Swin Transformer重磅升级！Swin V2：向更大容量、更高分辨率的更大模型迈进

原创 CV开发者都爱看的 极市平台 2021-11-19 22:00:00 手机阅读

收录于话题
#Transformer 45 #神经网络结构设计 25

↑ 点击蓝字 关注极市平台

壹伴图
极市平台
extreme
月发文数目：
月平均阅读：
文章工具
已发文
采集图文 合成多
采集样式 查看

作者丨happy
编辑丨极市平台

极市导读

针对SwinV1在更大模型方面存在的几点问题，Swin transformer V2提出了后规范化技术、对数空间连续位置偏置技术、大幅降低GPU占用的实现等得到了具有超高性能的SwinV2，刷新了多个基准数据集的指标。 >>加入极市CV技术交流群，走在计算机视觉的最前沿

Swin Transformer V2: Scaling Up Capacity and Resolution

Ze Liu* Han Hu*† Yutong Lin Zhuliang Yao Zhenda Xie Yixuan Wei Jia Ning
Yue Cao Zheng Zhang Li Dong Furu Wei Baining Guo
Microsoft Research Asia
{v-zeliu1,hanhu,t-yutonglin,t-zhuyao,t-zhxie,t-yixuanwei,v-jianing}@microsoft.com
{yuecao,zhez,lidong1,fuwei,bainguo}@microsoft.com

论文链接：https://arxiv.org/pdf/2111.09883.pdf

代码链接：https://github.com/microsoft/Swin-Transformer

> **SwinTransformer重磅升级**！MSRA提出SwinV2，朝着更大容量、更高分辨率的更大模型出发，在多个基准数据集(包含ImageNet分类、COCO检测、ADE20K语义分割以及Kinetics-400动作分类)上取得新记录。针对SwinV1在更大模型方面存在的几点问题，提出了后规范化技术、对数空间连续位置偏置技术、大幅降低GPU占用的实现等得到了具有超高性能的SwinV2，刷新了多个基准数据集的指标。

Abstract

本文提出一种升级版SwinTransformerV2，最高参数量可达3 Billion，可处理1536×1536尺寸图像。通过提升模型容量与输入分辨率，SwinTransformer在四个代表性基准数据集上取得了新记录：84.%@ImageNetV2、63.1 box 与54.4 max mAP@COCO、59.9mIoU@ADE20K以及86.8%@Kinetics-400(视频动作分类)。

所提技术可以广泛用于视觉模型缩放，Transformer的缩放技术再NLP语言建模中已得到广泛探索，但在视觉任务中尚未进行。主要是因为以下几点训练与应用难题：

- 视觉模型通常面临**尺度不稳定** 问题；
- 下游任务需要高分辨率图像，尚不明确**如何将低分辨率预训练模型迁移为高分辨率版本** ；
- 此外，当图像分辨率非常大时，**GPU显存占用** 也是个问题。

为解决上述问题，我们以SwinTransformer作为基线，提出了几种改进技术：

- 提出**后规范化(Post Normalization)技术** 与可缩放(Scaled)cosine注意力提升大视觉模型的稳定性;
- 提出**log空间连续位置偏置** 技术进行低分辨率预训练模型向高分辨率模型迁移。
- 此外，我们还共享了**至关重要的实现细节** ，它可以大幅节省GPU显存占用以使得大视觉模型训练变得可行。

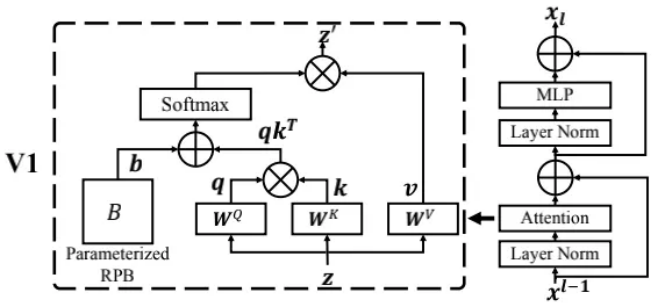
基于上述技术与自监督预训练，我们成功训练了一个包含3B参数量的SwinTransformer模型并将其迁移到不同的高分辨率输入的下游任务上，取得了SOTA性能。

Method

A Brief Review of Swin Transformer

Swin Transformer是一种通用的视觉骨干模型，在不同的视觉任务(包含图像分类、目标检测以及语义分割)上均取得了极强性能。Swin Transformer的主要思想：**为常规Transformer Encoder架构引入了几个重要的视觉信号先验信息** ，包含分层、局部以及平移不变形。基础Transformer单元提供了强建模能力，视觉信号先验信息使其对不同视觉任务极为友好。

Normalization Configuration 众所周知，规范化技术对于更深架构的训练非常重要。原始的SwinTransformer采用了常规的预规范化技术，见下图。



Relative position bias 它是原始SwinTransformer的一个关键成分，它引入了一个额外参数化偏置，公式如下：

$$Attention(Q,K,V)=\mathrm{SoftMax}(QK^T/\sqrt{d}+B)V$$

其中,$B\in R^{M^2\times M^2}$是每个head的相对位置偏置，它对于稠密识别任务非常重要。当进行不同分辨率模型迁移时，常规方案是对该偏置进行双三次插值近似。

Issues in scaling up model capacity and window resolution 在对SwinTransformer进行容量与窗口分辨率缩放过程中，我们发现以下两个问题：

- 容量缩放过程中的不稳定问题，见下图。

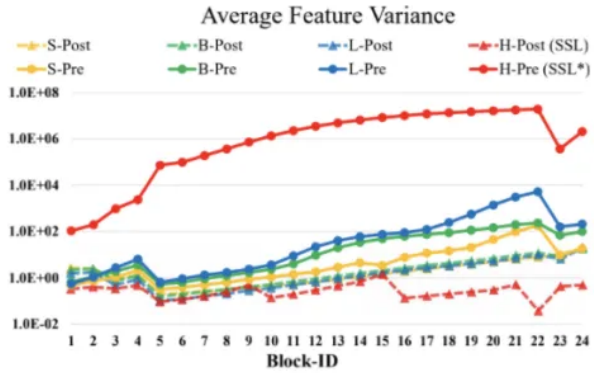


Figure 2. Signal Propagation Plot [5, 62] for various model sizes. The H-size models are trained at a self-supervisied learning stage and other sizes are trained by the classification task. * indicates that we use a 40-epoch model before it crashes.

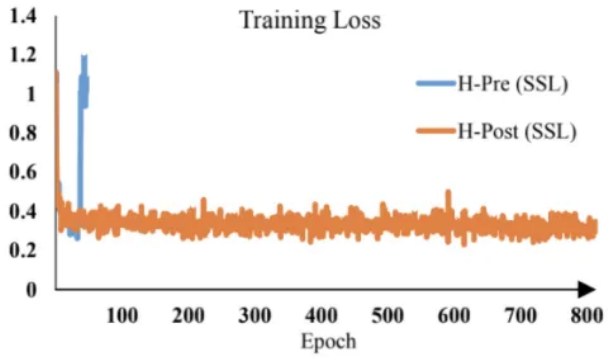


Figure 3. SwinV1-H versus SwinV2-H in training [59].

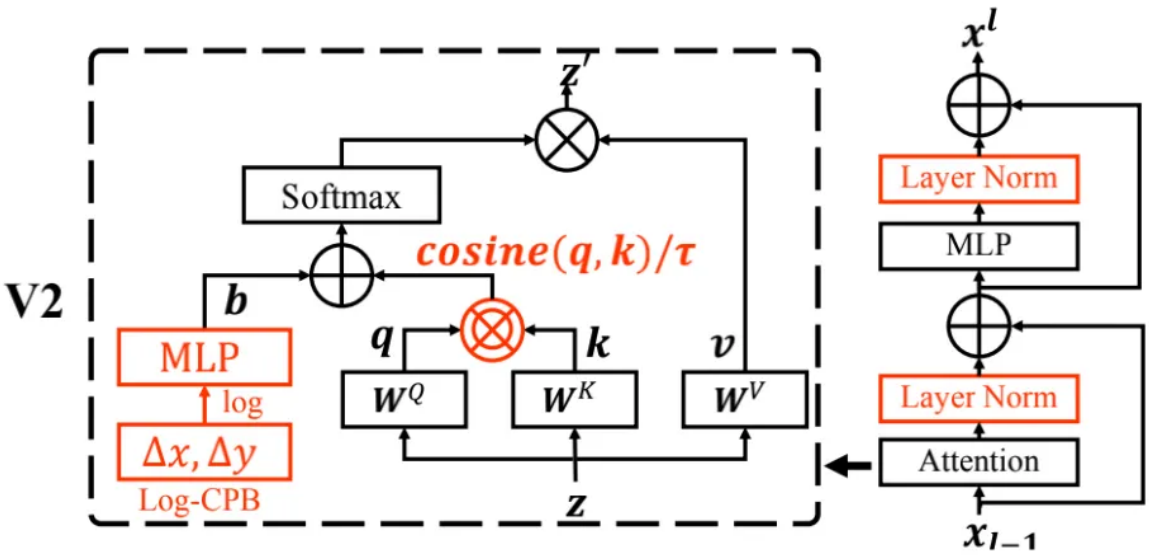
- 跨分辨率迁移时的性能退化问题，见下表。

method	ImageNet*	ImageNet†				COCO		ADE20k		
	W8, I256 top-1 acc	W12, I384 top-1 acc	W16, I512 top-1 acc	W20, I640 top-1 acc	W24, I768 top-1 acc	W16 AP^{box}	W32 AP^{box}	W16 mIoU	W20 mIoU	W32 mIoU
Parameterized position bias [35]	81.7	79.4/82.7	77.2/83.0	73.2/83.2	68.7/83.2	50.8	50.9	45.5	45.8	44.5
Linear-Spaced CPB	81.7 (+0.0)	82.0/82.9 (+2.6/+0.2)	81.2/83.3 (+4.0/+0.3)	79.8/83.6 (+6.6/+0.4)	77.6/83.7 (+8.9/+0.5)	50.9 (+0.1)	51.7 (+0.8)	47.0 (+1.5)	47.4 (+1.6)	47.2 (+2.7)
Log-Spaced CPB	81.8 (+0.1)	82.4/83.2 (+3.0/+0.5)	81.7/83.8 (+4.5/+0.8)	80.4/84.0 (+7.2/+0.8)	79.1/84.2 (+10.4/+1.0)	51.1 (+0.3)	51.8 (+0.9)	47.0 (+1.5)	47.7 (+1.9)	47.8 (+3.3)

Table 1. Comparison of different position bias computation approaches using Swin-T. * indicates the top-1 accuracy on ImageNet-1k trained from scratch. The models in * column will be used for testing on the ImageNet-1K image classification task using larger image/window resolutions, marked by †. For these results, we report both the results w.o./with fine-tuning. These models are also used for fine-tuning on COCO object detection and ADE20K semantic segmentation tasks.

Scaling up Model Capacity

正如上面所提到：原始SwinTransformer采用了预规范化技术。可以看到：**当对模型容量进行缩放时，深层的激活值会极大提升**。



事实上，在预规范化配置下，每个残差模块的输出激活值与主分支直接合并，导致主分支在更深层的幅值越来越大，进而导致训练不稳定。

Post Normalization 为缓解该问题，我们提出了Post Normalization(后规范化)：每个残差模块的输出先进行规范化再与主分支进行合并，因此主分支的幅值不会逐层累积。从上面的Figure2可以看到：使用后规范化的模型激活幅值更温和。

在最大的模型中，我们每6个Transformer模块额外引入一个LN单元以进一步稳定训练。

Scaled Cosine Attention 在原始自注意力计算过程中，像素对的像素性通过query与key的点积计算。我们发现：**在大模型中，某些模块与head的注意力图会被少量像素对主导** 。为缓解该问题，我们提出了Scaled Cosine Attention(SCA)，公式如下：

$$\mathrm{Sim}(q_i, k_j) = cos(q_i, k_j)/\tau + B_{ij}$$

Scaling Up Window Resolution

接下来，我们引入一种log空间连续位置偏置方法以使得相对位置偏置跨窗口分辨率平滑迁移。

Continuous Relative Position Bias 不同于直接对偏置参数直接优化，连续位置偏置方法采用了针对相对坐标的元网络：

$$B(\Delta x, \Delta y) = \mathcal{G}(\Delta x, \Delta y)$$

注：$\mathcal{G}$是一个很小的网络，比如2层MLP。它对任意相对坐标生成偏置参数，因而可以自然地进行任意可变窗口尺寸的迁移。在推理阶段，每个相对位置的偏置可以预先计算并保存，按照原始方式进行推理。

Log-space Coordinates 当跨大窗口迁移时，有较大比例的相对坐标范围需要外插。为缓解该问题，我们采用了对数空间坐标：

$$\hat{\Delta x} = sign(x) \cdot log(1 + |\Delta x|) \hat{\Delta y} = sign(y) \cdot log(1 + |\Delta y|)$$

通过对数空间坐标，在进行块分辨率迁移时，所需的外插比例会更小。比如，将8×8预训练模型向16×16迁移时，输入坐标范围从$[-7, 7] \times [-7, 7]$调整为$[-15, 15] \times [-15, 15]$，外插比例为$\frac{8}{7} = 1.14\times$。而采用对数空间坐标，输入坐标范围从$[-2.079, 2.079] \times [-2.079, 2.079]$调整为$[-2.773, 2.773] \times [-2.773, 2.773]$，外插比例为$0.33\times$。下表则给出了不同位置偏置下的迁移性能对比，可以看到：**当向更大窗口尺寸迁移时，对数空间连续位置偏置性能最佳** 。

method	ImageNet*	ImageNet†				COCO		ADE20k		
	W8, I256 top-1 acc	W12, I384 top-1 acc	W16, I512 top-1 acc	W20, I640 top-1 acc	W24, I768 top-1 acc	W16 AP^{box}	W32 AP^{box}	W16 mIoU	W20 mIoU	W32 mIoU
Parameterized position bias [35]	81.7	79.4/82.7	77.2/83.0	73.2/83.2	68.7/83.2	50.8	50.9	45.5	45.8	44.5
Linear-Spaced CPB	81.7 (+0.0)	82.0/82.9 (+2.6/+0.2)	81.2/83.3 (+4.0/+0.3)	79.8/83.6 (+6.6/+0.4)	77.6/83.7 (+8.9/+0.5)	50.9 (+0.1)	51.7 (+0.8)	47.0 (+1.5)	47.4 (+1.6)	47.2 (+2.7)
Log-Spaced CPB	81.8 (+0.1)	82.4/83.2 (+3.0/+0.5)	81.7/83.8 (+4.5/+0.8)	80.4/84.0 (+7.2/+0.8)	79.1/84.2 (+10.4/+1.0)	51.1 (+0.3)	51.8 (+0.9)	47.0 (+1.5)	47.7 (+1.9)	47.8 (+3.3)

Table 1. Comparison of different position bias computation approaches using Swin-T. * indicates the top-1 accuracy on ImageNet-1k trained from scratch. The models in * column will be used for testing on the ImageNet-1K image classification task using larger image/window resolutions, marked by †. For these results, we report both the results w.o./with fine-tuning. These models are also used for fine-tuning on COCO object detection and ADE20K semantic segmentation tasks.

Other Implementation

Implementation to save GPU memory 大分辨率输入与大容量模型存在的另一个问题是**GPU显存占用不可接受问题** 。我们采用了以下实现改善该问题：

- Zero-Redundancy Optimizer(ZeRO): 采用ZeRO优化器减少GPU显存占用，对整体训练速度影响极小；
- Activation check-pointing：采用checkpoint技术节省GPU占用，但会降低30%训练速度；
- Sequential Self-attention computation：采用串式计算，而非batch模式，对整体训练速度影响极小。

通过上述实现，我们可以在Nvidia A100-40G GPU训练参数量3B的模型(COCO检测与ImageNet分类，输入为1536×1536)。

Joining with a self-supervised approach

更大的模型需要更多地数据(data hungry)。为解决该问题，之前的大模型训练通过采用额外的数据或者自监督预训练。我们对这两种策略进行了组合

- 额外数据：我们对ImageNet-22K进行扩大五倍达到了70M数量；
- 自监督学习：我们采用了自监督训练以更好的进行数据挖掘。

通过上述训练方案，我们训练了一个具有3B参数量的SwinTransformer模型并在多个基准数据集上取得了SOTA性能。

Model Configurations

我们保持与SwinTransformer相同的stage、block以及通道配置得到了四个版本的SwinTransformerV2：

- SwinV2-T: C96, layer number= {2,2,6,2}
- SwinV2-S: C96, layer number= {2,2,18,2}
- SwinV2-B: C128, layer number= {2,2,18,2}
- SwinV2-L: C192, layer number= {2,2,18,2}

我们进一步对SwinV2进行更大尺寸缩放得到了658M与3B参数模型：

- SwinV2-H: C=352, layer number={2,2,18,2}
- SwinV2-G: C=512, layer number={2,2,42,2}

Experiments

本文主要在ImageNetV1、ImageNetV2、COCO检测、ADE20K语义分割以及Kinetics-400视频动作分类方面进行了实验。

Method	param	pre-train images	pre-train length (#im)	pre-train im size	pre-train time	fine-tune im size	ImageNet-1K-V1 top-1 acc	ImaegNet-1K-V2 top-1 acc
SwinV1-B	88M	IN-22K-14M	1.3B	224^2	$<30^\dagger$	384^2	86.4	76.58
SwinV1-L	197M	IN-22K-14M	1.3B	224^2	$<10^\dagger$	384^2	87.3	77.46
ViT-G [65]	1.8B	JFT-3B	164B	224^2	>30k	518^2	90.45	83.33
V-MoE [44]	14.7B*	JFT-3B	-	224^2	16.8k	518^2	90.35	-
CoAtNet-7 [11]	2.44B	JFT-3B	-	224^2	20.1k	512^2	**90.88**	-
SwinV2-B	88M	IN-22K-14M	1.3B	192^2	$<30^\dagger$	384^2	87.1	78.08
SwinV2-L	197M	IN-22K-14M	1.3B	192^2	$<20^\dagger$	384^2	87.7	78.31
SwinV2-G	3.0B	IN-22K-ext-70M	3.5B	192^2	$<0.5k^\dagger$	640^2	90.17	**84.00**

Table 2. Comparison with previous largest vision models on ImageNet-1K V1 and V2 classification. * indicates the sparse model; the "pre-train time" column is measured by the TPUv3 core days with numbers copied from the original papers. † That of SwinV2-G is estimated according to training iterations and FLOPs.

上表给出了ImageNet分类任务上的性能对比，可以看到：

- 在ImageNetV1数据上，SwinV2-G取得了90.17%的精度；

- 在ImageNetV2数据上，SwinV2-G取得了84.0%的精度，比之前最佳高0.7%；

- 相比SwinV1，SwinV2性能提升约0.4~0.8%。

Method	train I(W) size	test I(W) size	mini-val (AP) box	mini-val (AP) mask	test-dev (AP) box	test-dev (AP) mask
CopyPaste [17]	1280(-)	1280(-)	57.0	48.9	57.3	49.1
SwinV1-L [35]	800(7)	ms(7)	58.0	50.4	58.7	51.1
YOLOR [53]	1280(-)	1280(-)	-	-	57.3	-
CBNet [32]	1400(7)	ms(7)	59.6	51.8	60.1	52.3
DyHead [10]	1200(-)	ms(-)	60.3	-	60.6	-
SoftTeacher [60]	1280(12)	ms(12)	60.7	52.5	61.3	53.0
SwinV2-L (HTC++)	1536(32)	1100(32)	58.8	51.1	-	-
		1100 (48)	58.9	51.2	-	-
		ms (48)	60.2	52.1	60.8	52.7
SwinV2-G (HTC++)	1536(32)	1100(32)	61.7	53.3	-	-
		1100 (48)	61.9	53.4	-	-
		ms (48)	**62.5**	**53.7**	**63.1**	**54.4**

Table 3. Comparison with previous best results on COCO object detection and instance segmentation. I(W) indicates the image and window size. ms indicate multi-scale testing is employed.

上表比较了COCO检测任务上的性能，可以看到：**所提方案取得了63.1/54.4的box与mask mAP指标，比此前最佳高1.8/1.4** 。

Method	train I(W) size	test I(W) size	mIoU
SwinV1-L [35]	640(7)	640(7)	53.5*
Focal-L [61]	640(40)	640(40)	55.4*
CSwin-L [14]	640(40)	640(40)	55.7*
MaskFormer [8]	640(7)	640(7)	55.6*
FaPN [22]	640(7)	640(7)	56.7*
BEiT [3]	640(40)	640(40)	58.4*
SwinV2-L (UperNet)	640(40)	640(40)	55.9*
SwinV2-G (UperNet)	640(40)	640(40)	59.1
		896 (56)	59.3
		896 (56)	**59.9***

Table 4. Comparison with previous best results on ADE20K semantic segmentation. * indicates multi-scale testing is used.

上表比较了ADE20K语义分割任务上的性能，可以看到：**所提方案取得了59.9mIoU指标，比此前最佳高1.5** 。

Method	train I(W) size	test I(W) size	views	top-1
ViViT [1]	-(-)	-(-)	4×3	84.8
SwinV1-L [36]	480×480×16 (12×12×8)	480×480×16 (12×12×8)	10×5	84.9
TokenLearner [45]	256×256×64 (8×8×64)	256×256×64 (8×8×64)	4×3	85.4
Video-SwinV2-G	320×320×8 (20×20×8)	320×320×8 (20×20×8)	1×1	83.2
		384×384×8 (24×24×8)	1×1	83.4
		384×384×8 (24×24×8)	4×5	**86.8**

Table 5. Comparison with previous best results on Kinetics-400 video action classification.

上表比较了Kinetics-400视频动作分类任务上的性能，可以看到：**所提方案取得了86.8%的精度，比此前最佳高1.4%** 。

如果觉得有用，就请分享到朋友圈吧！

极市平台
专注计算机视觉前沿资讯和技术干货，官网：www.cvmart.net
624篇原创内容

公众号

Δ点击卡片关注极市平台，获取最新CV干货

公众号后台回复“transformer”获取最新Transformer综述论文下载～

极市干货

课程/比赛：珠港澳人工智能算法大赛丨保姆级零基础人工智能教程

算法trick：目标检测比赛中的tricks集锦丨从39个kaggle竞赛中总结出来的图像分割的Tips和Tricks

技术综述：一文弄懂各种loss function丨工业图像异常检测最新研究总结（2019-2020）

极市平台签约作者#

happy

知乎：AIWalker

AIWalker运营、CV技术深度Follower、爱造各种轮子

研究领域：专注low-level，对CNN、Transformer、MLP等前沿网络架构

保持学习心态，倾心于AI技术产品化。

公众号：AIWalker

作品精选

- 吊打一切现有版本的YOLO！旷视重磅开源YOLOX：新一代目标检测性能速度担当！
- YOLOv4团队开源最新力作！1774fps、COCO最高精度，分别适合高低端GPU的YOLO
- 图像增强领域大突破！以1.66ms的速度处理4K图像，港理工提出图像自适应的3DLUT

投稿方式：

添加小编微信Fengcall（微信号：fengcall19），备注：**姓名-投稿**

Δ长按添加极市平台小编

觉得有用麻烦给个在看啦~

阅读原文

喜欢此内容的人还喜欢

当Swin Transformer遇上DCN，清华可变形注意力Transformer模型优于多数ViT

磐创AI