How to help your ViT learn the inductive bias ?

Paper Reading by Yiwei Sun

2023.03.27

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

Inductive bias

归纳偏置：
Wikipedia：学习算法中，当学习器去预测其未遇到过的输入结果时，所做的一些**假设的集合**。
西瓜书：看作学习算法自身在庞大的假设空间中对假设进行选择的启发式或“价值观”。

CNN结构中蕴含的归纳偏置：

1. Locality：空间位置上的元素的相关性近大远小；控制复杂度。
2. Translation equivariance：相同的物体在不同的位置具有相同的响应；提高模型的泛化能力。

ViT中对归纳偏置的叙述

1. ViT相比CNN，缺少一些归纳偏置：仅有MLP是local且translationally equivariant，自注意力层是global的；
2. ViT中甚至图像本身的二维结构都被破坏：除了生成patch和对编码进行尺寸调整外，其余都以序列的形式组合。

带来的结果：由于缺少一定的归纳偏置，ViT在小数据集上训练后的效果不如CNN，但是随着数据集的规模增大，其也能够直接学习到相关特征。

Change the Backbone

ICCV2021

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

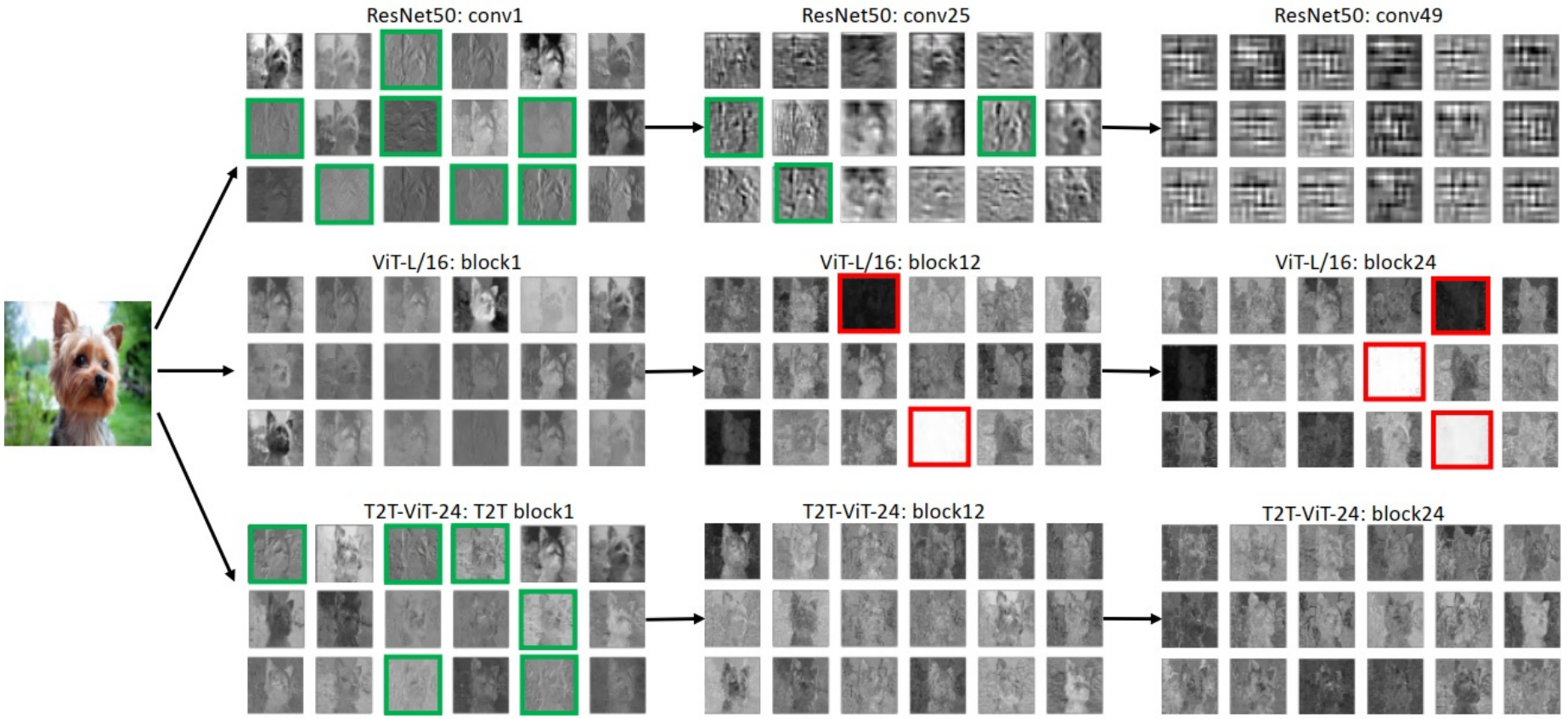
Li Yuan[1]*, Yunpeng Chen[2], Tao Wang[1,3]*, Weihao Yu[1], Yujun Shi[1],
Zihang Jiang[1], Francis E.H. Tay[1], Jiashi Feng[1], Shuicheng Yan[1]

[1] National University of Singapore [2] YITU Technology [3] Institute of Data Science, National University of Singapore

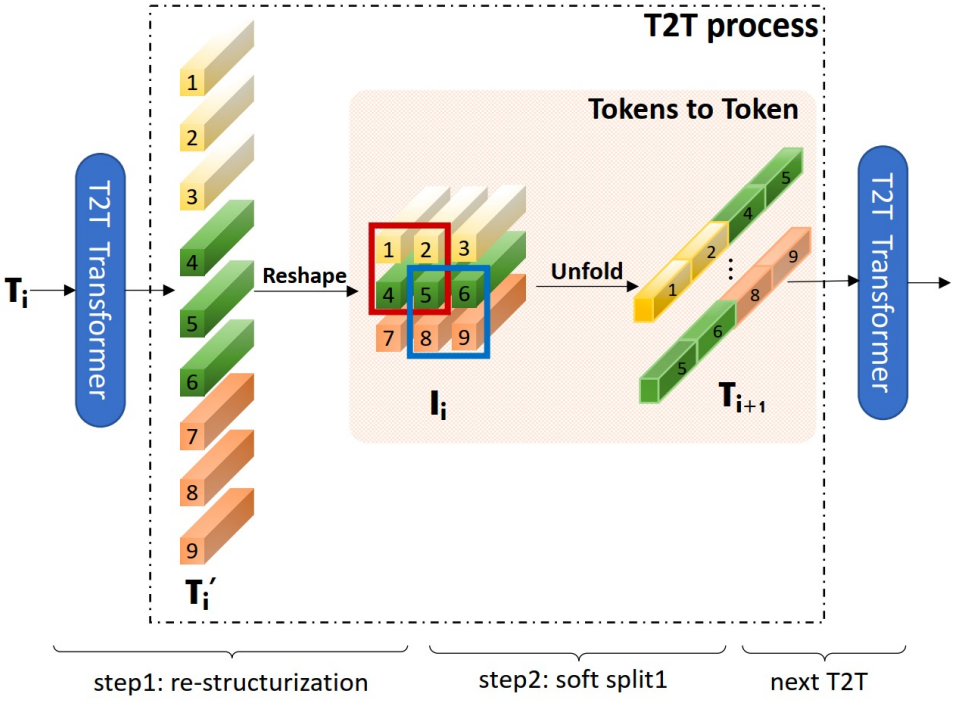
yuanli@u.nus.edu, yunpeng.chen@yitu-inc.com, shuicheng.yan@gmail.com

1. Hard split使得ViT无法对图像的局部结构（边、线条）进行建模；
2. ViT的Attention结构设计冗余，难以在有限的训练集中产生丰富的特征图。

Change the Backbone



Change the Backbone

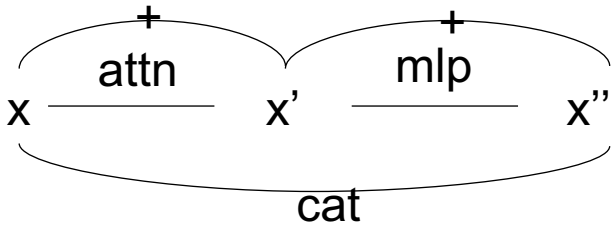


1. 输入图像/将tokens reshape成二维结构；
2. 一个kernel内的pixels/tokens按照channel维度拼接；
3. 将overlap的pixels/tokens送入Transformer中。

Change the Backbone

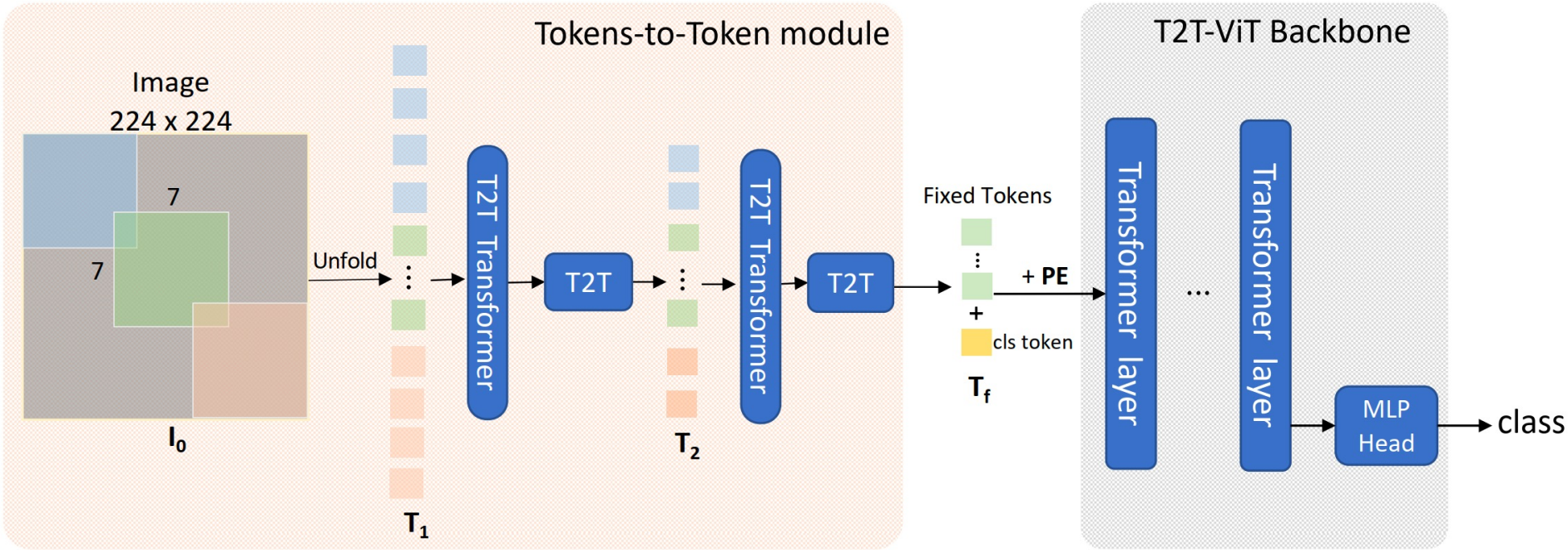
因为ViT中的通道数存在大量的冗余，因此借鉴了CNN的一些设计思想来优化ViT的结构：

1. Dense connection；
2. Deep-narrow and Shallow-wide；
3. Channel Attention；
4. More split heads in multi-head attention layer；
5. Ghost operation.



CNN to ViT	ViT-S/16 (Baseline)	78.1	48.6	10.1	8	768
	ViT-DN	79.0 (+0.9)	24.5	5.5	16	384
	ViT-SW	69.9 (-8.2)	47.9	9.9	4	1024
	ViT-Dense	76.8 (-1.3)	46.7	9.7	19	128-736
	ViT-SE	**78.4** (+0.3)	49.2	10.2	8	768
	ViT-ResNeXt	78.0 (-0.1)	48.6	10.1	8	768
	ViT-Ghost	73.7 (-4.4)	32.1	6.9	8	768

Change the Backbone



Adapter

ICLR2023:

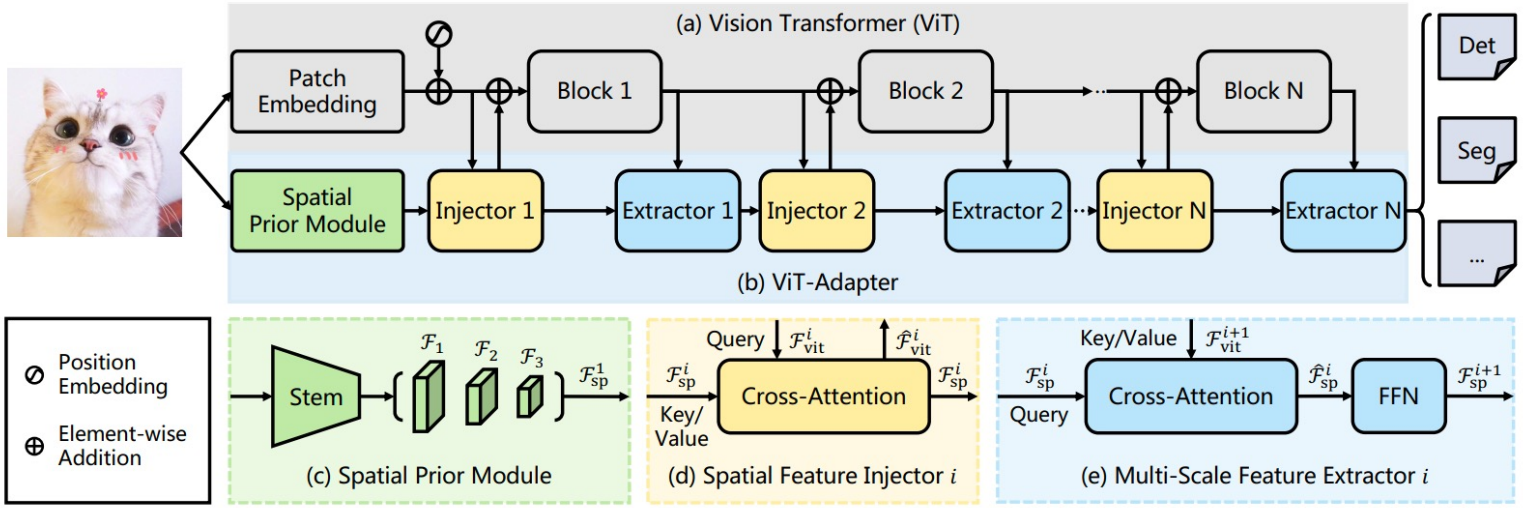
VISION TRANSFORMER ADAPTER FOR DENSE PREDICTIONS

Zhe Chen[1,2*], **Yuchen Duan**[2,3*], **Wenhai Wang**[2✉], **Junjun He**[2],
Tong Lu[1✉], **Jifeng Dai**[2,3], **Yu Qiao**[2]
[1]Nanjing University, [2]Shanghai AI Laboratory, [3]Tsinghua University
czcz94cz@gmail.com, {duanyuchen,wangwenhai,hejunjun}@pjlab.org.cn
lutong@nju.edu.cn, {daijifeng,qiaoyu}@pjlab.org.cn

Plain ViT缺乏与图像相关的先验知识，所以在下游任务（本文以密集预测任务为主）中与visual-specific transformer存在差距（收敛速度和性能）。

本文的目的是提出一种adapter来跨越他们之间的差距。

Adapter



SPM

卷积能更好地捕捉局部的空间信息，为了不影响ViT结构，将SPM设计为并行的CNN结构。

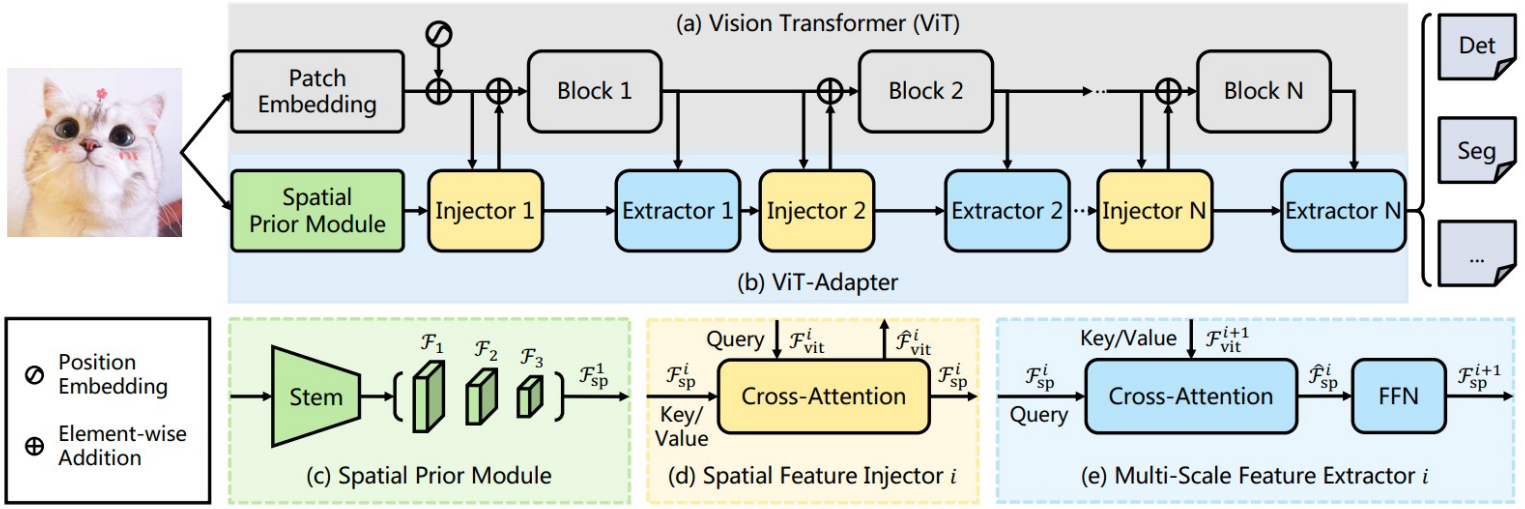
```
def forward(self, x):
    c1 = self.stem(x)
    c2 = self.conv2(c1)
    c3 = self.conv3(c2)
    c4 = self.conv4(c3)
    c1 = self.fc1(c1)
    c2 = self.fc2(c2)
    c3 = self.fc3(c3)
    c4 = self.fc4(c4)
```

Stem由3个卷积层和1个最大池化层组成

Conv2-3为核大小是3，步长为2的卷积，得到不同尺度的特征图

Fc1-4调整为相同的通道后展开为B C L

Adapter

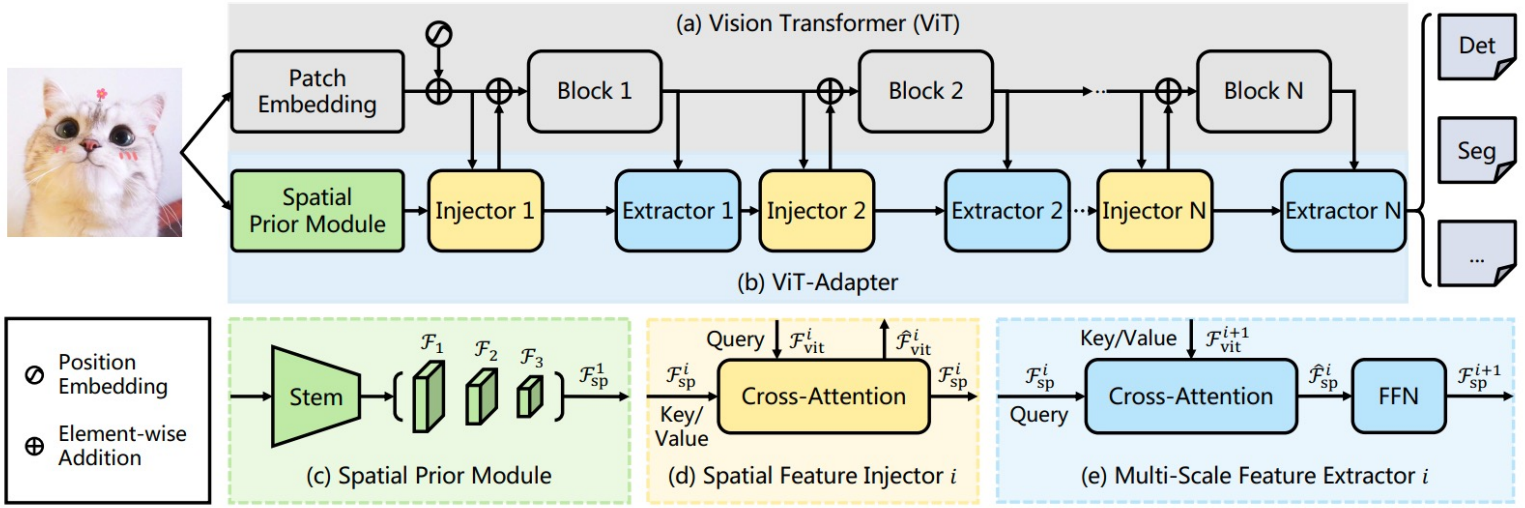


Injector

$$\hat{\mathcal{F}}^i_{\text{vit}} = \mathcal{F}^i_{\text{vit}} + \gamma^i \text{Attention}(\text{norm}(\mathcal{F}^i_{\text{vit}}), \text{norm}(\mathcal{F}^i_{\text{sp}}))$$

1. 采用交叉注意力对CNN特征和ViT特征进行耦合，这里采用MultiScaleDeformableAttention；
2. 加入可学习参数，平衡融合特征和原特征；

Adapter



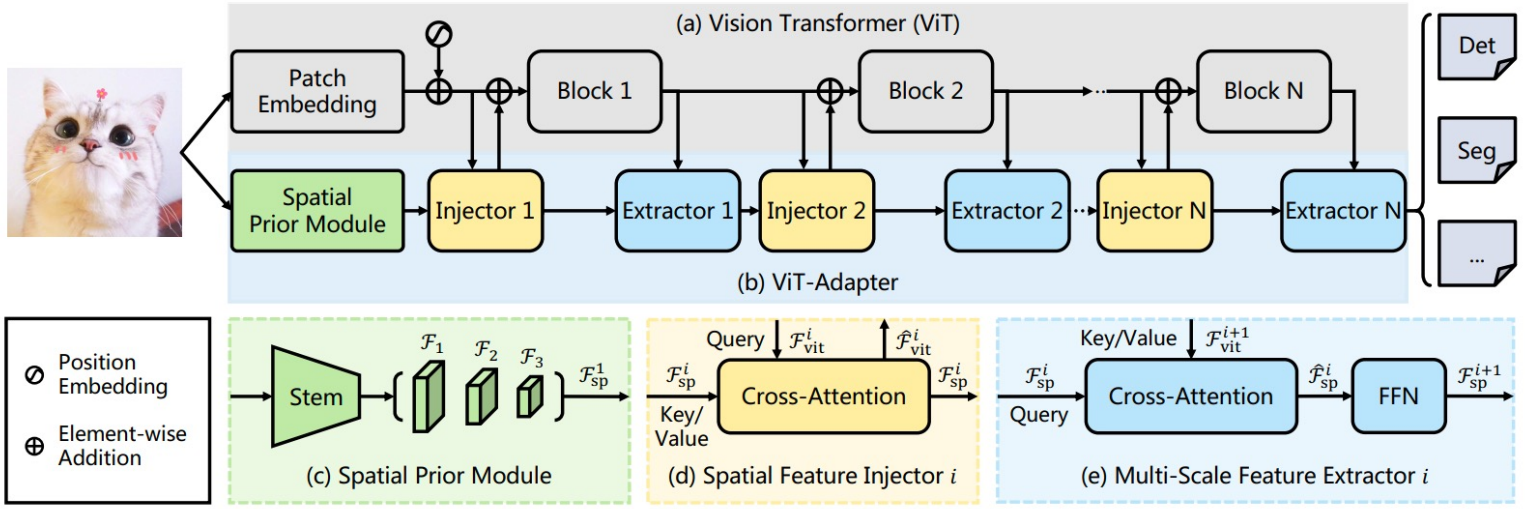
Extractor

提取多尺度的特征

$$\mathcal{F}_{\text{sp}}^{i+1} = \hat{\mathcal{F}}_{\text{sp}}^{i} + \text{FFN}(\text{norm}(\hat{\mathcal{F}}_{\text{sp}}^{i})),$$

$$\hat{\mathcal{F}}_{\text{sp}}^{i} = \mathcal{F}_{\text{sp}}^{i} + \text{Attention}(\text{norm}(\mathcal{F}_{\text{sp}}^{i}), \text{norm}(\mathcal{F}_{\text{vit}}^{i+1}))$$

Adapter

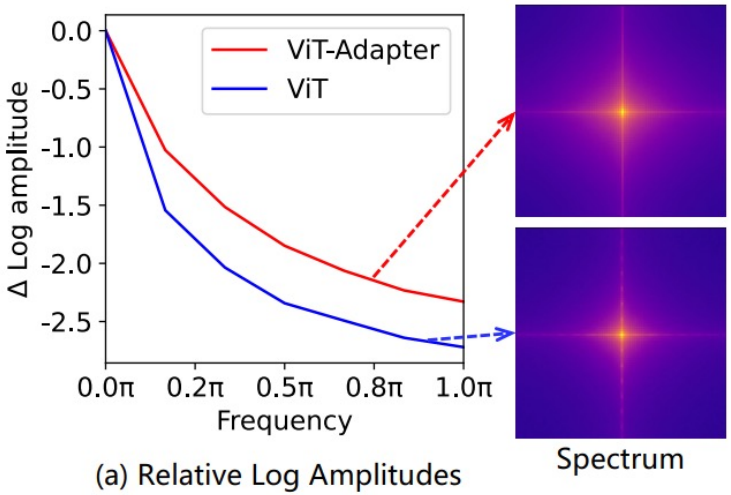


Extractor

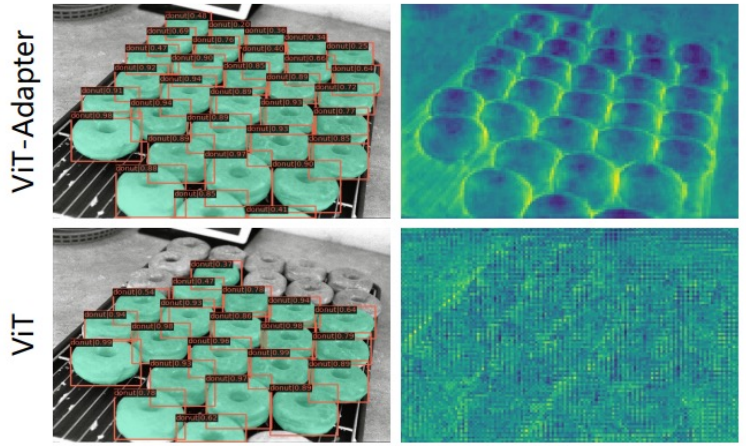
```python
class DWConv(nn.Module):
    def __init__(self, dim=768):
        super().__init__()
        self.dwconv = nn.Conv2d(dim, dim, 3, 1, 1, bias=True, groups=dim)

    def forward(self, x, H, W):
        B, N, C = x.shape
        n = N // 21
        x1 = x[:, 0:16 * n, :].transpose(1, 2).view(B, C, H * 2, W * 2).contiguous()
        x2 = x[:, 16 * n:20 * n, :].transpose(1, 2).view(B, C, H, W).contiguous()
        x3 = x[:, 20 * n:, :].transpose(1, 2).view(B, C, H // 2, W // 2).contiguous()
        x1 = self.dwconv(x1).flatten(2).transpose(1, 2)
        x2 = self.dwconv(x2).flatten(2).transpose(1, 2)
        x3 = self.dwconv(x3).flatten(2).transpose(1, 2)
        x = torch.cat([x1, x2, x3], dim=1)
        return x
```

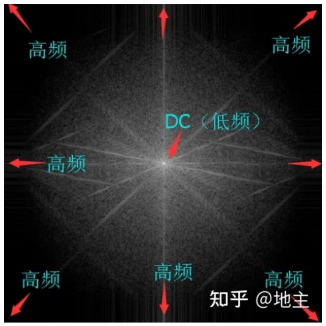
Adapter



(a) Relative Log Amplitudes



(b) Detection Results (c) Stride-8 Feature



ViT提取低频全局信息，CNN则提取高频信息（细节），Adapter将CNN的这种能力赋予了ViT。

Distillation

CVPR2022:

Co-advise: Cross Inductive Bias Distillation

Sucheng Ren[1,5] Zhengqi Gao[2] Tianyu Hua[3,5] Zihui Xue[4] Yonglong Tian[2]
Shengfeng He[1*] Hang Zhao[3,5*]
[1]South China University of Technology [2]Massachusetts Institute of Technology
[3]Tsinghua University [4]The University of Texas at Austin [5]Shanghai Qi Zhi Institute

1. 发现了：特定token与teacher对齐会有助于student的学习。
2. 发现了：影响学生模型性能的关键不是教师模型的性能而是教师模型蕴含的归纳偏置；

Distillation

介绍了两种教师模型：

Model	ImageNet(%)	A (%) ↑	R(%) ↑	C(mCE) ↓
Convolution				
ResNet-18	68.74	2.60	31.90	65.58
ResNet-34	72.62	3.45	35.17	60.26
ResNet-50	75.57	2.60	35.61	59.15
ResNet-101	77.00	6.03	38.77	54.33
ResNet-152	77.96	7.73	40.72	53.18
Involution				
RedNet-26	75.19	5.49	33.33	61.09
RedNet-38	76.88	6.88	34.80	58.15
RedNet-50	77.72	7.64	35.72	56.03
RedNet-101	78.35	9.03	36.30	54.78
RedNet-152	78.54	9.24	36.84	53.58

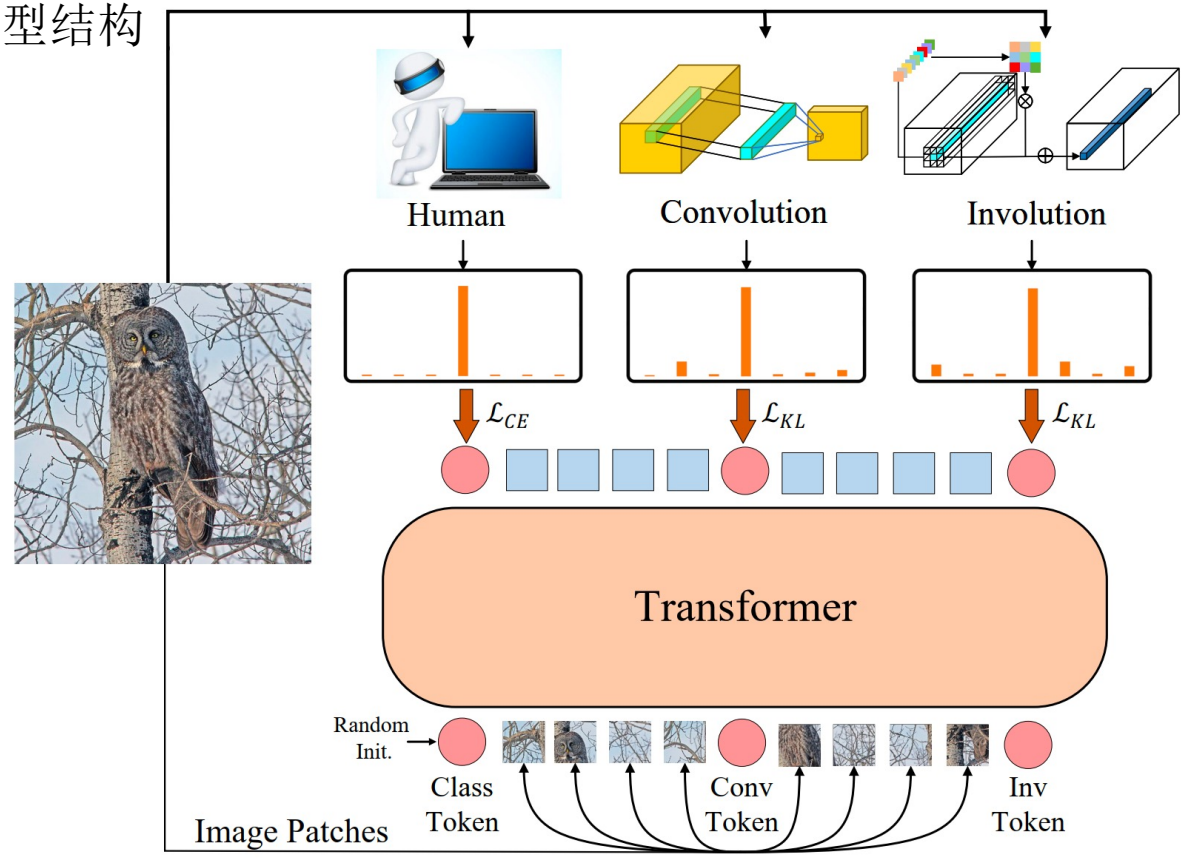
A、R、C：对ImageNet进行扰动产生的3个数据集

Acc mCE

对比ResNet50和RedNet26:
在ImageNet上有相似的性能，但在其中特定的数据集上缺有着不同的表现，因此得出结论：两种模型关注不同的特征模式，可以给学生带来不同的知识。

Distillation

模型结构



什么是对齐？
分类token-GT
卷积token-CNN教师
逆卷积token-INN教师

特定token与特定标签进行优化

token初始化方式会影响其学习能力，因此：
Cls token：高斯初始化
Conv token：卷积特征图+平均池化
Inv token：反卷积特征图+平均池化

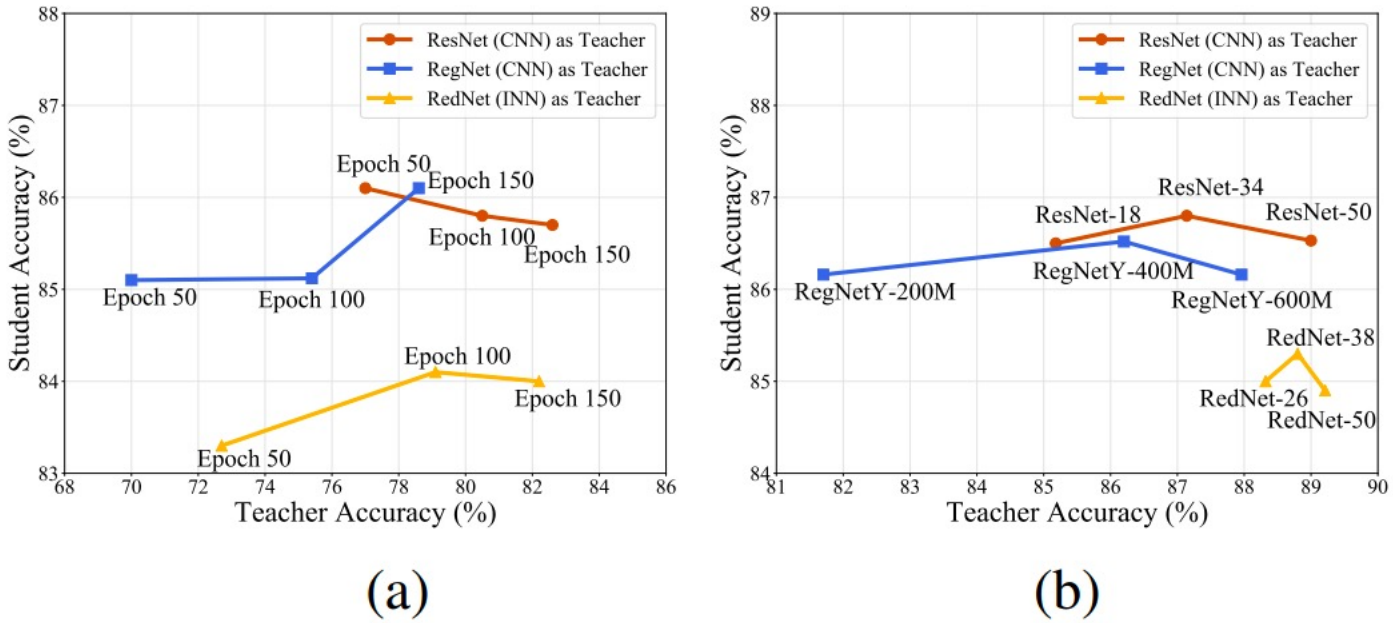
```
x = torch.cat((cls_tokens, conv_token, inv_token, x_conv+x_inv), dim=1)
```

Distillation

Student		Teacher		Top-1 (%)
Model	Token	ResNet-18	RedNet-26	
Transformer-Ti	1			81.8
Transformer-Ti	1	✓		81.9
Transformer-Ti	1		✓	80.7
Transformer-Ti	1	✓	✓	83.5
Transformer-Ti (Ours)	3	✓	✓	88.0

证明了单独token同时完成分类、蒸馏会在一定程度上有阻碍作用。

Distillation



1. 图a和图b中教师模型的性能得到显著提升，但是学生模型的性能变化不大，这说明教师模型的准确性不是影响学生模型性能的主要因素；
2. 具有相同性能的不同教师给学生带来的性能提升是不同的，这说明教师的内在学习模式对学生的表现有很大的影响。

Distillation

Model		ImageNet ↑	A ↑	R↑	C ↓
Random w/o KD	Conv Token	79.80	18.36	42.35	41.36
	Inv Token	79.80	18.35	42.35	41.35
Random w/ KD	Conv Token	81.43	16.18	45.08	39.58
	Inv Token	81.89	18.80	44.43	40.95
Align w/o KD	Conv Token	81.72	24.89	41.88	38.54
	Inv Token	81.74	24.88	41.76	38.56
Align w/ KD	Conv Token	82.11	23.58	47.41	38.11
	Inv Token	82.51	25.15	46.81	38.04

Distillation

Student	Teacher			Top-1 (%)
	ResNet-18	ResNet-50	RedNet-26	
ResNet-18				85.1
ResNet-50				89.0
RedNet-26				89.2
Transformer-Ti				81.8
Transformer-Ti	✓			86.5
Transformer-Ti		✓		86.6
Transformer-Ti			✓	85.0
Transformer-Ti	✓✓			87.2
Transformer-Ti	✓	✓		87.0
Transformer-Ti (Ours)	✓		✓	88.0

含有不同归纳偏置的教师模型共同辅导学生学习会带来更大的性能提升。

Distillation

Student	Teacher		Top-1 (%)
	ResNet-18	RegNet-26	
ResNet-10			81.5
ResNet-10	✓		83.0
ResNet-10		✓	82.6
ResNet-10	✓	✓	83.4
Mixer-Ti			80.5
Mixer-Ti	✓		81.6
Mixer-Ti		✓	80.9
Mixer-Ti	✓	✓	82.3
Transformer-Ti			81.8
Transformer-Ti	✓		86.5
Transformer-Ti		✓	85.0
Transformer-Ti (Ours)	✓	✓	88.0

Student	ResNet-18	RedNet-26	Top-1 (%)
ResNet-10	0.261	0.274	83.4
Mixer-Ti	0.358	0.313	82.3
CiT-Ti conv token	0.255	0.290	87.1
CiT-Ti inv token	0.254	0.154	87.7

两个猜想：

1. 模型本身需要有少量的归纳偏置，以免学习时步入极端；
2. 模型本身需要有一定的能力和模型容量，以免阻碍学习的程度；

1. ResNet10:本身有很强的归纳偏置，与INN的偏置存在一些冲突；
2. Mixer：Pure MLP，对应归纳偏置极少的模型；
3. Transformer：归纳偏置少，注意力层不仅执行卷积同时也与逆卷积有密切关系；