Transformer在CV领域的发展

Transformer在CV领域有两个基础架构，分别是Vision Transformer(ViT)和Swin Transformer,其他的工作是基于目前这两个Transformer框架进行。

- Vision Transformer

paper:arxiv

code:code:pytorch

Key contribution:将自然语言处理领域的Transformer首次大规模应用于计算机视觉的图像分类任务中，并取得了多项SOTA。同时在迁移Transformer的过程中保证了其原始性，揭开了Transformer在计算机视觉广泛应用的高潮。

Related papers:

Area	Title	Key contribution	link	code
3D	PoinTr: Diverse Point Cloud Completion with Geometry-Aware Transformers	a geometry-aware transformer for the set-to-set translation problem	paper(oral)	github
3D	SnowflakeNet: Point Cloud Completion by Snowflake Point Deconvolution with Skip-Transformer	predicting the complete point clouds via skip transformer	paper	-
3D	PlaneTR: Structure-Guided Transformers for 3D Plane Recovery	predicting a sequence of plane instances considering the context and structure clues	paper	-
3D	Point Transformer	construct self-supervised networks for 3d tasks; vector attention	paper	-
3D	Transformer-Based Attention Networks for Continuous Pixel-Wise Prediction	conv+attention; attention based on gates	paper	github
3D detection	Group-Free 3D Object Detection via Transformers	no grouping of local points; compute object features from all the points in the point cloud	paper	-
Audio-visual	The Right to Talk: An Audio-Visual Transformer Approach	visual self-attention, audio self-attention + cross attention	paper	-
Classification	Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet	reorginize tokens in VIT	paper	github
Classification/detection	Rethinking Spatial Dimensions of Vision Transformers	spatial dimension reduction is also useful to transformers	paper	github
Classification/detection	Conformer: Local Features Coupling Global Representations for Visual Recognition	combine conv+transformer	paper	github
Classification/detection	LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference	decreasing resolutions; attention bias	paper	-
Classification/detection	Vision Transformer with Progressive Sampling	sample interesting patches	paper	github
Classification/detection	AutoFormer: Searching Transformers for Visual Recognition	search a new one-shot architecture	paper	-
Classification/detection	Swin Transformer: Hierarchical Vision Transformer using Shifted Windows	hierarchical transformer; shifted windows	paper	github

Classification/detection	Rethinking and Improving Relative Position Encoding for Vision Transformer	design a new position encoding method for 2D images	paper	github
Cross-modal	Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers	interpreting bi-modal and encoder-decoder transformers	paper	github
Detection	Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions	training dense predictions; progressive shrinking pyramids	paper	github
Detection	Fast Convergence of DETR with Spatially Modulated Co-Attention	regression-aware co-attention in DETR by constraining co-attention responses to be high near initially estimated bounding box locations	paper	github
Detection	Conditional DETR for Fast Training Convergence	learns a conditional spatial query from the decoder embedding for decoder multi-head cross-attention	paper	github
Few-shot segmentation	Simpler is Better: Few-shot Semantic Segmentation with Classifier Weight Transformer	only meta-learn the classification part	paper	github
Image quality assessment	MUSIQ: Multi-scale Image Quality Transformer	using original image scale; add scale embeddings	paper	github
Scene understanding; video	Spatial-Temporal Transformer for Dynamic Scene Graph Generation	spatial encoder+temporal decoder; flexible to varying video lengths	paper	-
segmentation	SOTR: Segmenting Objects with Transformers	combine conv+transformer; twin attention	paper	github
Self-supervised learning	An Empirical Study of Training Self-Supervised Vision Transformers	ablation study of vit in self-supervised learning	paper	-
Self-supervised learning	Emerging Properties in Self-Supervised Vision Transformers	emerging semantic segmentation of an image; excellent classifiers. DINO	paper	github
Stroke prediction	Paint Transformer: Feed Forward Neural Painting with Stroke Prediction	predicting strokes via transformer; self-training transformer	paper	github
Visual tracking	Learning Spatio-Temporal Transformer for Visual Tracking	end-to-end bbox prediction for visual tracking	paper	github
masked image modeling	Masked Autoencoders Are Scalable Vision Learners	BERT in vision computer	paper	github

- Swin Transformer

paper:arxiv

code:code:pytorch

Key contribution:随着Vision Transformer的提出在计算机视觉领域掀起了一番热潮，但是Vision Transformer并没有令不同的local information进行交互，没有利用到CNN的局部相关性。同时Vision Transformer无法在目标检测领域直接使用，因为图像的分辨率大小决定了Transformer架构大小。为了解决这些问题，作者提出将相对位置编码结合窗口移位的方式，并在人脸识别，图像分类，目标检测，语义分割多个计算机视觉应用领域取得了SOTA效果。是目前最完善的计算机视觉领域大一统框架。

Area	Title	Key contribution	link	code
Self-supervised learning	Self-Supervised Learning with Swin Transformers	MoCo v2+BYOL;use Swin-T as backbone	paper	github
Video	Video Swin Transformer	use Swin-T as backbone in Video;advocate an inductive bias of locality	paper	github

Classification/detection	Swin Transformer: Hierarchical Vision Transformer using Shifted Windows	MLP was added to the original article of swin transformer	paper	github
Detection	End-to-End Semi-Supervised Object Detection with Soft Teacher	using pseudo tags to assist target detection;soft teacher mechanism;a box jittering;	paper	-
Masked image modeling	SimMIM: A Simple Framework for Masked Image Modeling	MAE+Swin-T	paper	github
Face recognition	StyleSwin: Transformer-based GAN for High-resolution Image Generation	Swin Transformer for StyleGAN	paper	-
Face recognition	FaceX-Zoo: A PyTorh Toolbox for Face Recognition	Swin Transformer for Face Recognition; is a pytorch tool in face recognition	paper	github
Person reID	-	Swin Transformer for person reID	-	github
Image Restoration	SwinIR: Image Restoration Using Swin Transformer	Swin Transformer for Image Restoration	paper	github
Classification/detection	Swin Transformer V2: Scaling Up Capacity and Resolution	a post normalization technique and a scaled cosine attention;the second version of swin transformer	paper	github
Classification/detection	PyramidTNT: Improved Transformer-in-Transformer Baselines with Pyramid Architecture	Pyramid feature extraction architecture+Swin-T	paper	github
Detection	HRFormer: High-Resolution Transformer for Dense Prediction	HRNet+Swin-T	paper	github
Detection/semantic segmentation/model compression	CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows	Swin-T+Cross-Shaped Window self-attention	paper	github
Crowd localization	Congested Crowd Instance Localization with Dilated Convolutional Swin Transformer	Swin-Transformer for crowd localization	paper	github
Medical imaging	COVID-19 Detection in Chest X-ray Images Using Swin-Transformer and Transformer in Transformer	Swin-Transformer for COVID-19 Detection	paper	-