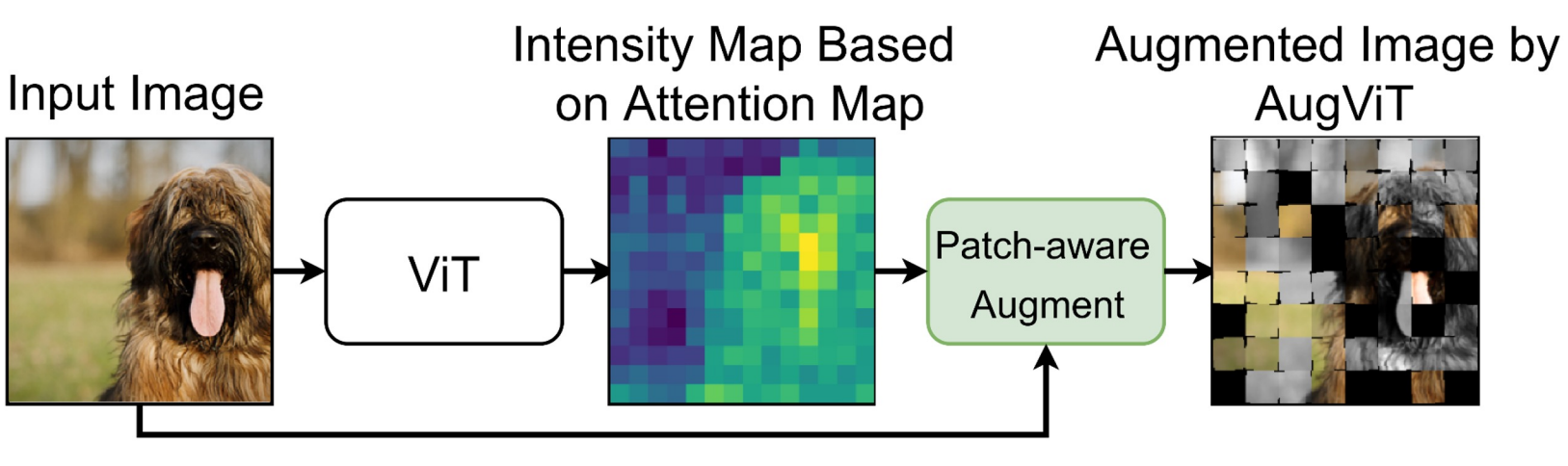
The Proposed AugViT Framework

- **Core idea: Leverage attention map as an indicator to guide augmentation in each patch**
  - Patches with **higher** attention will use **less** intense augmentation to preserve their features



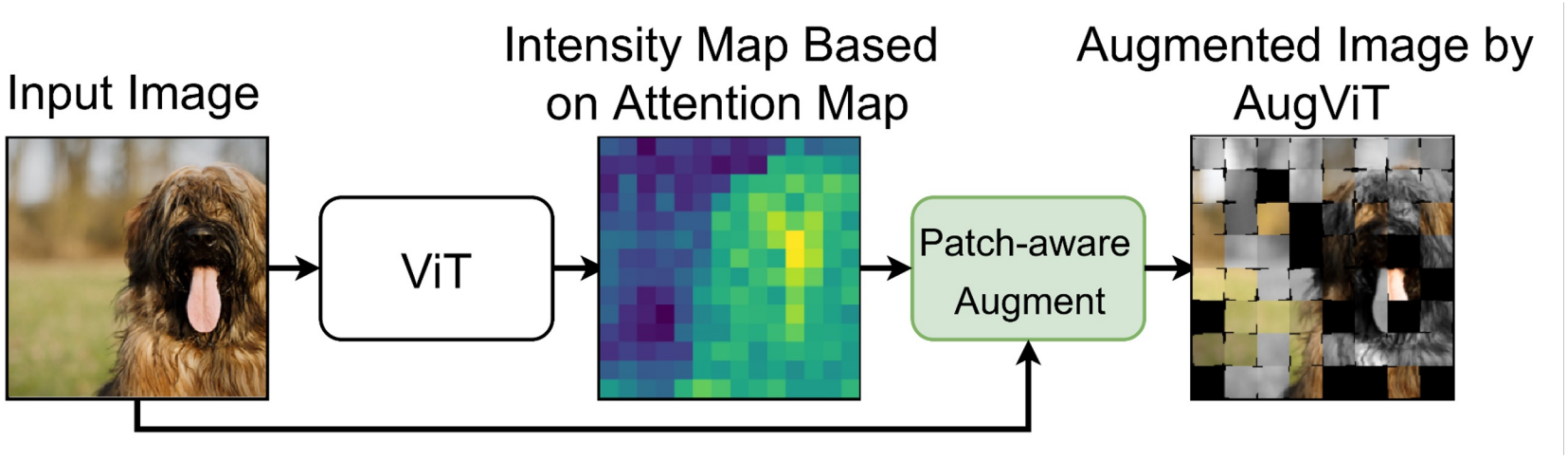
The Proposed AugViT Framework

- **Core idea: Leverage attention map as an indicator to guide augmentation in each patch**
  - Patches with **higher** attention will use **less** intense augmentation to preserve their features



The Proposed AugViT Framework

- **Core idea: Leverage attention map as an indicator to guide augmentation in each patch**
  - Patches with **higher** attention will use **less** intense augmentation to preserve their features
- **Key enablers of AugViT**
  - A set of **patch-aware** augmentation techniques
  - An **attention-to-augmentation-intensity** mapper



Our Contributions

- Propose AugViT, an **input-adaptive data augmentation** framework to boost ViTs' achievable task accuracy
- Integrate two enablers
  - A set of **patch-aware** augmentation techniques
  - An **attention-to-augmentation-intensity** mapper
- Consistently boost ViTs' accuracy-efficiency trade-off across **two tasks and ten representative ViT models**

Our Contributions

- Propose AugViT, an **input-adaptive data augmentation** framework to boost ViTs' achievable task accuracy

- **Integrate two enablers**
  - **A set of patch-aware augmentation techniques**
  - **An attention-to-augmentation-intensity mapper**

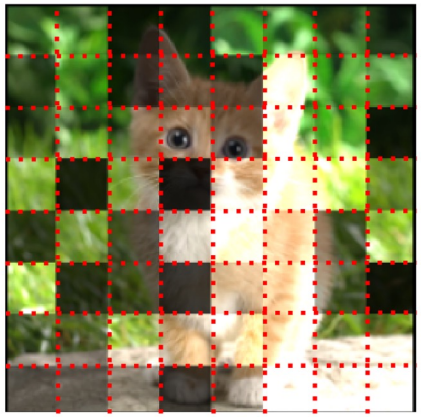
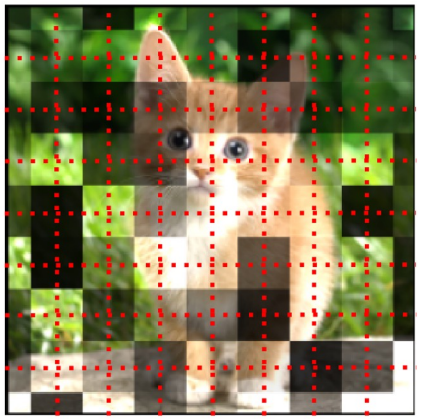
- Consistently boost ViTs' accuracy-efficiency trade-off across **two tasks and ten representative ViT models**

Does Patch-awareness Affect Training?

- Setting
  - **Vanilla**: Original color jitter
  - **SP**: **Same** augmentation within each patch
  - **DP**: **Different** augmentation within each patch

Original *Vanilla* *SP* *DP*



Importance of Patch-aware Augmentation

- Setting
  - **Vanilla**: Original color jitter
  - **SP**: **Same** augmentation within each patch
  - **DP**: **Different** augmentation within each patch

Patch-awareness matters

SP achieves 0.16%~0.28% higher accuracy than DP/vanilla

Augment	Vanilla	SP	DP
Acc. (Std)	79.81 (0.07)	79.97 (0.05)	79.69 (0.04)

DeiT-Small@ImageNet

Enabler 1: Patch-aware Augmentation

- **Challenge**: How to augment each patch properly?

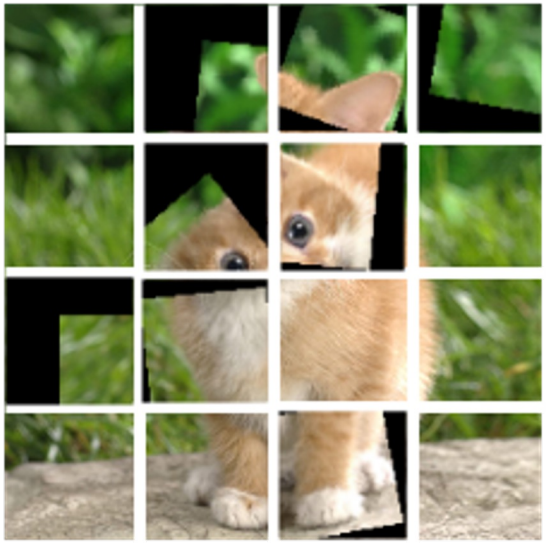
Enabler 1: Patch-aware Augmentation

- **Challenge**: How to augment each patch properly?

Original

Image-wise (Baseline)

Patch-aware (Ours)



Enabler 1: Patch-aware Augmentation

- **Challenge**: How to augment each patch properly?
- **Observations**: Patch-aware augmentation is **sensitive to spatial changes** (**blue** and **red** patches)

Original

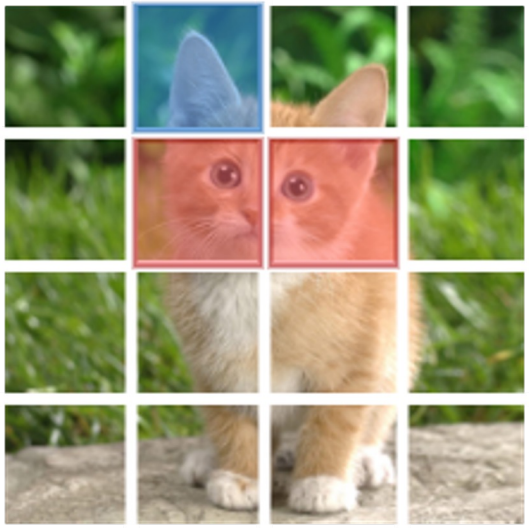
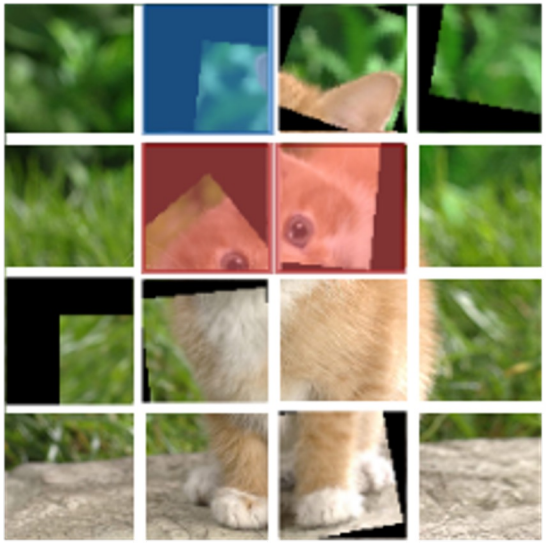


Image-wise (Baseline)

Patch-aware (Ours)



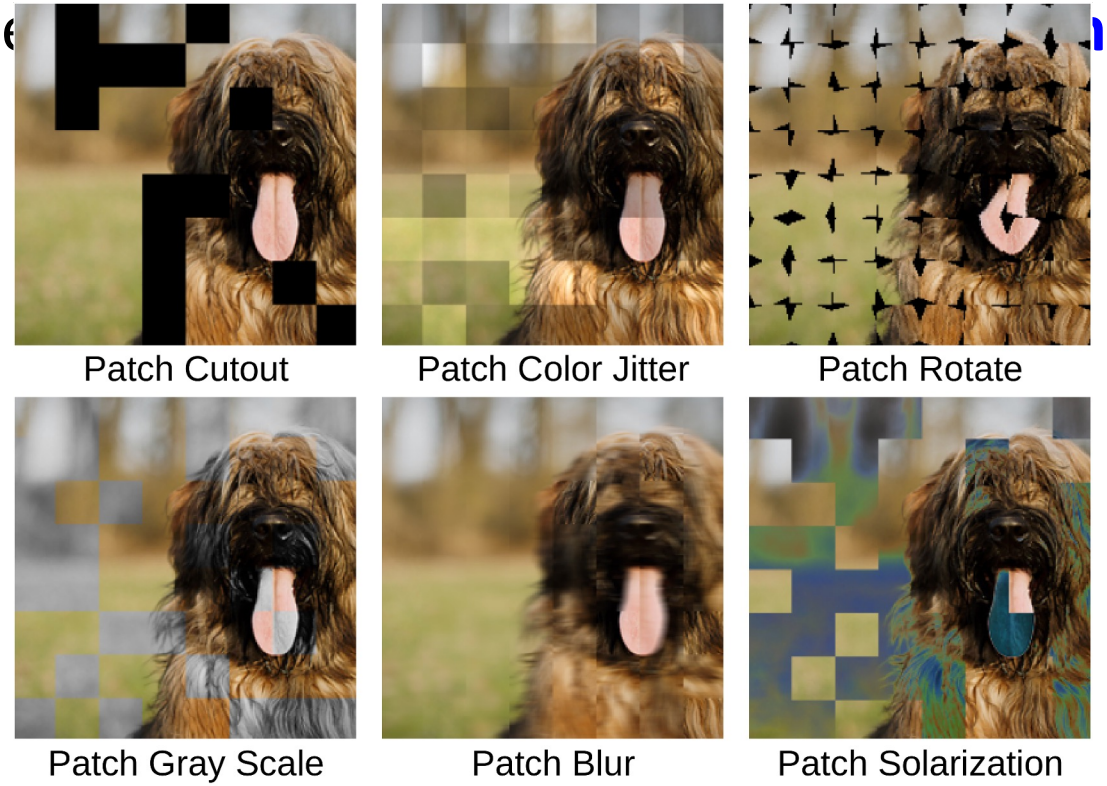
Enabler 1: Patch-aware Augmentation

- **Challenge**: How to augment each patch properly?
- **Observations**: Patch-aware augmentation is **sensitive to spatial changes** (**blue** and **red** patches)
- **Our answer**: Avoid drastic changes in spatial information
  - Preserve **key features** in the patch (**blue**)
  - Maintain inter-patch **relationships** (**red**)

Original *Image-aware (Baseline)* *Patch-aware (Ours)*

Enabler 1: Patch-aware Augmentation

- **Our answer**: Avoid drastic changes in spatial information
  - Preserve **key features** in the patch
  - Maintain Inter-patch **relationships**
- Propose **niques**



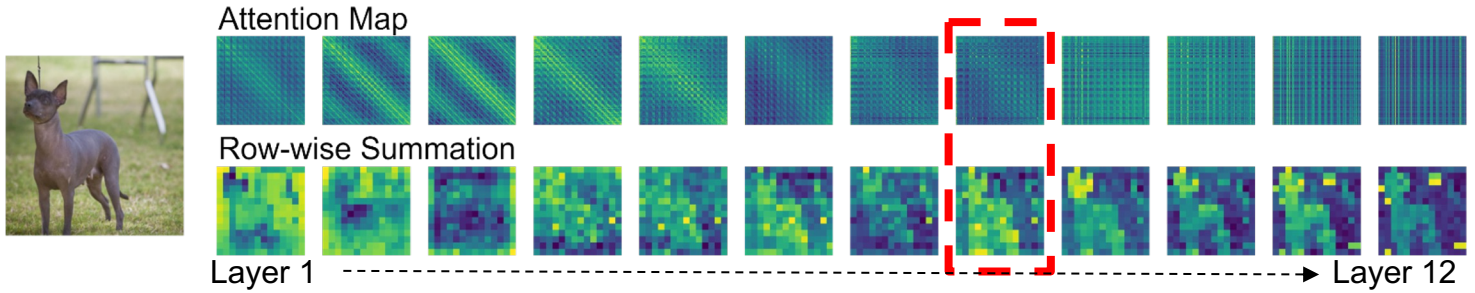
Enabler 2: Attention to Intensity Mapper

- **Challenge**: How to guide augmentation intensity with attention map?

Enabler 2: Attention to Intensity Mapper

- **Challenge**: How to guide augmentation intensity with attention map?
- **Our solution**: a three-stage pipeline
  - Select a representative attention map: **Enabler 2-1**
  - Preserve high-attention patches: **Enabler 2-2**
  - Optimize attention to intensity mapping: **Enabler 2-3**

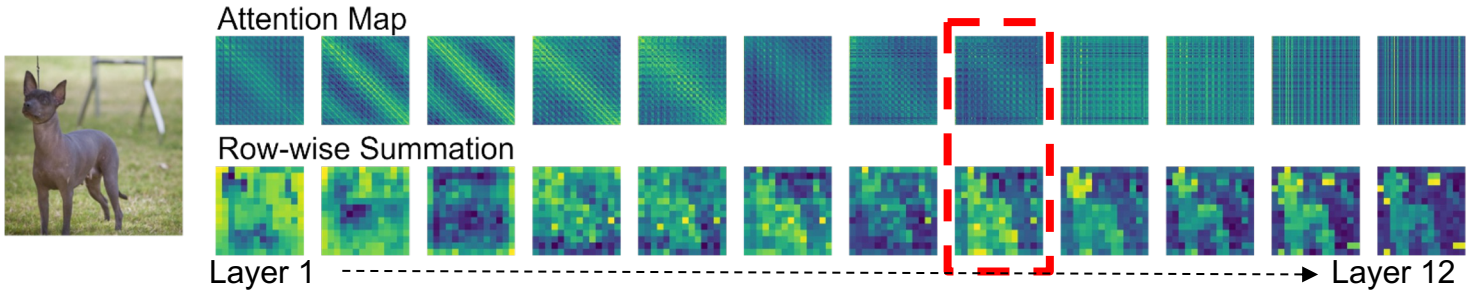
Enabler 2-1: Analysis on Attention Map



Visualization of DeiT-Small's attention maps at different layers

Enabler 2-1: Analysis on Attention Map

- **Certain attention maps** can better **identify the object** in the image

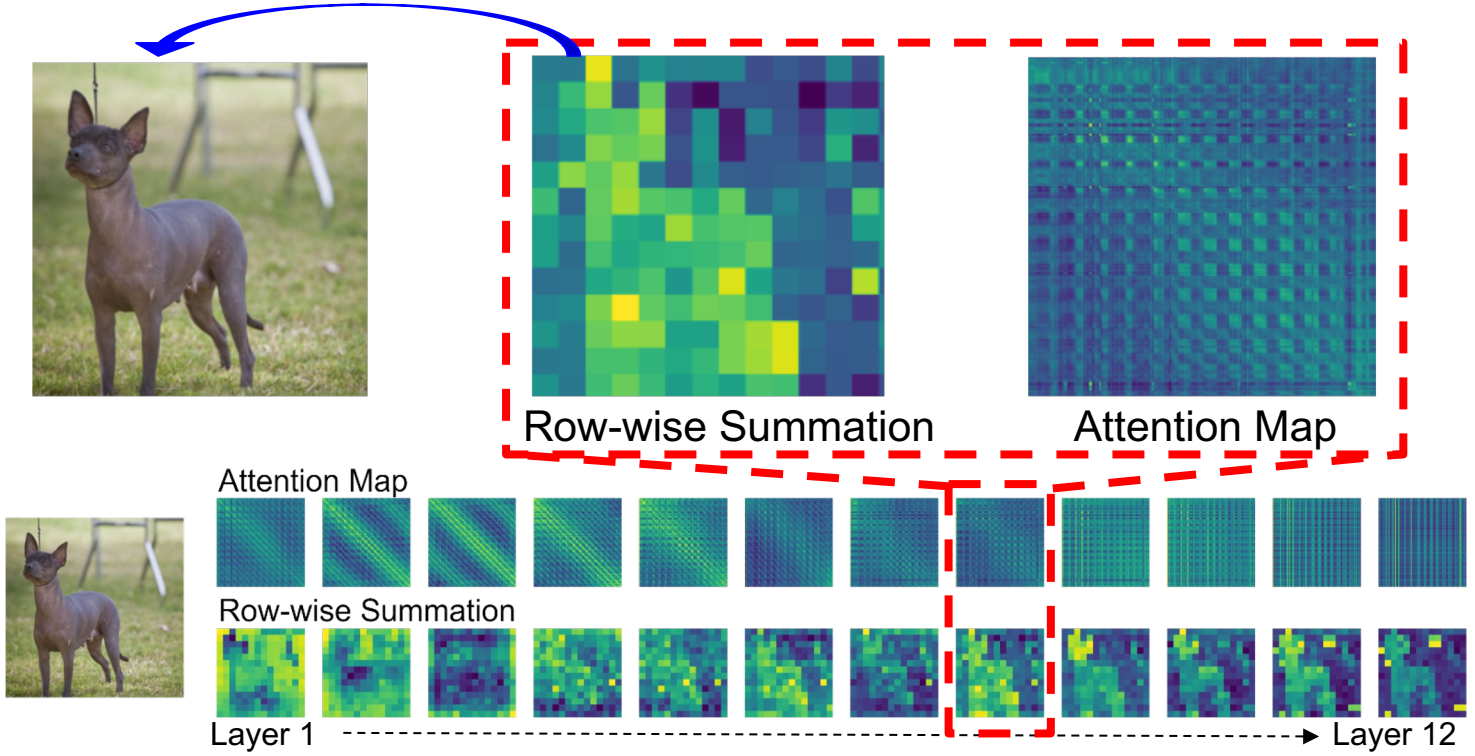


Visualization of DeiT-Small's attention maps at different layers

Enabler 2-1: Analysis on Attention Map

- **Certain attention maps** can better **identify the object** in the image

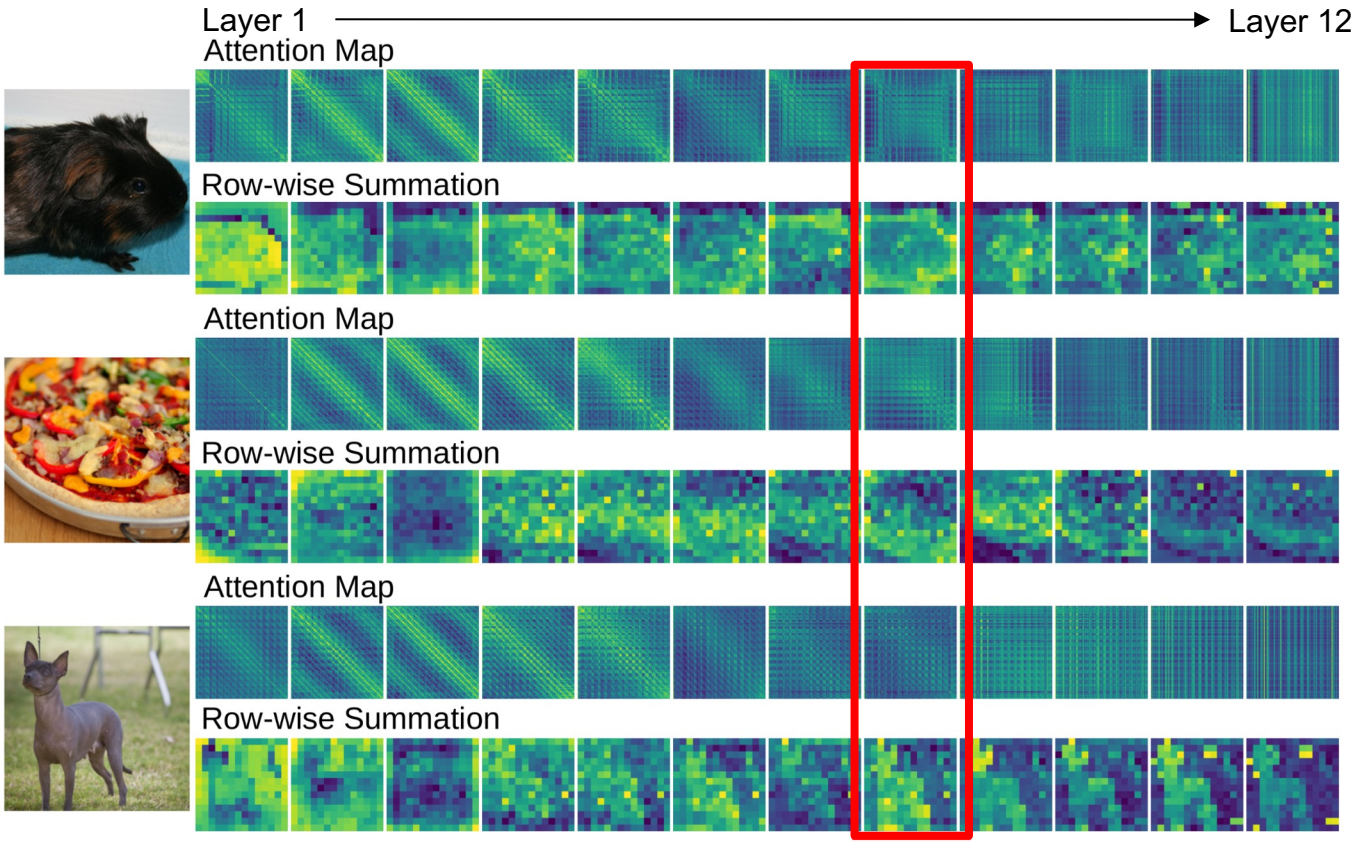
High value patches correlates well with the **location of the object**



Visualization of DeiT-Small's attention maps at different layers

Enabler 2-1: Analysis on Attention Map

- Given different inputs, such attention maps appear **at same layer** in the same ViT model

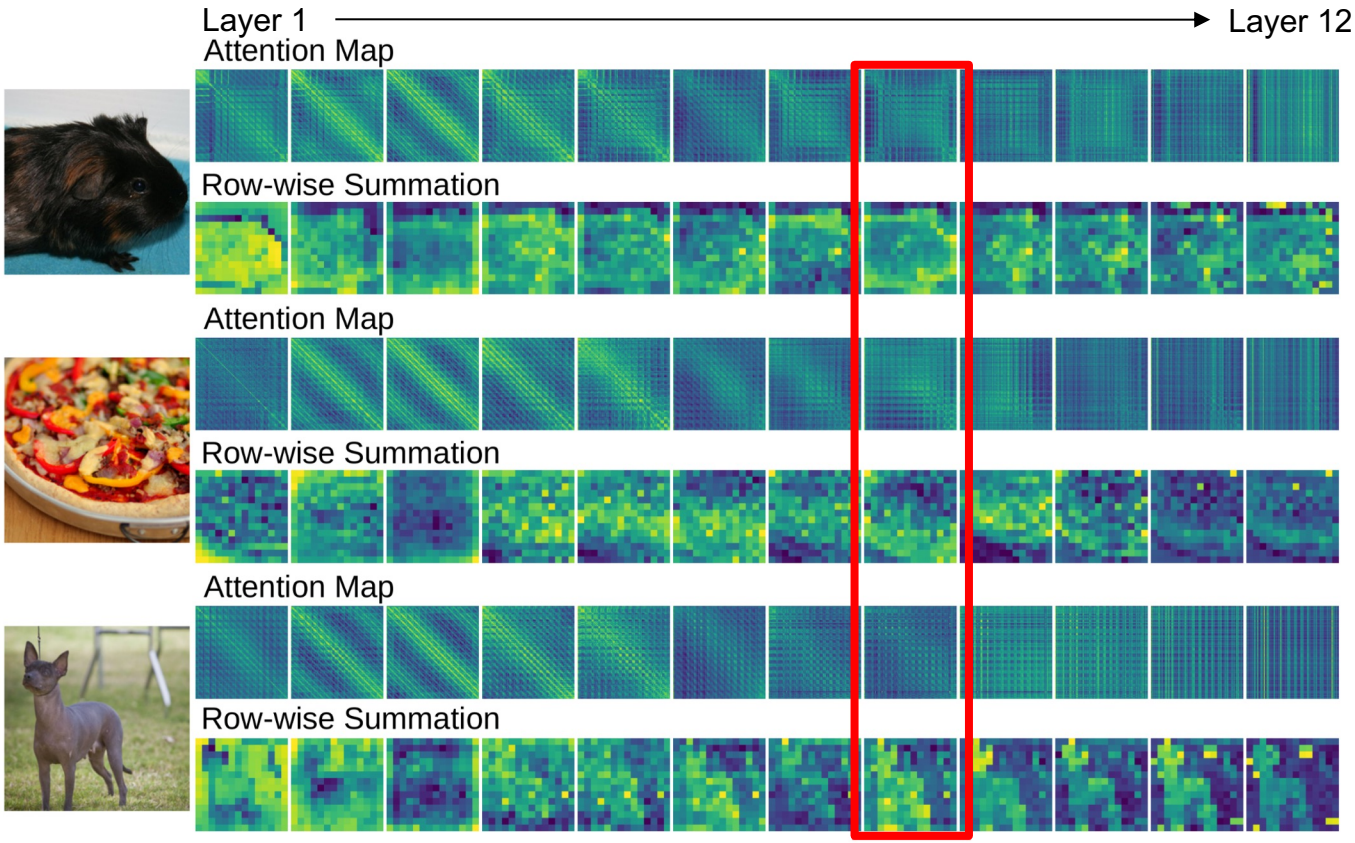


Visualization of DeiT-Small's attention maps at different layers

Enabler 2-1: Analysis on Attention Map

- Given different inputs, such attention maps appear **at same layer** in the same ViT model

➡ We use this attention map in AugViT



Visualization of DeiT-Small's attention maps at different layers

Enabler 2-2: Insights on Augmentation Intensity

- Setting
  - **Vanilla**: **Image-wise** augmentation in *[H. Touvron, ICML'21]*

Enabler 2-2: Insights on Augmentation Intensity

- Setting
  - **Vanilla**: **Image-wise** augmentation in *[H. Touvron, ICML'21]*
  - **Uniform**: Augment each patch with **random** intensity

Enabler 2-2: Insights on Augmentation Intensity

- Setting
  - **Vanilla**: **Image-wise** augmentation in *[H. Touvron, ICML'21]*
  - **Uniform**: Augment each patch with **random** intensity
  - **Same**: **Higher** attention –> **higher** aug. intensity

Enabler 2-2: Insights on Augmentation Intensity

- Setting
  - **Vanilla**: **Image-wise** augmentation in *[H. Touvron, ICML'21]*
  - **Uniform**: Augment each patch with **random** intensity
  - **Same**: **Higher** attention –> **higher** aug. intensity
  - **Inverse**: **Higher** attention –> **lower** aug. intensity

Enabler 2-2: Insights on Augmentation Intensity

- Setting
  - **Vanilla**: **Image-wise** augmentation in *[H. Touvron, ICML'21]*
  - **Uniform**: Augment each patch with **random** intensity
  - **Same**: **Higher** attention –> **higher** aug. intensity
  - **Inverse**: **Higher** attention –> **lower** aug. intensity

Mapping	Vanilla	Uniform	Same	Inverse
Acc (%)	79.8	80.1	80.1	80.3

DeiT-Small@ImageNet

Enabler 2-2: Insights on Augmentation Intensity

- Setting
  - **Vanilla**: **Image-wise** augmentation in *[H. Touvron, ICML'21]*
  - **Uniform**: Augment each patch with **random** intensity
  - **Same**: **Higher** attention –> **higher** aug. intensity

High-attention patches should be preserved

Inverse achieves 0.2%~0.5% higher accuracy than baselines

Mapping	Vanilla	Uniform	Same	Inverse
Acc (%)	79.8	80.1	80.1	80.3

DeiT-Small@ImageNet

Enabler 2-3: Attention to Intensity Mapping Function

- Gumbel softmax mapping function

$$s_i = \frac{\exp\left[(log(\alpha_i) + g_i)/\tau\right]}{\sum_i \exp\left[(log(\alpha_i) + g_i)/\tau\right]},$$

Enabler 2-3: Attention to Intensity Mapping Function

- Gumbel softmax mapping function

$$s_i = \frac{\exp\left[\left(log(\alpha_i) + g_i\right)/\tau\right]}{\sum_i \exp\left[\left(log(\alpha_i) + g_i\right)/\tau\right]},$$

Random variable

Temperature

Enabler 2-3: Attention to Intensity Mapping Function

- Gumbel softmax mapping function

$$s_i = \frac{\exp\left[(log(\alpha_i) + g_i)/\tau\right]}{\sum_i \exp\left[(log(\alpha_i) + g_i)/\tau\right]},$$

- Setting
  - **GS**: α_i is the row-wise sum in attention map
  - **Inv-GS**: α_i is the **reciprocal** of row-wise sum in attention map

Enabler 2-3: Attention to Intensity Mapping Function

- Gumbel softmax mapping function

$$s_i = \frac{\exp\left[(log(\alpha_i) + g_i)/\tau\right]}{\sum_i \exp\left[(log(\alpha_i) + g_i)/\tau\right]},$$

- Setting
  - **GS**: α_i is the row-wise sum in attention map
  - **Inv-GS**: α_i is the **reciprocal** of row-wise sum in attention map

Inv-GS maps the attention to aug. intensity better

Inv-GS achieves 0.6%~0.1% higher accuracy than baselines

Mapping	Vanilla	Uniform	Same	Inverse	GS	**Inv-GS**
Acc (%)	79.8	80.1	80.1	80.3	79.8	**80.4**

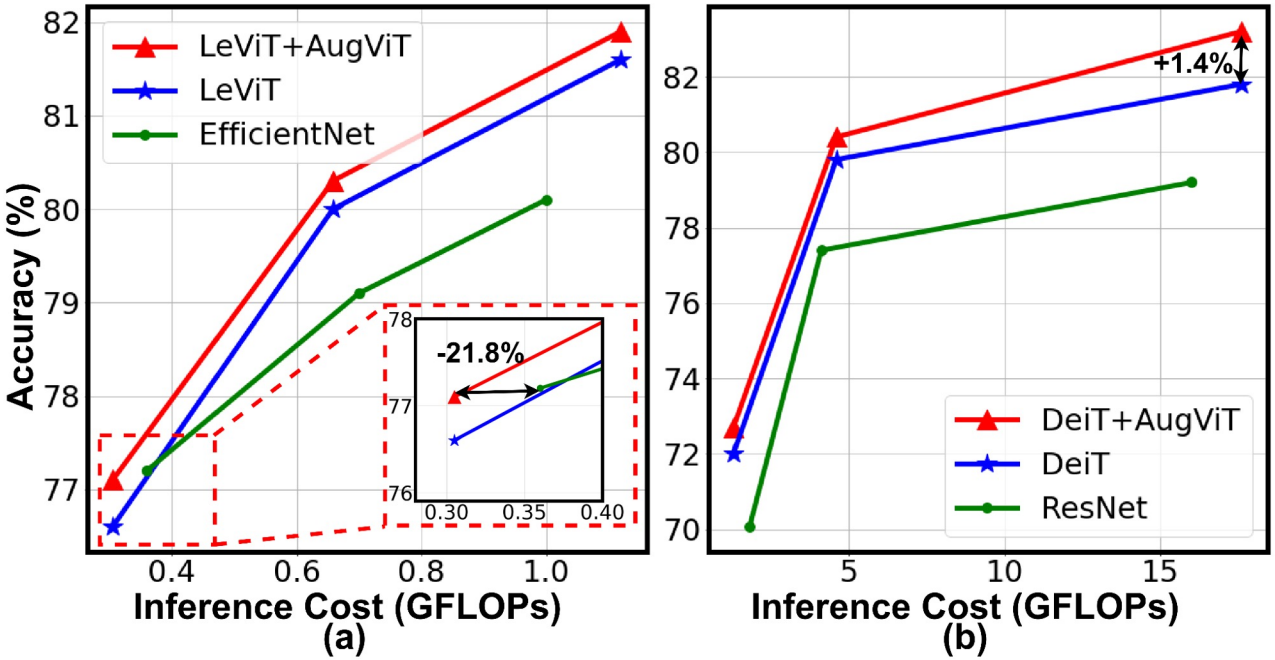
DeiT-Small@ImageNet

Evaluation Settings

- **Ten models** on **two tasks**
  - **Image classification** on ImageNet *[J. Deng, CVPR'09]*
    - Five variants of LeViT *[H. Graham, ICCV'21]*
    - Three variants of DeiT *[H. Touvron, ICML'21]*
    - Swin-Tiny *[Z. Liu, ICML'21]* and PVT-Small *[W. Wang, ICCV'21]*
  - **Object detection** on COCO *[T. Lin, ECCV'14]*
    - DeiT-Small *[H. Touvron, ICML'21]* and Swin-Tiny *[Z. Liu, ICML'21]*
- **Two SOTA ViT dedicated data augmentation baselines**
  - DeiT *[H. Touvron, ICML'21]*
  - DeiT-III *[H. Touvron, ECCV'22]*

AugViT Boosts ViTs' Accuracies

- **AugViT boosts ViTs' accuracies on ImageNet classification**
  - **+0.3% ~ 1.4% accuracy** across different variants of **DeiT**
  - **+0.3% ~ 0.6% accuracy** across different variants of **LeViT**
  - **+0.5%** on **LeViT-128S,** achieving a comparable accuracy with EfficientNet-B0 while **saving 21.8% FLOPs**



AugViT is Effective Across Various Tasks

- AugViT on average boosts DeiT-Small's accuracy on **object detection**@COCO by a **0.2% higher AP**

Model	Augmentation	AP^{box}	AP^{box}_{50}	AP^{box}_{75}
DeiT-Small	DeiT [33]	48.0	67.2	51.7
	AugViT	**48.3**	**67.4**	**51.8**

AugViT is Effective Across Various Tasks

- AugViT on average boosts DeiT-Small's accuracy on **object detection**@COCO by a **0.2% higher AP**

Model	Augmentation	AP^{box}	AP^{box}_{50}	AP^{box}_{75}
DeiT-Small	DeiT [33]	48.0	67.2	51.7
	AugViT	**48.3**	**67.4**	**51.8**

Please refer to our paper for more results !

Copyright Notice

www.tinyml.org