COMPARING VISION TRANSFORMERS AND CONVOLUTIONAL NETS FOR SAFETY CRITICAL SYSTEMS

MICHAŁ FILIPIUK, VASU SINGH

MARCH 1ST, SAFEAI 2022

VISION TRANSFORMERS AND ITS DESIRABLE PROPERTIES FOR SAFETY APPLICATIONS

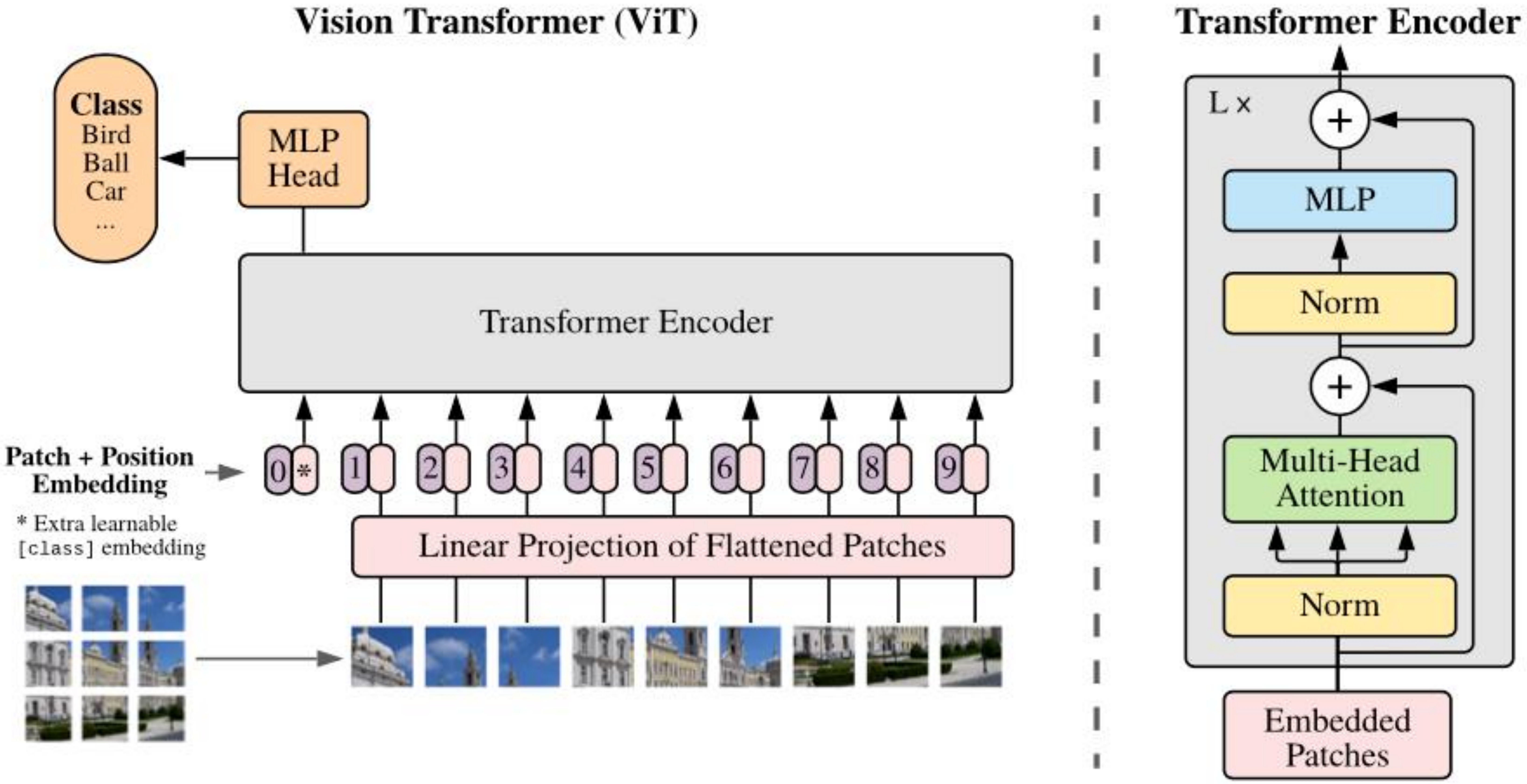


Figure 1. Vision transformer diagram. Source: Google ViT GitHub repository

Reusability

Dosovitskiy et al. 2020

Robustness

Bhojanapalli et al. 2021, Naseer et al. 2021

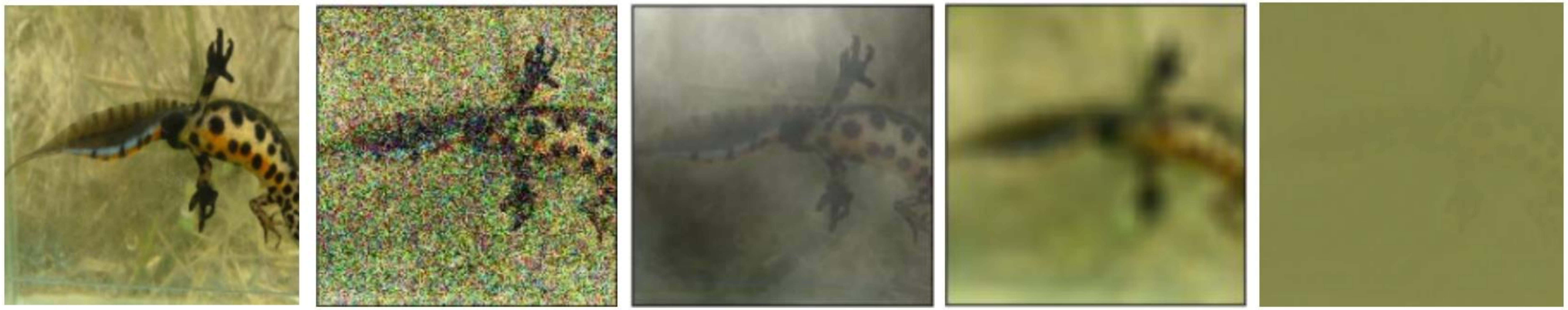
Detection of distribution shift

Fort et al. 2021

Redundancy

Raghu et al. 2021

IMAGENET-C EXPERIMENT



(a) Original image (b) Gaussian noise (c) Fog (d) Defocus Blur (e) Contrast

Figure 2. Sample image and four different corruptions for robustness tests

Model	Original data		Gaussian noise		Defocus blur		Contrast		Fog	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CNN	0.7826	0.9464	0.3780	0.5914	0.3536	0.5762	0.3792	0.6076	0.4700	0.7518
ViT	0.8326	0.9684	**0.5330**	**0.7532**	0.3966	0.5852	0.1980	0.3008	0.6036	0.7744
CNN + ViT	**0.8416**	**0.9726**	0.5130	0.7522	**0.4340**	**0.6634**	**0.4010**	**0.6210**	**0.6276**	**0.8646**

Table 2. Comparison of accuracy for CNN, ViT, and ensemble for ImageNet-C corruptions

ONGOING AND FUTURE WORK

- Revisiting image classification ensembling research:
  - Numerous architectures (pure CNNs, ViT, hybrid approaches)
  - Different model sizes
  - Various pre-/training methods
- Extending the research towards other problems like object detection or image segmentation
- Leveraging transformer architecture in detection of OOD samples in AV environment
- Investigating creating redundant design in resource-constrained systems

NVIDIA®