XAI

Explainability of Vision Transformers: A Comprehensive Review and New Perspectives

Common Attribution Methods

Method	Paper	Dataset	Info
GradCam	Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization	general public images	Image Captioning and VQA achieved using CNN + LSTM on locally activated regions
Integrated Gradients	Axiomatic Attribution for Deep Networks	ImageNet	GoogleNet arch used
ViT Shapley	LEARNING TO ESTIMATE SHAPLEY VALUES WITH VISION TRANSFORMERS	ImageNette, MURA, musculoskeletal radiographs, Oxford IIIT-Pets	expensive (not in production)
LRP	Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers	CIFAR-10, ImageNET, MIT Places	based on firsto order Taylor expansion

Attention Based Methods

Method	Paper	Dataset	Info
Raw Attention	Self-attention for raw optical Satellite Time Series Classification	Crop Satelite Data	LSTM-RNN, MS-ResNet, TempCNN, hyper para selection - HyperOpt, classification task
Attention Rollout & Flow	Quantifying Attention Flow in Transformers	subject-verb agreement	information flow in the network with a DAG, maximum flow algorithms, pretrained Bert
Partial LRP	AttnLRP: Attention-Aware Layer-Wise Relevance Propagation for Transformers	SQuAD v2 (QA), Wikipedia summaery	handles atrribution flow through non linear components better
Beyond Attention	Transformer Interpretability Beyond Attention Visualization	ImageNet, Movie Review, ERASER	relevancy propagation rule applicable to both positive and negative attributions, integrate the attention and the relevancy scores, and combine the integrated results for multiple attention blocks.

Method	Paper	Dataset	Info
GradSam	Grad-SAM: Explaining Transformers via Gradient Self-Attention Maps	Stanford Sentiment Tree (SST), AgNews, IMDB, MultiRC	
Beyond Intuition	Beyond Intuition:Rethinking Token Attributions inside Transformers	ImageNet, Movie Review, Newgroups Review	attention perception and reasoning feedback, head-wise and token-wise approximations, single & bi modality

Pruning-based Methods

Method	Paper	Dataset	Info
IA-Red^2	IA-RED2: Interpretability-Aware Redundancy Reduction for Vision Transformers	ImageNet, Kinetics-400	mutlihead interpreter, both model-agnostic and task-agnostic
Vision DiffMask	VISION DIFFMASK: Faithful Interpretation of Vision Transformers with Differentiable Patch Masking	CIFAR-10, ImageNet	activations of the model's hidden layers
X Pruner	X-Pruner: eXplainable Pruning for Vision Transformers	CIFAR-10, ILSVRC-12	novel explainability-aware mask, DeiT, Swin transformer
EViT	NOT ALL PATCHES ARE WHAT YOU NEED: EXPEDITING VISION TRANSFORMERS VIA TOKEN REORGANIZATIONS	ImageNet, JFT-300M	attentive token indentifiication, inattentive token fusion

Inherently Explainable Methods

Method	Paper	Dataset	Info
ViT-CX	ViT-CX: Causal Explanation of Vision Transformers	ImageNette, MURA	vit feature maps -> mask, high degree of redunduncy, clustering of masks
ViT-Net	ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder	CUB-200-2011, Stanford Cars, Stanford Dogs	neural tree encoder
R-Cut	R-Cut: Enhancing Explainability in Vision Transformers with Relationship Weighted Out and Cut	ImageNet, LRN	

Other

Method	Paper	Dataset	Info
eX-ViT	eX-ViT: A Novel explainable vision transformer for weakly supervised semantic segmentation	PASCAL VOC 2012, MS COCO 2014	Explainable Multi-Head Attention (E-MHA) module and Attribute-guided Explainer (AttE) module