EXPLORING DIVERSITY IN NEURAL ARCHITECTURES FOR SAFETY

MICHAŁ FILIPIUK, VASU SINGH

JULY 25TH, AISAFETY 2022

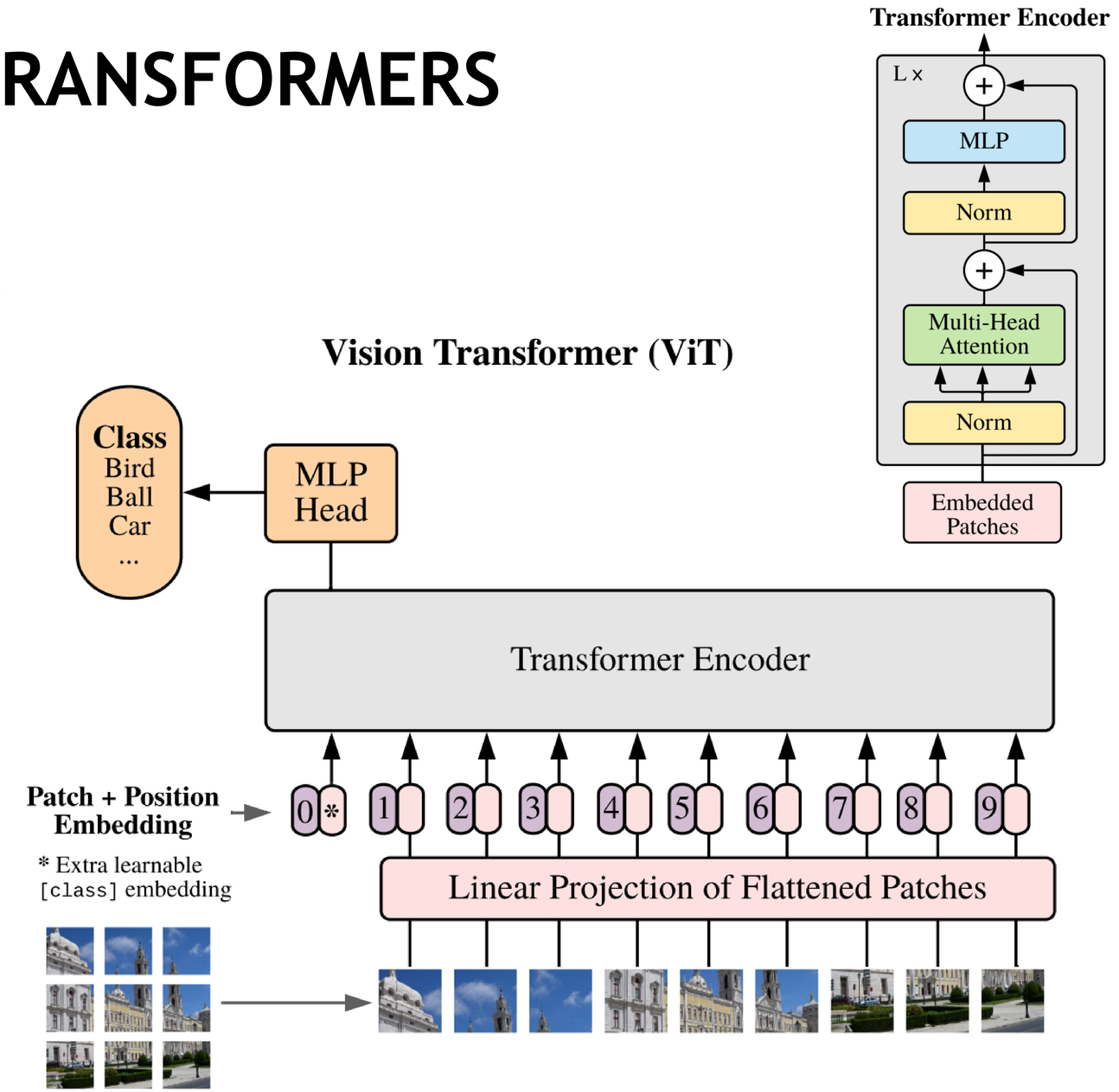
INTRODUCTION

- New architectures for computer vision problems like Vision Transformers and MLP-Mixers emerged, challenging CNNs position
- They are proven to perform differently to CNNs (Bhojanapalli et al. 2021, Raghu et al. 2021), but how quantitatively this diversity in computation translates to CV problems in unknown
- We investigate how differently are various architectures performing on image classification using a few diversity metrics and how we can benefit from this diversity by ensembles

NVIDIA.

VISION TRANSFORMERS

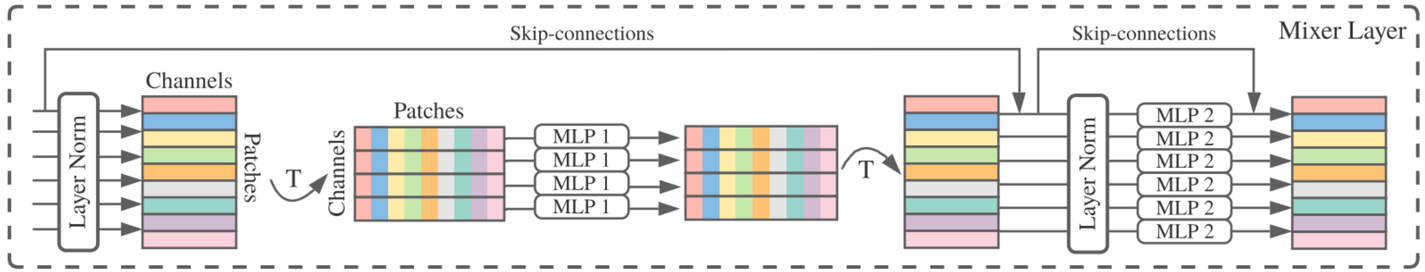
- October 2020: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale published
- Pros:
  - No restriction on input size
  - Global perception field from the first layer
- Cons:
  - Needs more data to learn
  - Quadratic complexity wrt number of patches
  - Lack of translation equivariance
  - No locality of perception

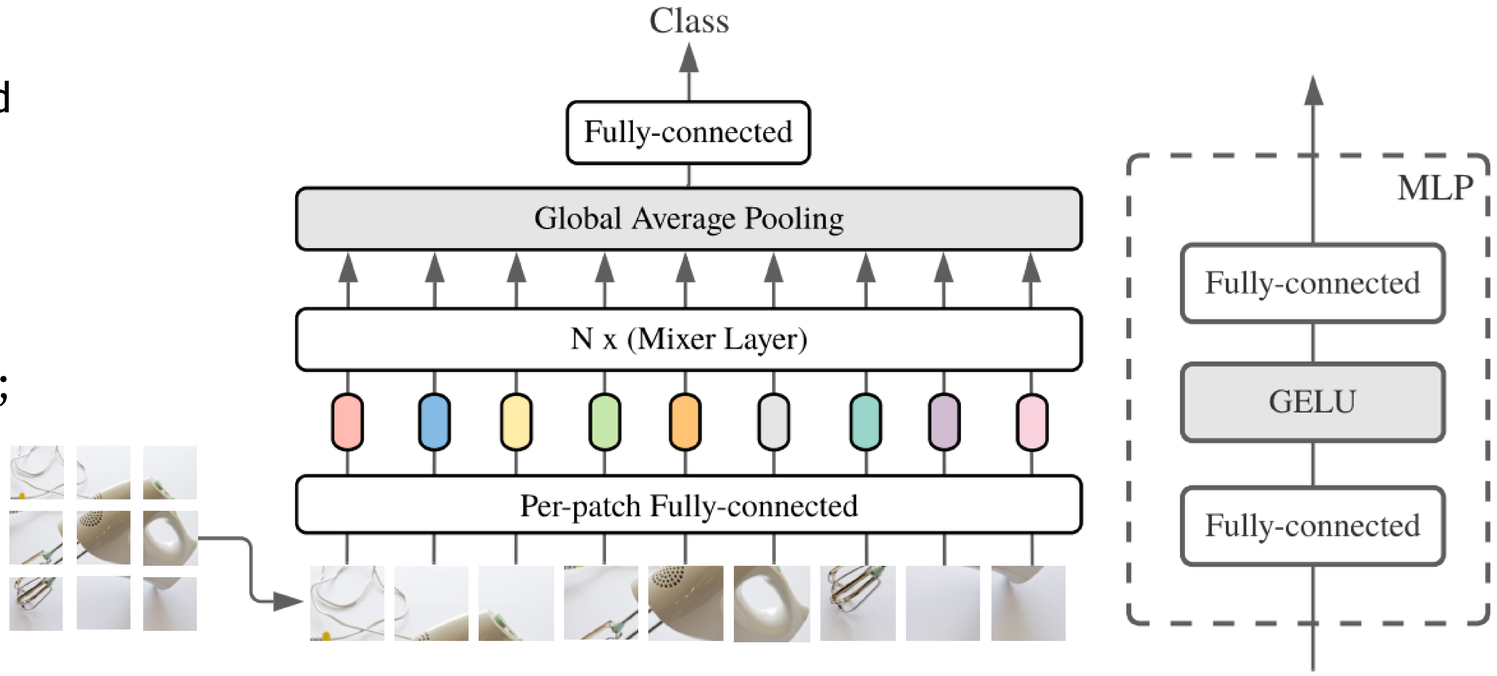


Source: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

MLP MIXERS

- May 2021: MLP-Mixer: An all-MLP Architecture for Vision published
- Two ways of mixing:
  - Mixing across patches, within channels
  - Mixing within patches, across channels
- Pros:
  - Competitive performance to CNNs and ViTs
  - Global perception field from start
  - Linear complexity wrt number of patches
- Cons:
  - Needs more even more data to learn; otherwise falls behind CNNs and ViTs
  - Fixed image input size
  - Lack of translation equivariance
  - Little interest in community





Source: MLP-Mixer: An all-MLP Architecture for Vision

DIVERSITY METRICS FOR CLASSIFICATION

Error Consistency (Geirhos et al. 2020, lower value, more diverse model combination):

$$\kappa = \frac{c_{obs} - c_{exp}}{1 - c_{exp}}$$

c_{obs} – observed error overlap (rate of images classified correctly/incorrectly by both models)

$c_{exp} = acc_{model1} acc_{model2} + (1 - acc_{model1})(1 - acc_{model2})$ – expected error overlap

Diversity metrics (Ortega et al. 2021, higher values, more diverse model combinations):

0/1-loss diversity: $\mathbb{E}_v\left[\mathbb{V}_\varrho\left(\mathbb{1}(f(x;\theta) \neq y)\right)\right]$

Cross-Entropy-loss diversity: $\mathbb{E}_v\left[\mathbb{V}_\varrho\left(\frac{p(y|x,\theta)}{\sqrt{2} max_\theta p(y|x,\theta)}\right)\right]$

$\mathbb{E}_v$ – expected value over the dataset

$\mathbb{V}_\varrho$ - variance over the ensemble models

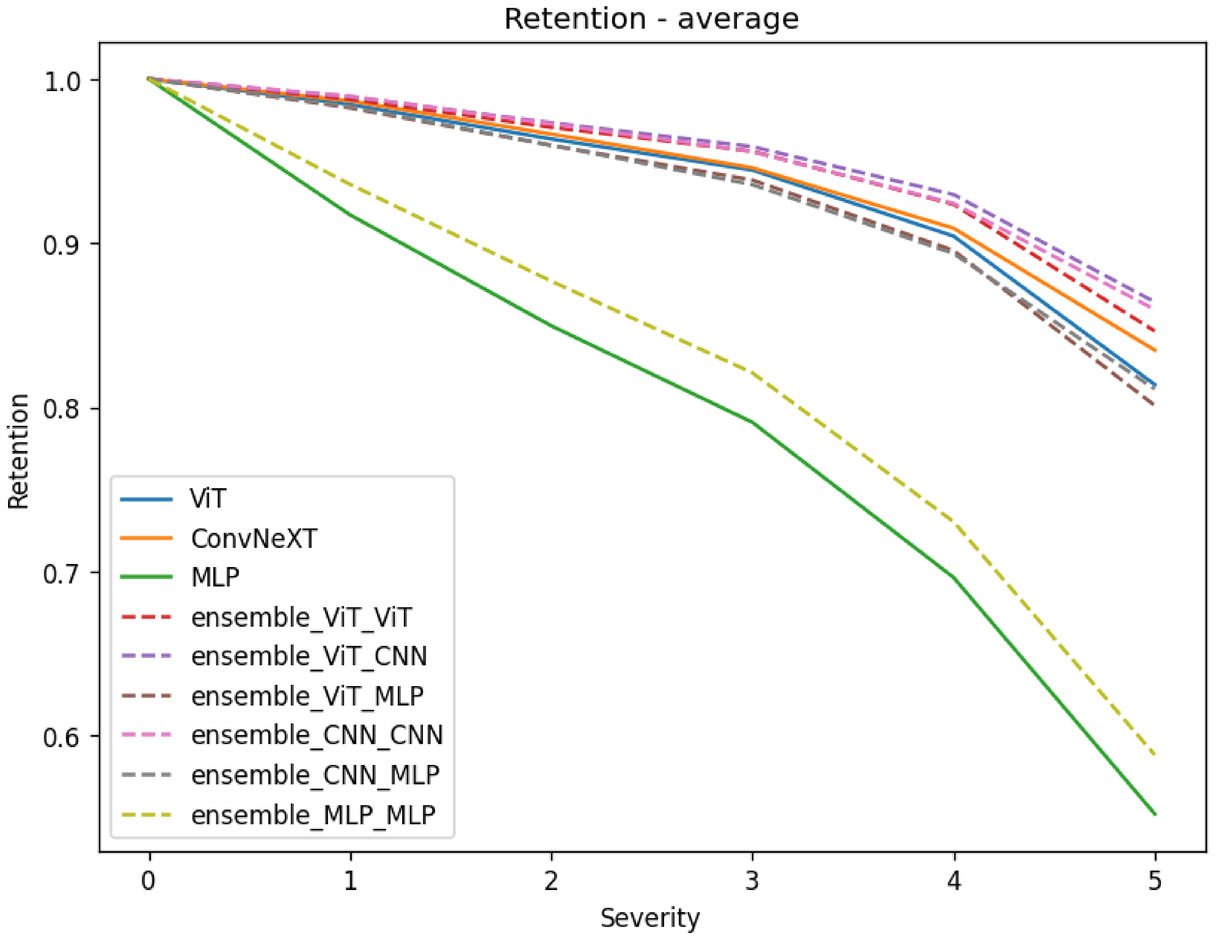
0/1-diversity counts rate of samples, where one model classifies correctly while the other doesn't.
CE-diversity is an average over the dataset of variances of predicted probabilities of ensemble's models.

EXPERIMENTAL SETUP

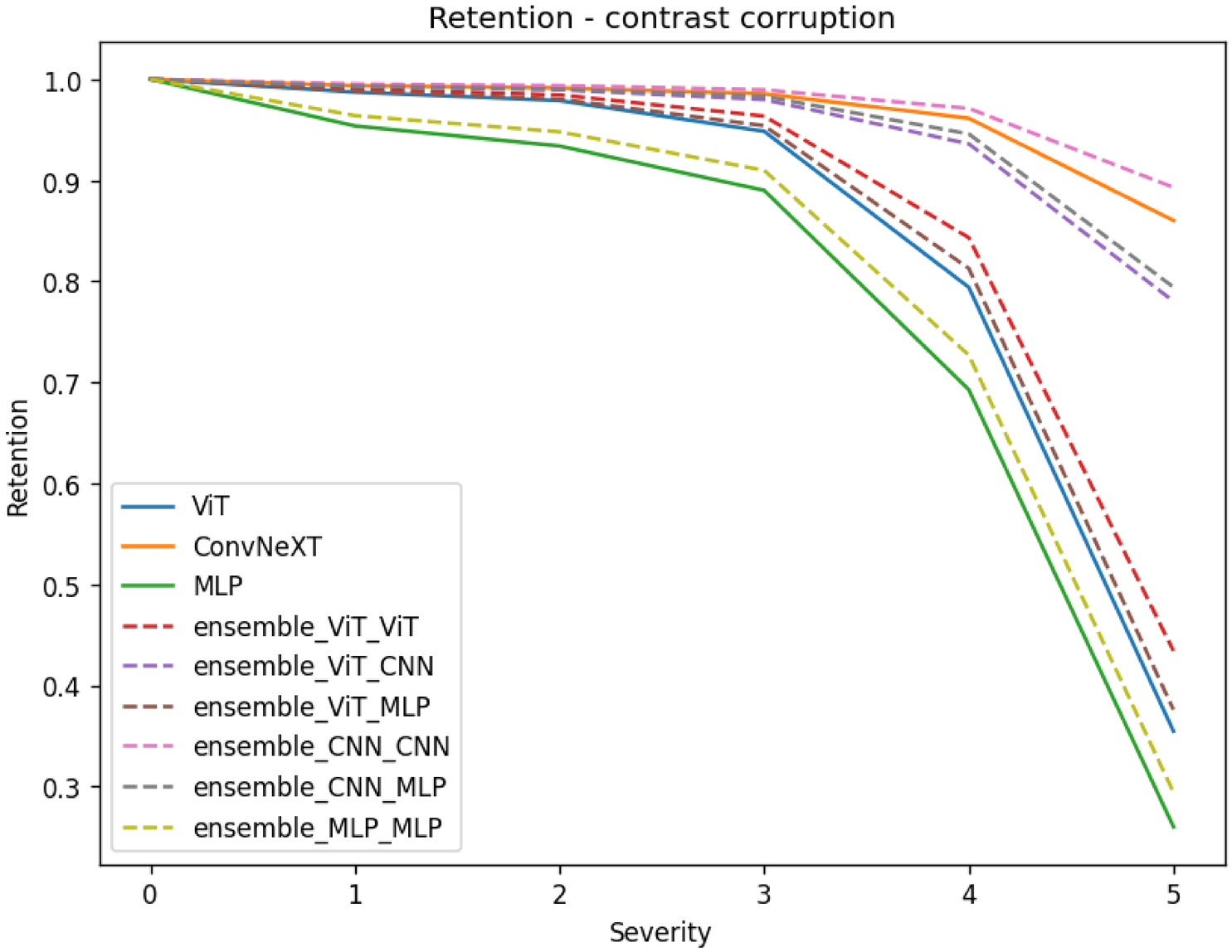
- 6 Models:
  - ConvNeXt-Base: 89M params
  - ConvNeXt-XL: 350M params
  - ViT-Base/8: 86M params
  - ViT-Large/16: 307M params
  - MLP-Base/16: 59M params
  - MLP-Large/16: 207M params
- 30 ensembles (created by averaging softmax outputs)
- Model input: 224x224px images
- All models were pretrained on ImageNet-21K and finetuned to ImageNet-1K (they differ in data augmentation)
- Data used: ImageNet-1k validation set and ImageNet-C
- Metrics:
  - Top10 accuracy
  - Retention (accuracy at corrupted input divided by original accuracy)
  - Diversity metrics
  - Error consistency
  - 0-1 Diversity components

RETENTION COMPARISON OVER DIFFERENT CORRUPTIONS

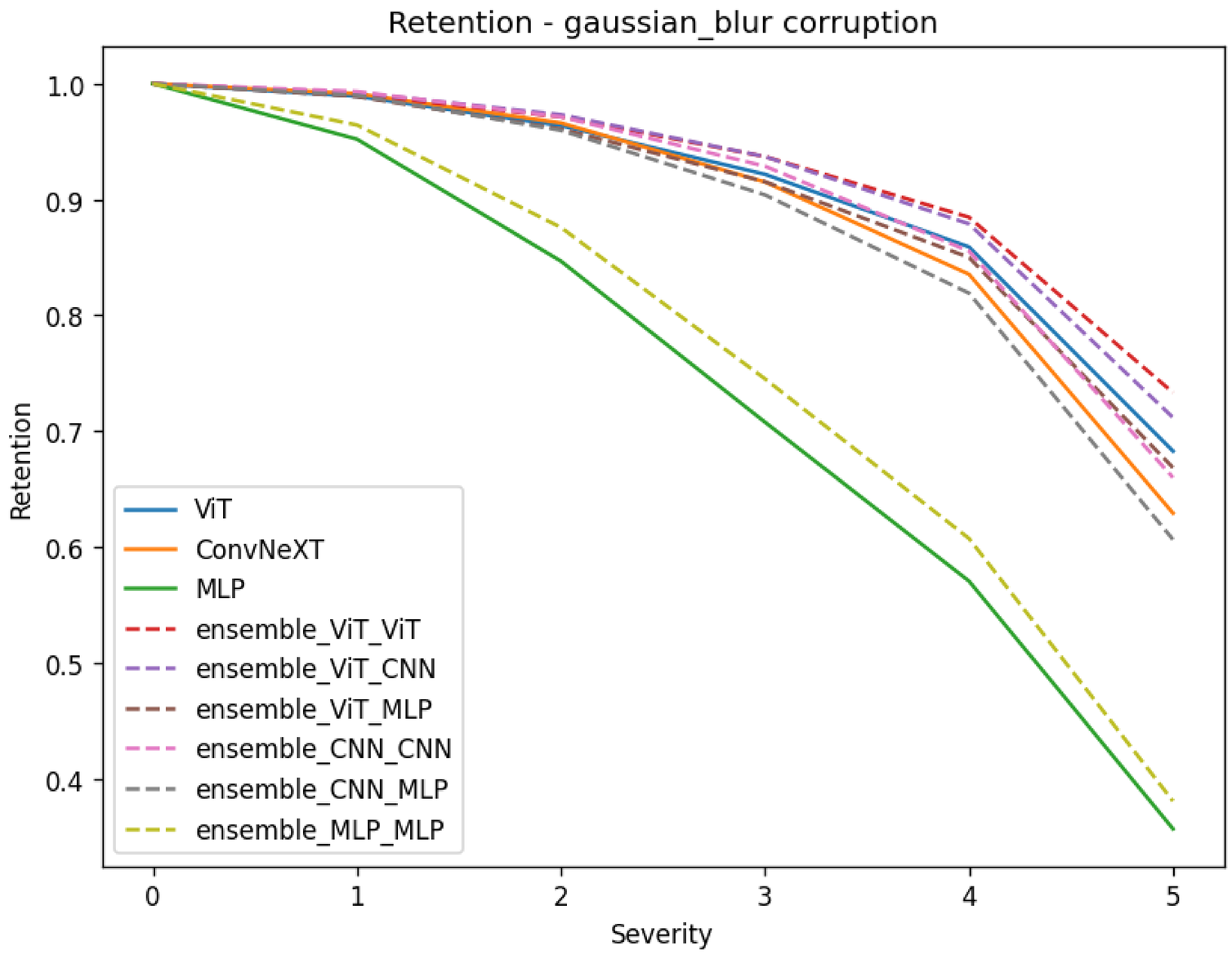
- CNN-ViT ensemble performs well for every corruption – even when one of its components performs poorly
- Publicly available MLP-Mixers are much less robust to IM-C corruptions than ViTs and CNNs



RETENTION COMPARISON OVER CONTRAST CORRUPTION



RETENTION COMPARISON OVER GAUSSIAN BLUR CORRUPTION



OBSERVATIONS ON DIVERSITY AND ROBUSTNESS

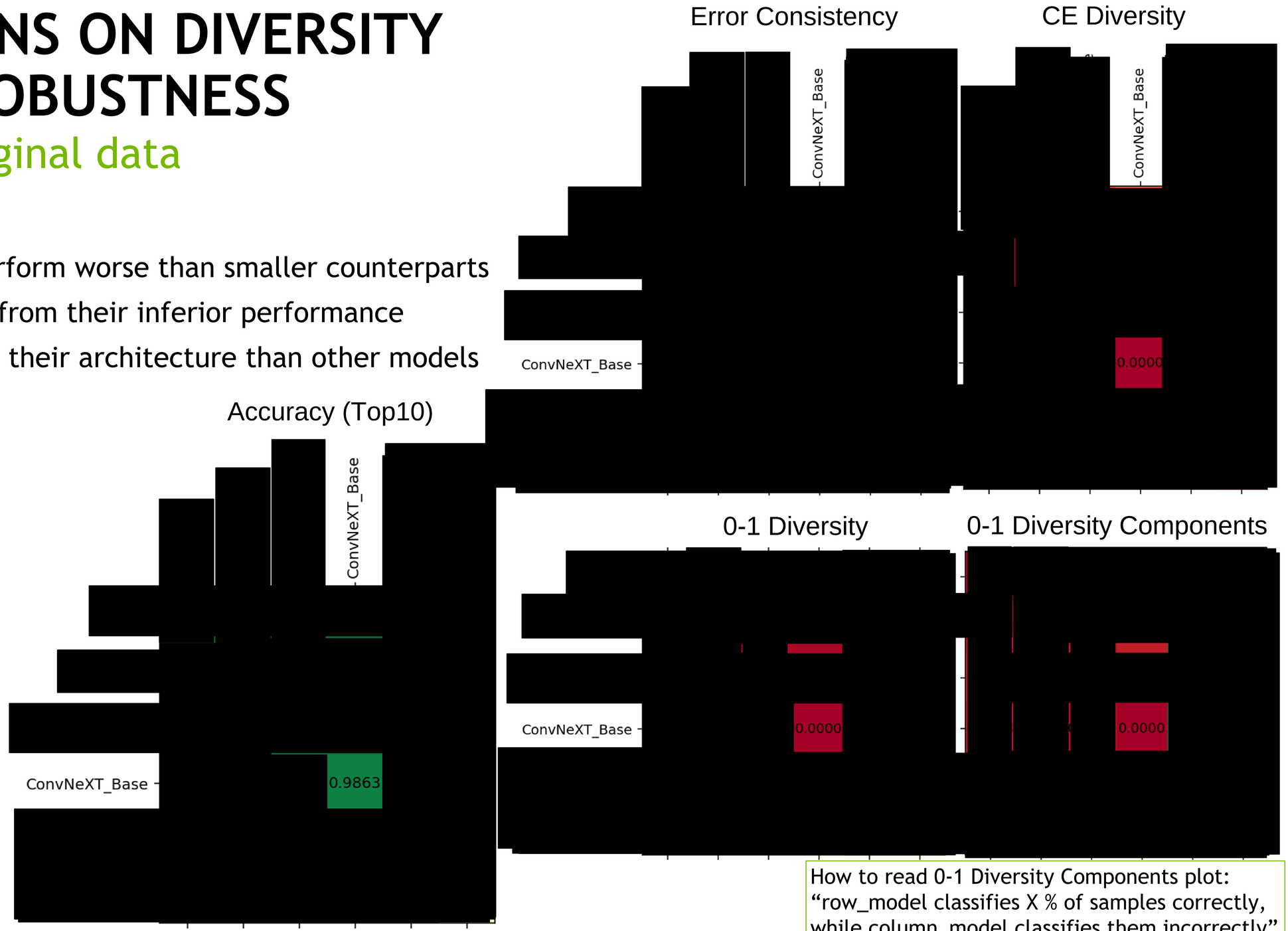
Original data

Key takeaways:

- Large ViT and MLP-Mixer perform worse than smaller counterparts
- MLP-Mixers diversity comes from their inferior performance
- CNNs are less diverse within their architecture than other models

Interesting ensembles comparison:

- ConvNeXt-XL + ConvNeXt-B vs. ViT-B + ConvNeXt-B
- ConvNeXt-XL + ConvNeXt-B vs. ViT-L + ConvNeXt-B
- ViT-L + ViT-B vs. ConvNeXt-B + ViT-B



How to read 0-1 Diversity Components plot: “row_model classifies X % of samples correctly, while column_model classifies them incorrectly”

OBSERVATIONS ON DIVERSITY AND ROBUSTNESS

Gaussian blur 5

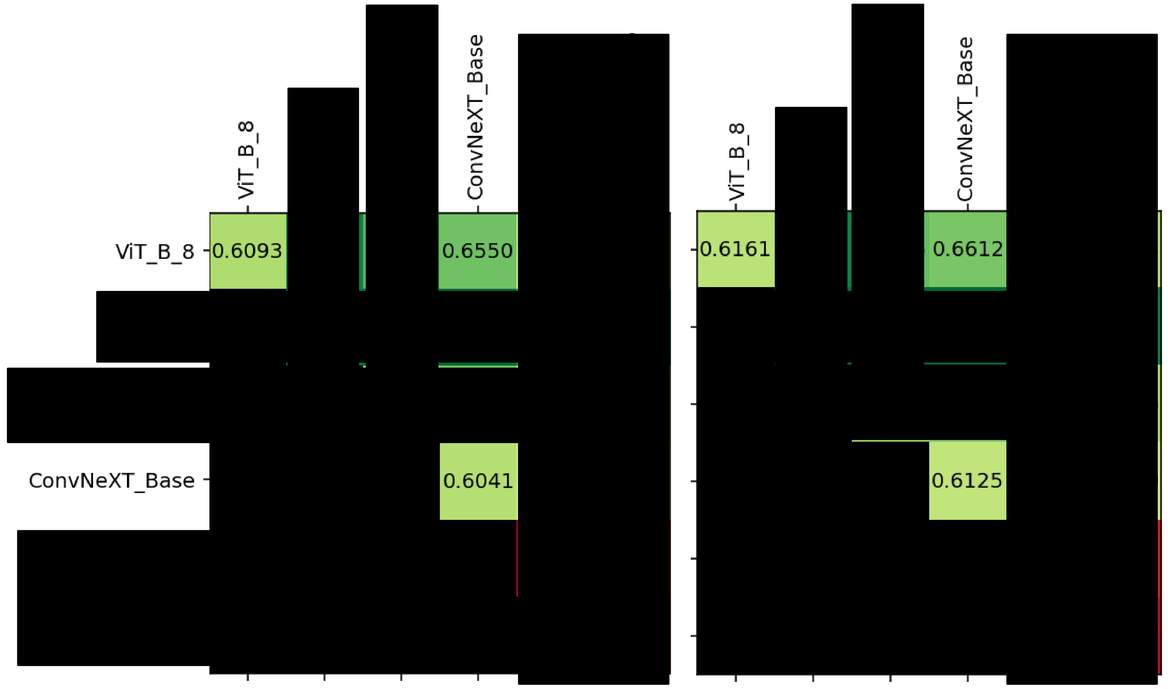
Key takeaways:

- Accuracy drop isn't significant when ensembling with poorly performing model
- 0-1 Diversity Components for ViT-B and CNNs are quite symmetric, explaining a big accuracy boost

Interesting ensembles comparison:

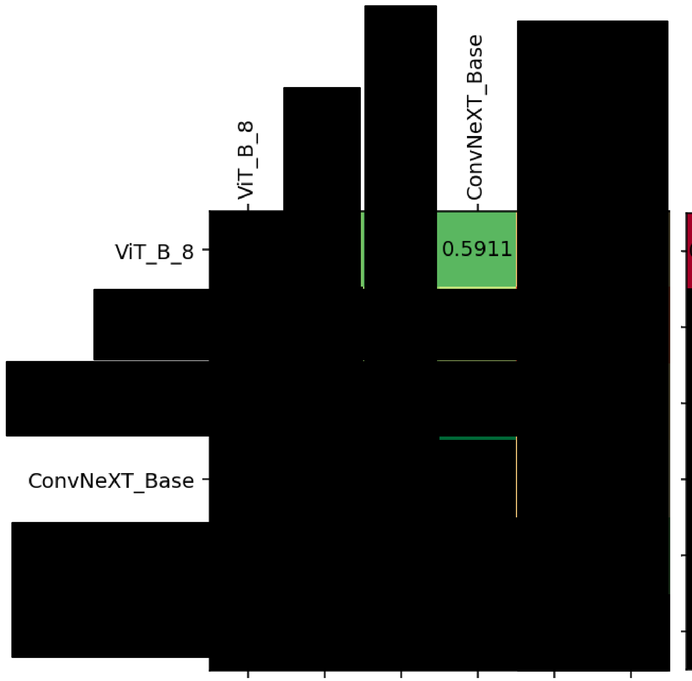
- CNN-XL+CNN-B vs. ViT-B+CNN-B
- CNN-XL+CNN-B vs. CNN-XL+ViT-B
- ViT-L+ViT-B vs. ViT-L +CNN-B

Accuracy (Top10)

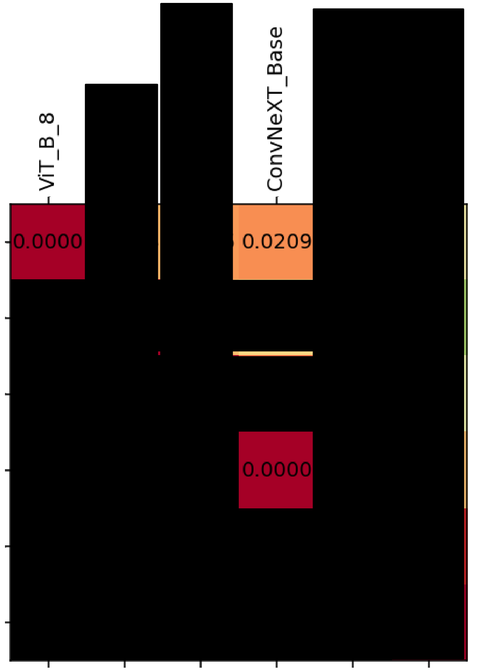


Retention (Top10)

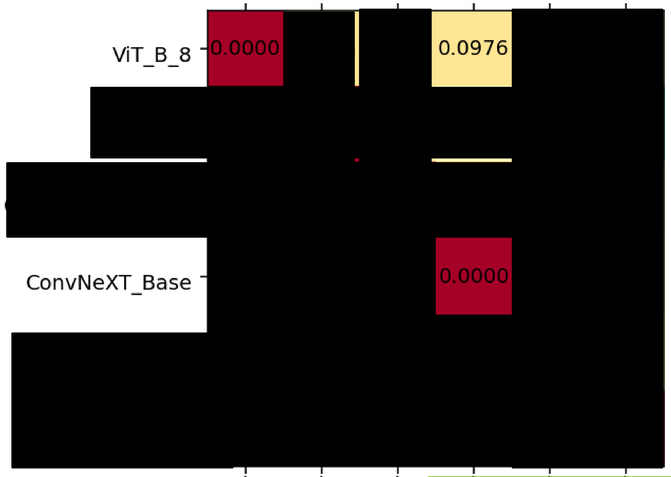
Error Consistency



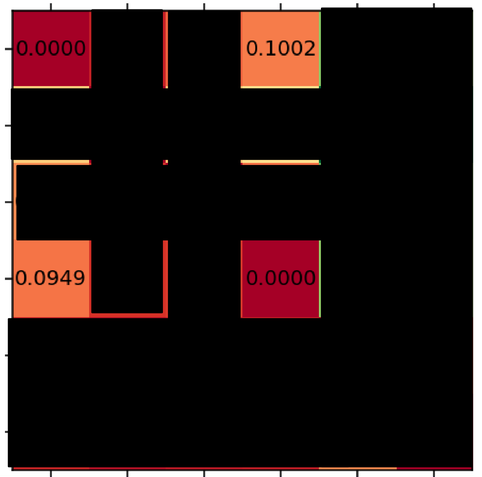
CE Diversity



0-1 Diversity



0-1 Diversity Components



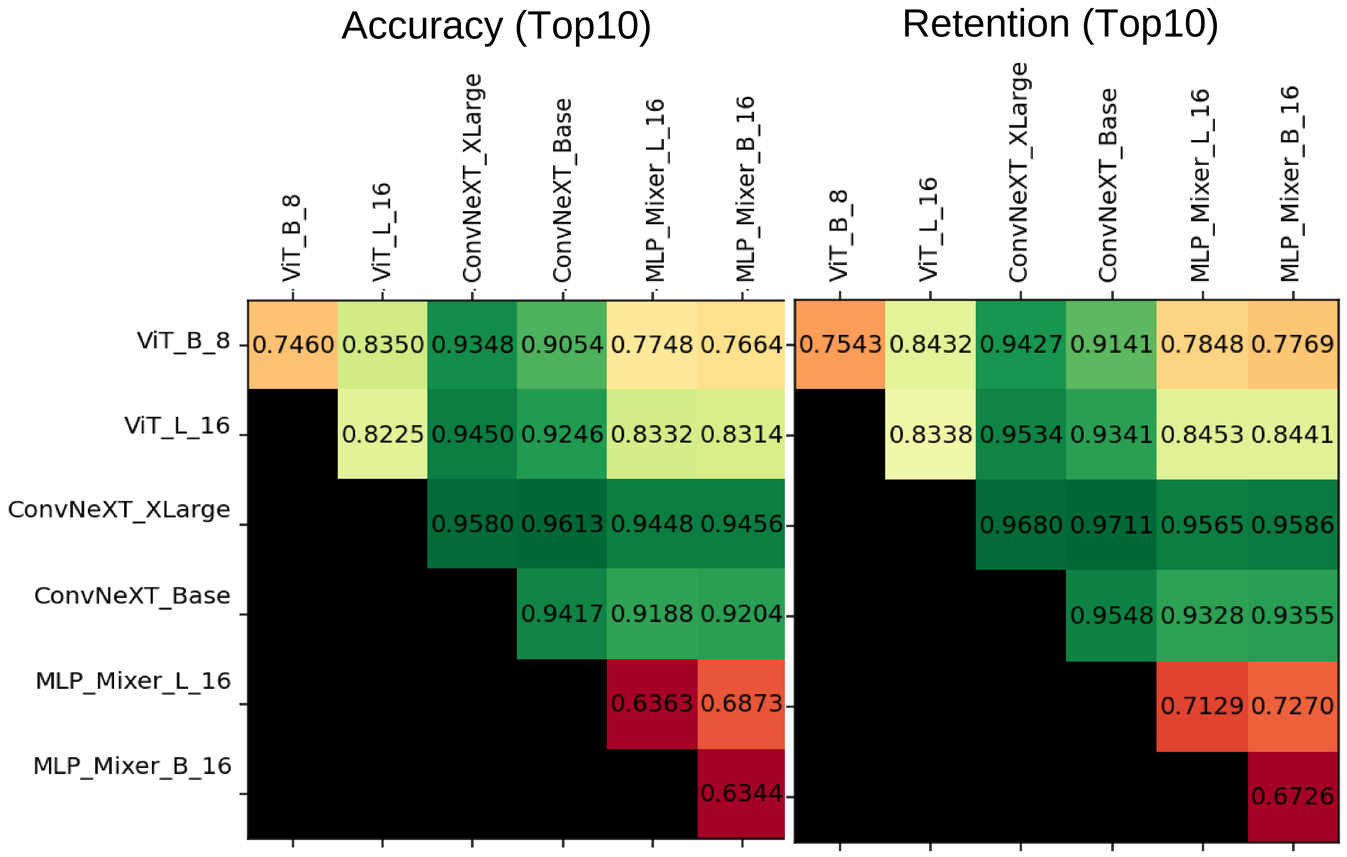
How to read 0-1 Diversity Components plot:
"row_model classifies X % of samples correctly, while column_model classifies them incorrectly"

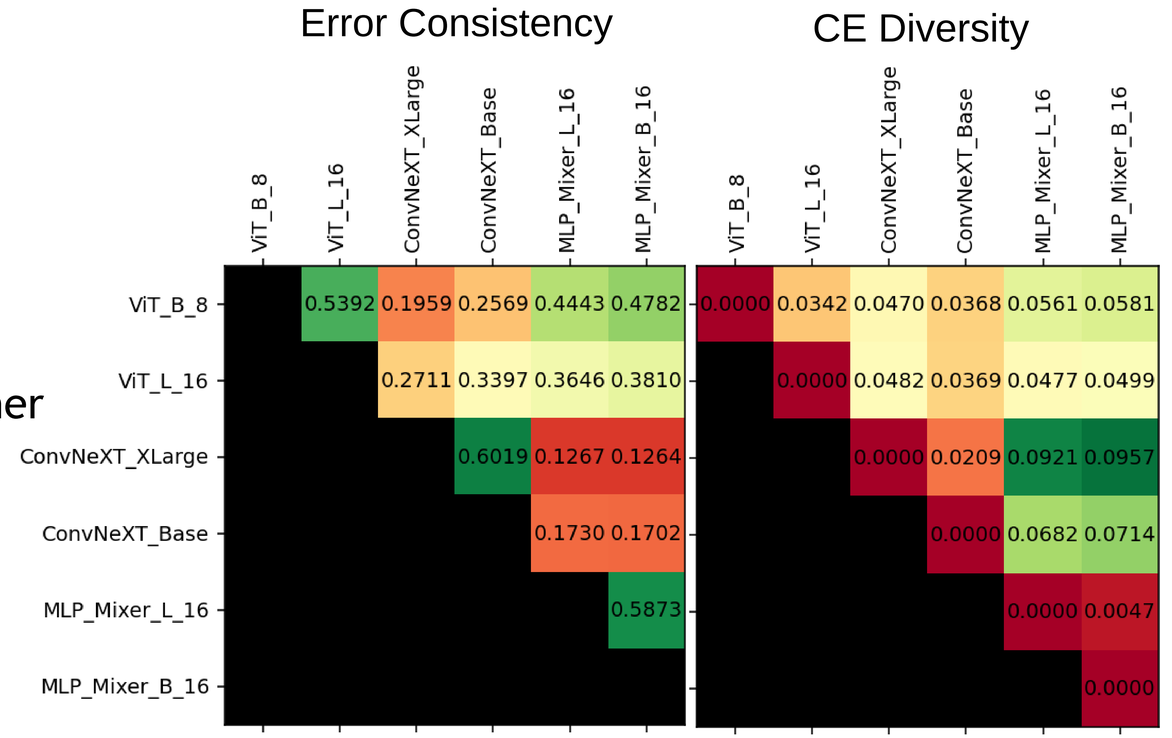
OBSERVATIONS ON DIVERSITY AND ROBUSTNESS

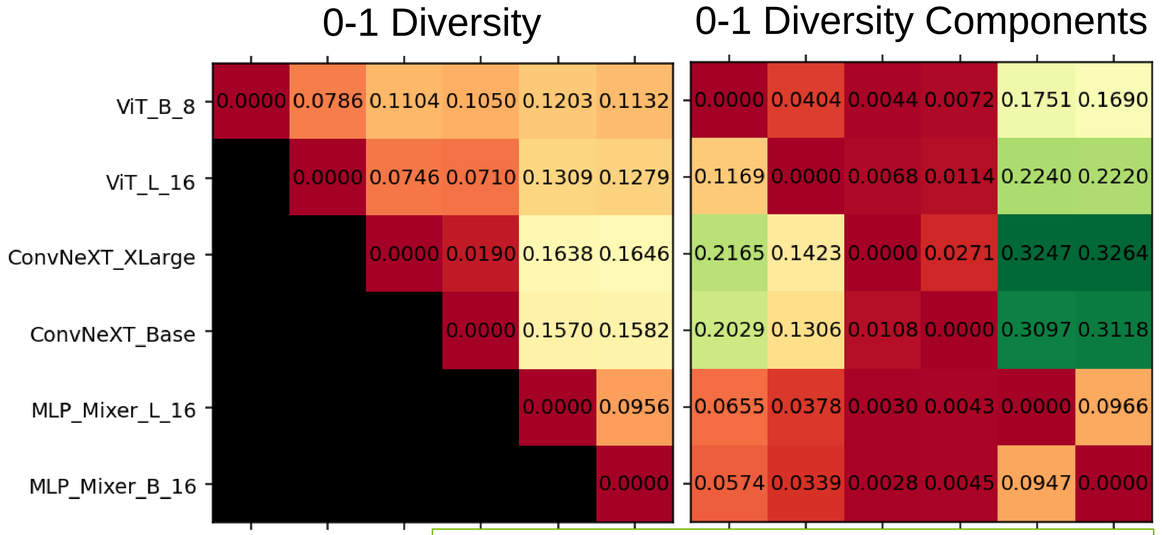
Contrast 4

Key takeaways:

- For some corruptions, one architecture can dominate every other







How to read 0-1 Diversity Components plot:
“row_model classifies X % of samples correctly, while column_model classifies them incorrectly”

CONCLUSION

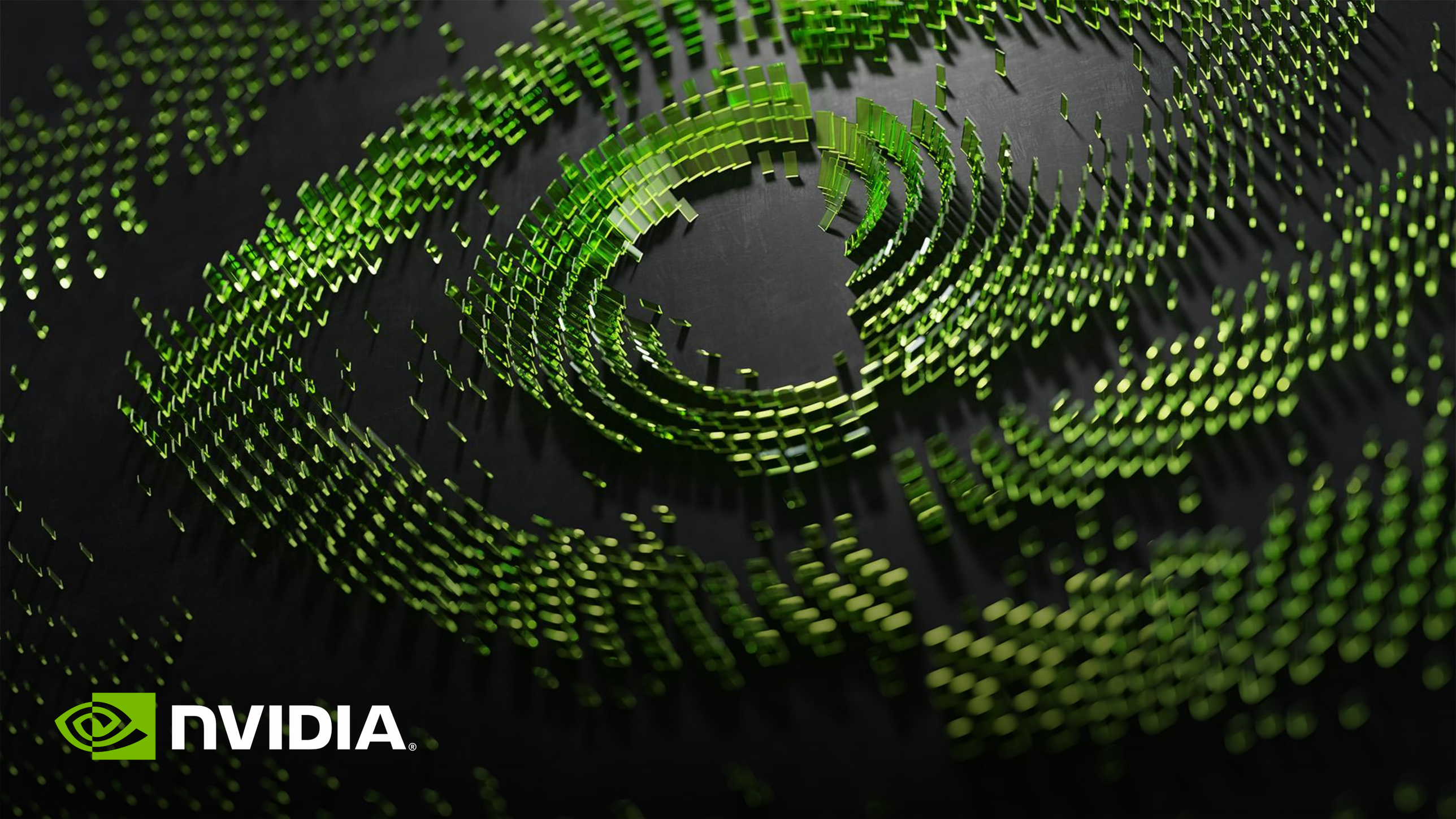
- Simple ensemble manages to aggregate strengths of the underlying architectures
- Diversity metrics and error consistency explains:
  - which models classify differently
  - why some ensembles performs better than others
- However, diversity of classification ≠ potential for ensemble performance gain

NVIDIA.

FUTURE WORK

- Creating new metric for assessing ensemble potential
- Scaling the research to more various models e.g. pretrained on different datasets and more levels of accuracy
- Improving the ensemble technique
- Extending this work to different problems like object detection or image segmentation

NVIDIA.


NVIDIA