arXiv:2303.02331v2 [cs.CV] 1 Jul 2024

Trai利用延迟空间合并实现视觉 Transformer 的免训练加速

Jung Hwan Heo [1] **Seyedarmin Azizi** [1] **Arash Fayyazi** [1] **Massoud Pedram** [1]

摘要

令牌合并已成为一种新范式，能够在无需任何重新训练或微调的情况下加速视觉 Transformer（ViTs）的推理。为推进视觉 Transformer 中**免训练**加速的前沿，我们通过引入以下视角来改进令牌合并：1）激活异常值；2）分层表示。通过仔细分析视觉 Transformer 中的注意力行为，我们刻画了收敛注意力现象的延迟出现，这使得令牌合并在视觉 Transformer 的底部块中并不理想。此外，我们利用分层处理方案增强令牌合并，以捕获视觉令牌之间的多尺度冗余。结合这两项洞见，我们构建了一个统一的推理框架，称为**DSM**：**延迟空间合并**。我们在各种视觉 Transformer 模型规模（Tiny 至 Huge）和任务（ImageNet-1k 及迁移学习）上对 DSM 进行了广泛评估，在可忽略不计的损失下，最高实现了 1.8× 浮点运算次数减少和 1.6×吞吐量加速，并且比现有方法快两个数量级。

1. 引言

Transformer（Vaswani 等人，2017年）已成为一种通用骨干架构，推动了语言建模（Devlin 等人，2019年）、语音识别（Tian 等人，2020年）以及计算机视觉（Dosovitskiy 等人，2020年）领域的巨大进步。与卷积神经网络（CNN）相比，视觉 Transformer（ViT）具有较少的归纳偏置，能够从大规模预训练中受益。现代自监督模型（如掩码自编码器）在 ImageNet-1k 上的 Top-1准确率最高可达 90.94%（Wortsman 等人，2022）。

然而，由于模型规模庞大，视觉 Transformer 的高效部署仍然面临挑战。主要的一类工作致力于根据各种重要性指标剪除任务无关令牌，这些指标包括令牌嵌入（Yin 等人，2022）、注意力分数（Liang 等人，2022）以及轻量级神经网络预测器（Rao 等人，2021）。最近，一种新提出的令牌合并方案实现了无需训练的令牌缩减方法（Bolya 等人，2023年）。虽然先前工作能够有效加速视觉 Transformer 推理，但在实践中使用这些技术仍然面临挑战。以往的方法需要从头训练（Liang 等人，2022）、使用额外参数进行微调（Rao 等人，2021），或通过额外的损失函数进行优化，从而延长实际训练时间（Yin 等人，2022）。这些复杂性增加了计算预算和工程投入，阻碍了相关技术的轻松应用。令牌合并方案可以通过无需训练的模式避免这些问题（Bolya 等人，2023年），但会造成不可忽视的准确率损失，最终仍需从头训练才能达到有竞争力的性能。

为进一步推动通过令牌合并实现的无需训练的视觉 Transformer 加速，我们借鉴了近期关于大型 Transformer 中激活离群值的研究（Darcet 等人，2023年；Xiao 等人，2023年），以及一种有原则的分层处理技术（Jarrett 等人，2009；Lee 等人，2009；Krizhevsky 等人，2009）。结合延迟合并和分层合并方案，DSM 形成了一种强大的令牌合并技术，兼顾 Transformer 注意力机制和多尺度冗余。我们的贡献总结如下：

- 我们发现，近期在视觉 Transformer 中发现的高范数令牌离群值（Darcet 等人，2023年）源于语言模型中的注意力汇聚行为（Xiao 等人，2023年）。通过仔细研究视觉 Transformer 中的注意力机制，我们发现了一种有趣的现象，并将其称为延迟收敛注意力。

- 受以下观察启发：1）在底层 Transformer 块中进行令牌合并并不可取；2）分层图像处理能够捕捉多尺度交互。我们提出了一种名为延迟空间合并（DSM）的统一推理框架。

- 我们在 ImageNet-1k 和迁移学习任务上，对不同规模（Tiny ∼ Huge）的视觉Transformer 和 DeiT 模型进行了全面评估。在准确率下降不超过 1% 的情况下，我们的框架在 NVIDIA A6000 GPU 上实现了 1.8× 浮点运算次数减少和 1.6× 加速。

[1]南加州大学（USC）。通讯作者：Jung Hwan Heo < jhjohnheo@gmail.com>。

发表于 *ICML 2024 2nd* 高效基础模型系统研讨会（*ES-FoMo*），奥地利维也纳。PMLR 235，2024。版权所有 © 2024，作者所有。

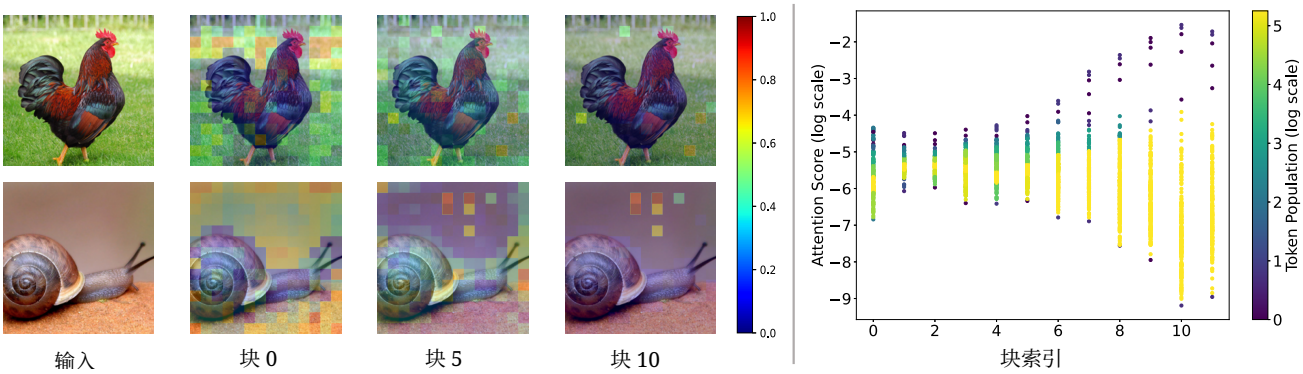


图*1*。 **ViT-S上高范数激活异常值与注意力汇聚之间的联系**。 *左*： 低信息背景令牌逐渐汇聚了大部分注意力分数。*右*： 此类离群标记获得的注意力值高出数个数量级。值得注意的是，我们观察到了离群标记开始出现的延迟，这促使我们进一步研究注意力行为。

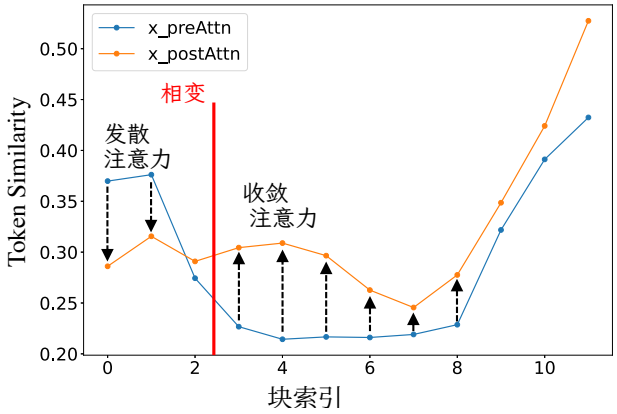


图*2*。**延迟收敛注意力示意图**。 前几个注意力块的相似度逐渐降低（使用公式 2 在 DeiT-S 上计算），随后在网络的其余部分逐渐升高。当令牌开始变得相似（收敛）时对其进行合并是理想的做法，这促成了延迟合并方案。

2. 延迟空间合并

追踪视觉 Transformer 中的注意力汇聚。 近期，人们在视觉 Transformer 中观察到了高范数激活异常值，它们充当寄存器，用于汇集全局信息（Darcet 等人，2023年；Bondarenko 等人，2023年）。我们从语言模型中的注意力汇聚行为（Xiao 等人，2023年）获得启发，以追踪此类离群值的来源。如图 1所示，我们首先验证视觉 Transformer 中的高范数离群令牌与注意力汇聚行为有关。尽管注意力分数的初始分布近乎均匀，但它们会逐渐集中到少数几个背景令牌上，导致离群汇聚令牌的分数与其他令牌的分数相差数个数量级。有趣的是，我们观察到，在注意力汇聚开始显现之前存在一段初始延迟。这自然引出了两个问题：这种延迟为何存在？它又会如何影响令牌合并？

2.1. 延迟合并

基础令牌合并。 视觉 Transformer 中的 Transformer 块由一个多头注意力（MHA）层和一个前馈网络（FFN）层组成。对于深度为 L的网络中的第 l个 Transformer 块，前向传播表示为

$$\bar{\mathbf{X}}^l = \mathbf{X}^l + \mathrm{MHA}(\mathbf{X}^l), \mathbf{X}^{l+1} = \bar{\mathbf{X}}^l + \mathrm{FFN}(\bar{\mathbf{X}}^l), \quad (1)$$

其中，$\mathbf{X}^l \in \mathbb{R}^{N\times C}$ 是包含 N 个令牌的输入序列，每个令牌的嵌入维度为 C。令牌合并应用于每个 Transformer块中 MHA 与前馈网络之间的位置。给定一个包含 n 个令牌的序列（MHA 层输出），记为 $\bar{\mathbf{X}}^l = [x_1, ..., x_n]$，构造一个加权完全二部图，该图由两组节点（令牌）组成：$\mathbb{A} = [x_1, x_3, ..., x_{n-1}]$ 和 $\mathbb{B} = [x_2, x_4, ..., x_n]$。令牌 $a \in \mathbb{A}$ 与令牌 $b \in \mathbb{B}$之间的边表示 a 与b的嵌入之间的余弦相似度。随后应用加权二部图匹配算法，以确定具有最大权重和的 $r \leq n/2$ 条边集合。与这些 r 条边关联的令牌通过按通道加权平均进行合并。最后，将两组 $\mathbb{A}$ 和 $\mathbb{B}$ 合并，得到一个包含 r 个令牌的截断令牌序列。

表征收敛注意力。现在，我们研究注意力汇聚的延迟如何影响令牌相似度分布。直观来说，理想的令牌合并场景是令牌彼此最为相似时（即避免在令牌彼此不相似时强行合并）。为量化令牌之间的相似程度，我们采用了文本生成中广泛使用的令牌相似度指标（Zhang 等人，2019年）：

$$Sim = \frac{1}{n\,(n-1)} \sum_{i\neq j} \frac{x_i^T x_j}{||x_i||_2 ||x_j||_2}, \quad (2)$$

更高的 Sim 分数表示该层中的令牌

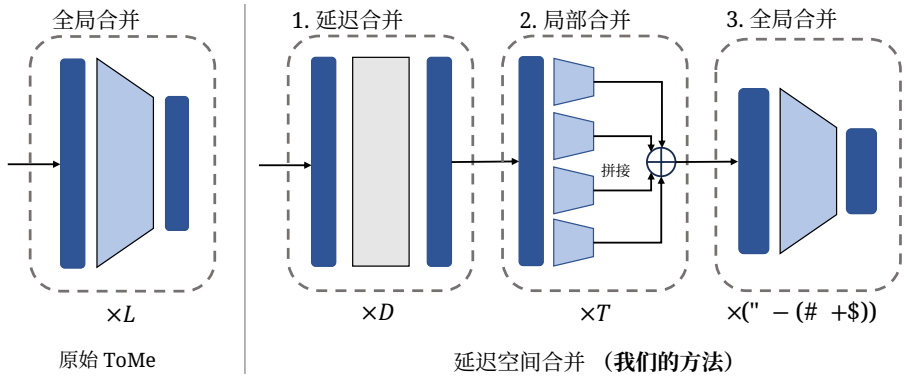


图 *3.* **左：** Bolya 等人提出的原始令牌合并（ToMe）Bolya 等人会对所有 L Transformer 块中的所有令牌进行全局合并。 **右：** 受收敛注意力和空间感知原则的启发，我们提出的延迟空间合并（DSM）对 ToMe 进行了扩展：在最初的 D 个块中不进行合并，在 T 个块中进行局部合并，然后在网络的其余部分进行全局合并。

具有相似的嵌入。有趣的是，如图2所示，初始块中的令牌在注意力过程中变得不那么相似（发散注意力），而在视觉Transformer中的某个点之后（从块 2 开始发生相变），令牌会持续变得更加相似（收敛注意力）。合并正在分化的令牌会适得其反，因此我们将合并延迟到令牌嵌入稳定下来，呈现出收敛注意力行为之后。

2.2. 空间合并

分层图像处理是一项基础技术，广泛应用于计算机视觉建模的诸多领域，从语义分割（Long 等人，2015年）、目标检测（Jarrett 等人，2009；Lin 等人，2017年），到基于神经辐射场的三维渲染（Barron 等人，2021）。我们首次将分层表示原则引入令牌合并。其核心思路是捕捉视觉令牌之间的多尺度交互，从而能够以更细的粒度执行相似性（特征冗余）搜索。

图像中彼此语义关系更强的相邻像素；例如，动物图片中的身体部位彼此相连，此时空间邻近性与语义相似性高度相关。我们不进行全局相似令牌搜索，而是将搜索空间限制在局部窗口内。输入令牌可表示为尺寸为 (H, W) 的二维网格，并将其划分为四个尺寸相等、维度为 w的窗口。为降低复杂度，我们在所有实验中将初始窗口大小设为 $w = 7$，因为它可以很好地整除包含 14×14 个令牌的网格（分辨率为 224×224，采用常见的 16 补丁大小）。当令牌数量不能被 w整除时，我们会在右下方进行填充，以保留二维结构。

我们不是采用固定的窗口大小 w，而是在每个块中逐步增大窗口大小。这一做法基于以下直觉：由于位置嵌入是在块 0 之前添加的，因此位置相似性在网络的较早阶段最为相关。因此，我们逐块增大窗口，直到窗口等效于全局合并（此时窗口与剩余令牌二维网格一样大）。窗口可以堆叠起来，以高效地并行合并令牌。由于令牌合并是针对批次中的每个样本独立执行的，因此可以将窗口维度融合到批次维度中：B: (B, H, W) $\rightarrow$ (B, H/w, w, W/w, w) $\rightarrow$ (B* H/w * W/w,w, w)。窗口堆叠操作的高效内核实现是可行的，正如（Liu 等人，2021）所示。

2.3. 统一推理框架

如图 3所示，DSM 在基础词元合并技术中加入了延迟合并和局部合并。对于深度为 L 的网络，我们延迟 D 个块，再以窗口大小 w对 T 个块进行局部合并，并对网络其余部分执行全局合并。我们调整的唯一超参数是 r，即单个令牌合并层中要减少的令牌数量；超参数设置将在第 B 节中进一步讨论。

表*1*。**与先前工作的比较**。我们的框架在性能上具有竞争力，速度则快两个数量级。端到端训练时间是在单个 8 GPU 节点上测得的。

	Top-1	GFLOP	训练轮数	端到端（小时）
DeiT-S	79.8	4.6	0	0
DynamicViT（Rao 等人，2021）	79.3	2.9	30	44.8
SPViT（Kong 等人，2021）	79.3	2.6	60	–
A-ViT（Yin 等人，2022）	78.6	2.9	100	76.4
E-ViT（Liang 等人，2022）	79.1	2.6	300	154.4
ATS (Fayyaz 等人，2022年)	79.7	2.9	30	–
ToMe(Bolya 等人，2023年)	79.4	2.7	300	102.2
空间合并**(我们的方法)**	79.3	2.8	0	0
DSM **(我们的方法)**	78.6	2.5	0	0

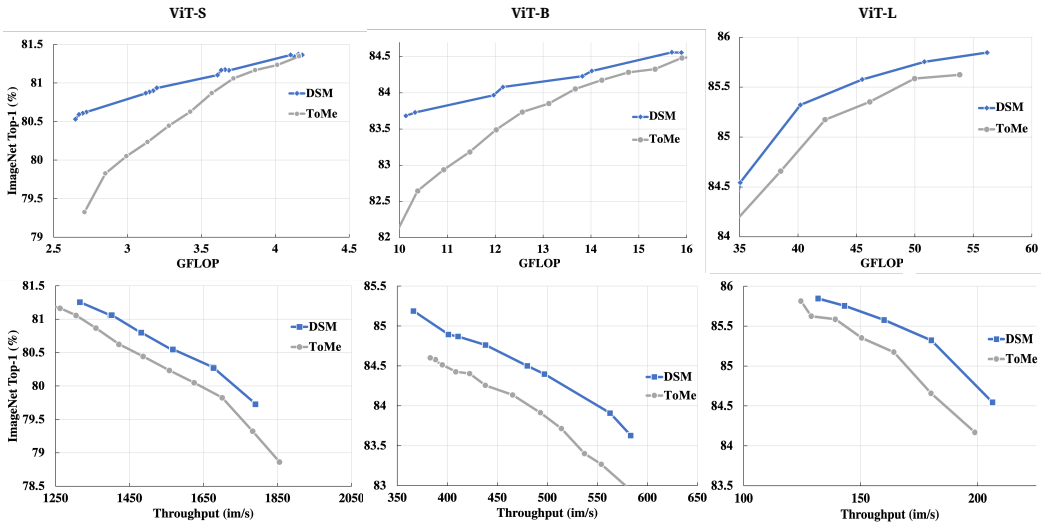


图*4*。 **模型扫描**。 我们以无需训练 的方式，将我们的推理框架应用于多个最先进的视觉Transformer模型。DSM 的超参数由网络架构固定，仅改变参数r，以在 ImageNet-1k 上生成 Top-1 准确率与 GFLOP 的曲线。

	Oxford-IIIT Pet			Flowers102			FGVC-Aircraft			CIFAR-100		
	Top-1 准确率	GFLOP	im/s	Top-1 准确率	GFLOP	im/s	Top-1 准确率	GFLOP	im/s	Top-1 准确率	GFLOP	im/s
基线	92.12	17.57	404.53	98.13	17.57	407.27	81.12	17.57	408.88	91.10	17.57	402.51
$r = 4$	92.04	16.10	384.51	98.19	16.10	384.54	80.98	16.10	388.25	90.96	16.10	383.1
$r = 8$	92.01	14.46	431.37	97.95	14.46	430.18	80.95	14.46	436.20	91.01	14.46	431.72
$r = 12$	91.74	13.27	468.78	97.69	13.27	466.17	81.07	13.27	471.10	90.90	13.27	469.10
$r = 16$	91.55	11.96	517.05	97.45	11.96	513.75	80.38	11.96	518.86	90.76	11.96	517.76
$r = 20$	91.32	10.16	612.50	97.14	10.16	609.61	80.20	10.16	612.11	90.25	10.16	613.92

表 *2*。**迁移学习** 。使用 DSM 加速的微调版 ViT-B 在各种数据集上始终实现 1.5× 倍加速。

3. 实验

我们在 ImageNet-1K（Russakovsky 等人，2015年）上开展实验，以评估我们的方法在分类任务中加速现成视觉 Transformer 的效果。我们使用 DeiT（Touvron 等人，2021）以及采用 AugReg 训练的视觉 Transformer 模型（Steiner 等人，2021），以检验我们的方法在不同骨干网络和训练方法上的泛化能力。我们使用 Torchprofiler[1] 库测量浮点运算次数。推理吞吐量在 Nvidia RTX A6000 GPU 上测量，批量大小固定为 32，并对 50 次运行取平均值。

如表1所示，得益于免训练方法，我们的 DSM 在实现具有竞争力的性能的同时，速度比现有方法快两个数量级。例如，E-ViT（Liang 等人，2022）单次运行约需 154 个单 GPU 小时。由于该方法需要针对每种目标加速比都运行一次，因此在不同资源限制下部署的成本很快就会变得难以承受。

在图4中，我们将框架应用于现成的视觉Transformer-[S、B、L]，并关闭

[1]https://github.com/zhijian-liu/torchprofile

架，分辨率为 224px，补丁大小为 16。对于每个模型，我们将 DSM 与 ToMe 进行基准测试。我们改变 r，以构建两条帕累托曲线，比较 Top-1准确率与乘加运算次数和吞吐量。请注意，我们使用 DSM 时会扫描更高的 r 值，以匹配 ToMe 的计算负载。

我们可以看到，我们的框架始终比 ToMe 表现更好，尤其是在较小的模型上。值得注意的是，对于 ViT-S 和 ViT-B，我们分别能在准确率损失不超过 1% 的情况下节省 45% 和 42% 的浮点运算次数。与原始 ToMe 相比，该方法能将准确率提高 1% 以上。然而，DSM 的效果与模型规模呈反比，在 ViT-L 上几乎没有提升。与 DSM 在节省浮点运算次数方面的显著效果相比，其吞吐量提升相对有限。我们认为，这是因为额外的数据移动（例如排序、填充和修改张量维度）带来了不可忽视的开销。有趣的是，这种开销在较大的模型上不那么明显，因为 DSM 曲线向右移动，呈现出更好的权衡空间。对于执行大量矩阵乘法的大型模型而言，计算而非数据移动会成为瓶颈。因此，局部合并所带来的内存输入输出开销在较大模型上不太明显。

参考文献

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., and Srinivasan, P. P. Mip-nerf：一种用于抗锯齿神经辐射场的多尺度表示。载于*IEEE/CVF*国际计算机视觉会议论文集，第 5855–5864 页，2021。

Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., and Kolesnikov, A. 知识蒸馏：优秀的教师耐心且始终如一。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10925–10934 页，2022。

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. 令牌合并：让你的视觉Transformer运行得更快。载于国际学习表征会议，2023年。

Bondarenko, Y., Nagel, M., and Blankevoort, T. 可量化的 Transformer：通过让注意力头不执行任何操作来消除离群值。*arXiv* 预印本 *arXiv:2306.12929*，2023年。

Darcet, T., Oquab, M., Mairal, J., andBojanowski, P. 视觉变换器需要寄存器。*arXiv* 预印本 *arXiv:2309.16588*，2023年。

Dettmers, T., Lewis, M., Belkada, Y. 和 Zettlemoyer, L. Llm. int8（）：面向大规模 Transformer 的 8 位矩阵乘法。*arXiv* 预印本 *arXiv:2208.07339*，2022。

Devlin, J. 等。BERT：面向语言理解的深度双向变换器预训练。载于北美计算语言学协会分会会议：人类语言技术。计算语言学协会，2019年。

Dosovitskiy, A.、Beyer, L.、Kolesnikov, A.、Weissenborn, D.、Zhai, X.、Unterthiner, T.、Dehghani, M.、Minderer, M.、Heigold, G.、Gelly, S. 等。一张图像胜过 16×16 个词：面向大规模图像识别的 Transformer。*arXiv* 预印本 *arXiv:2010.11929*，2020年。

Fayyaz, M.、Koohpayegani, S. A.、Jafari, F. R.、Sengupta, S.、Joze, H.R. V.、Sommerlade, E.、Pirsiavash, H. 和 Gall, J. 用于高效视觉变换器的自适应令牌采样。载于 欧洲计算机视觉会议，第 396–414 页。施普林格，2022。

Han, S.、Mao, H. 和 Dally, W. J. 深度压缩：通过剪枝、训练量化和霍夫曼编码压缩深度神经网络。*arXiv* 预印本 *arXiv:1510.00149*，2015年。

He, K.、Zhang, X.、Ren, S. 和 Sun, J. 用于图像识别的深度残差学习。载于*IEEE* 计算机视觉与模式识别会议论文集，第 770–778 页，2016年。

Heo, J. H., Kim, J., Kwon, B., Kim, B., Kwon, S. J[illegible]d Lee, D. 重新思考通道维度，以分离用于大语言模型低比特权重量化的离群值。*arXiv* 预印本 *arXiv:2309.15531*，2023年。

Hinton, G. E.、Vinyals, O. 和 Dean, J. 提炼神经网络中的知识。*CoRR*，abs/1503.02531，2015年。

Jarrett, K., Kavukcuoglu, K., Ranzato, M. 和 LeCun, Y. 用于目标识别的最佳多阶段架构是什么？见*2009*年*IEEE*第*12*届国际计算机视觉会议，第2146–2153页。IEEE，2009。

Kim, S., Gholami, A., Yao, Z., Mahoney, M. W. 和 Keutzer, K. I-BERT：纯整数BERT量化。见 Meila, M. 和 Zhang, T.（编），第*38*届国际机器学习会议论文集，*ICML 2021*，*2021*年*7*月*18*日至*24*日，线上活动，第139卷，收录于机器学习研究论文集，第5506–5518页。PMLR，2021。网址：http://proceedings.mlr.press/v139/kim21d.html。

Kong, Z., Dong, P., Ma, X., Meng, X., Niu, W., Sun, M., Ren, B., Qin, M., Tang, H. 和 Wang, Y. SPViT：通过软令牌剪枝加速视觉变换器。*arXiv*预印本 *arXiv:2112.13890*，2021。

Kong, Z., Dong, P., Ma, X., Meng, X., Niu, W., Sun, M., Shen, X., Yuan, G., Ren, B., Tang, H., Qin, M. 和 Wang, Y. SPViT：通过延迟感知软令牌剪枝加速视觉变换器。载于 Avidan, S., Brostow, G., Ciss´e, M., Farinella, G. M. 和 Hassner, T.（编），计算机视觉 – *ECCV 2022*，第 620–640 页，Cham，2022。施普林格·自然瑞士。

Krizhevsky, A., Hinton, G. 等。从小型图像中学习多层特征。2009。

Kurtic, E., Campos, D., Nguyen, T., Frantar, E., Kurtz, M., Fineran, B., Goin, M. 和 Alistarh, D. TheoptimalBERT surgeon：面向大语言模型的可扩展且准确的二阶剪枝。*CoRR*，abs/2203.07259，2022。doi：10.48550/arXiv.2203.07259。URLhttps://doi.org/10.48550/arXiv.2203.07259。

Langley, P. 撰写机器学习论文。载于 Langley, P.（编），第*17*届国际机器学习会议论文集（*ICML 2000*），第1207–1216页，美国加利福尼亚州斯坦福，2000年。Morgan Kaufmann。

Lee, H., Grosse, R., Ranganath, R., 和 Ng, A. Y. 用于可扩展分层表示无监督学习的卷积深度置信网络。载于第*26*届年度国际机器学习会议论文集，第609–616页，2009。

Liang, W., Yuan, Y., Ding, H., Luo, X., Lin, W., Jia, D., Zhang, Z., Zhang, C., 和 Hu, H. 无需微调即可加速用于密集预测的大规模视觉变换器。神经信息处理系统进展，35:35462–35477，2022。

Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., and Han, S. AWQ：用于大语言模型压缩和加速的激活感知权重量化。*arXiv*预印本 *arXiv:2306.00978*，2023年。

Lin, T.-Y., Doll´ar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. 用于目标检测的特征金字塔网络。发表于*IEEE*计算机视觉与模式识别会议论文集，第2117–2125页，2017年。

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin变换器：采用移位窗口的层次化视觉变换器。发表于*IEEE/CVF*国际计算机视觉会议论文集，第10012–10022页，2021。

Long, J.、Shelhamer, E. 和 Darrell, T. 用于语义分割的全卷积网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 3431–3440 页，2015年。

Marin, D.、Chang, J. R.、Ranjan, A.、Prabhu, A.、Rastegari, M. 和 Tuzel, O. 视觉变换器中的令牌池化。*CoRR*，abs/2110.03860，2021。网址：https://arxiv.org/abs/2110.03860。

Miller, E. 注意力偏移一位。2023年。网址：https://www.evanmiller.org/attention-is-off-by-one.html。

Radosavovic, I.、Kosaraju, R. P.、Girshick, R.、He, K. 和 Doll´ar, P. 设计网络设计空间。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10428–10436 页，2020年。

Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., and Hsieh, C.-J. Dynamicvit：高效视觉变换器与动态令牌稀疏化。神经信息处理系统进展，34:13937–13949，2021。

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 等。ImageNet大规模视觉识别挑战赛。国际计算机视觉杂志，115: 211–252，2015年。

Ryoo, M. S., Piergiovanni, A. J., Arnab, A., Dehghani, M., 和 Angelova, A. TokenLearner：面向视频的自适应时空标记化。载于 Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., 和 Vaughan, J. W.（编），神经信息处理系统进展第*34*卷：*2021*年神经信息处理系统大会，*NeurIPS 2021*，*2021*年*12*月*6*日至*14*日，线上举行，第12786–12797页，2021年。网址：https://proceedings.neurips.cc/paper/2021/hash/6a30e32e56fce5cf381895dfe6ca7b6f-Abstract.html。

Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M. W., and Keutzer, K. Q-BERT：基于海森矩阵的 BERT 超低精度量化。载于 第三十四届人工智能促进协会人工智能会议，*AAAI 2020*，第三十二届人工智能创新应用会议，*IAAI 2020*，第十届*AAAI*人工智能教育进展研讨会，*EAAI 2020*，美国纽约州纽约市，*2020*年*2*月7日至*12*日，第8815–8821页。AAAI出版社，2020年。网址：https://ojs.aaai.org/index.php/AAAI/article/view/6409。

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J. 和 Beyer, L. 如何训练你的视觉Transformer？视觉Transformer中的数据、增强和正则化。*CoRR*，abs/2106.10270，2021年。网址：https://arxiv.org/abs/2106.10270。

Tan, M. 和 Le, Q. EfficientNet：重新思考卷积神经网络的模型缩放。载于国际机器学习会议，第6105–6114页。PMLR，2019年。

Tian, Z., Yi, J., Bai, Y., Tao, J., Zhang, S. 和 Wen, Z. 用于端到端语音识别的同步变换器。载于*2020*年*IEEE*国际声学、语音与信号处理会议，*ICASSP 2020*，西班牙巴塞罗那，*2020*年*5*月*4*日至*8*日，第7884–7888页。IEEE，2020年。doi：10.1109/ICASSP40776.2020. 9054260。网址：https://doi.org/10.1109/ICASSP40776.2020.9054260。

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., 和 J´egou, H. 通过注意力训练数据高效的图像变换器并进行蒸馏。载于 Meila, M. 和 Zhang, T.（编），第*38*届国际机器学习会议论文集，*ICML 2021*，*2021*年*7*月*18*日至*24*日，线上活动，第139卷，机器学习研究论文集，第10347–10357页。PMLR，2021。网址：http://proceedings.mlr.press/v139/touvron21a.html。

Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, A. N.、Kaiser, Ł. 和 Polosukhin, I. 注意力就是你所需要的一切。神经信息处理系统进展，30，2017年。

Voita, E.、Talbot, D.、Moiseev, F.、Sennrich, R. 和 Titov, I. 分析多头自注意力：专门化的注意力头承担了主要工作，其余部分可以进行剪枝。在 Korhonen, A.、Traum, D. R. 和 M`arquez, L.（编），计算语言学协会第 *57* 届会议论文集，*ACL 2019*，意大利佛罗伦萨，*2019*年 *7* 月 *28* 日至 *8* 月 *2* 日，第 *1* 卷：长篇论文，第 5797–5808 页。计算语言学协会，2019年。doi: 10.18653/v1/p19-1580。URL: https://doi.org/10.18653/v1/p19-1580。

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., 和 Schmidt, L. 模型融合：对多个微调模型的权重进行平均，可在不增加推理时间的情况下提高准确率。载于 Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., 和 Sabato, S.（编），第*39*届国际机器学习大会论文集，第162卷，机器学习研究论文集，第23965–23998页。PMLR，2022年7月17–23日。网址: https://proceedings.mlr.press/v162/wortsman22a.html。

Xiao, G., Lin, J., Seznec, M., Demouth, J., 和 Han, S. Smoothquant：面向大语言模型的准确高效训练后量化。*arXiv*预印本 *arXiv:2211.10438*，2022。

Xiao, G., Tian, Y., Chen, B., Han, S., 和 Lewis, M. 借助注意力汇聚实现高效的流式语言模型。*arXiv*，2023年。

Yin, H., Vahdat, A., Alvarez, J. M., Mallya, A., Kautz, J., 和 Molchanov, P. A-ViT：用于高效视觉变换器的自适应令牌。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10809–10818页，2022。

Zhang, T.、Kishore、V.、Wu、F.、Weinberger、K. Q. 和 Artzi、Y. BERTScore：使用 BERT 评估文本生成。*arXiv* 预印本 *arXiv:1904.09675*，2019年。

A. 相关工作

A.1. 高效 Transformer

在降低神经网络高昂的计算成本并实现其在资源受限环境中的高效部署方面，已经取得了显著进展。在算法层面，模型量化（Shen 等人，2020；Kim 等人，2021；Xiao 等人，2022）、模型剪枝（Han 等人，2015年；Voita 等人，2019；Kurtic 等人，2022）和知识蒸馏（Hinton 等人，2015；Beyer 等人，2022）等方法日益受到关注。与权重剪枝正交的是，令牌压缩（Yin 等，2022；Rao 等人，2021；Liang 等人，2022）表明，Transformer 的输入可以在推理时进行动态剪枝。在本文中，我们聚焦于自适应令牌压缩技术，其令牌缩减决策取决于输入图像。

A.2. 令牌压缩

令牌剪枝通过丢弃不太重要的令牌来加速 ViT 模型的推理。许多先前工作都致力于识别这类令牌冗余。例如，DynamicViT（Rao 等人，2021）使用 Gumbel-Softmax 分布训练令牌重要性预测器。A-ViT（Yin 等人，2022）通过引入一种惩罚未剪枝令牌的损失函数来学习令牌的重要性。E-ViT（Liang 等人，2022）使用 [CLS] 令牌的注意力分数作为重要性启发依据。尽管这些方法在部署后有效，但它们需要从头开始进行高成本的重新训练，或基于模型检查点进行微调。相比之下，我们的工作着重于彻底消除这些使用障碍（表 3）。

令牌合并将令牌组合起来，而非剪枝。先前的工作尝试使用自定义启发式方法（Kong 等人，2022）或可学习的多层感知机投影（如 TokenLearner（Ryoo 等人，2021）），将不重要的令牌融合为单个令牌。令牌池化也被提出作为一种通过合并实现的下采样方法（Marin 等人，2021）；然而，其基于迭代 k 均值的方法速度较慢，且与现成模型不兼容。ToMe（Bolya 等人，2023年）是近期提出的一种令牌合并模块，它采用二分图匹配算法，在无需重新训练的情况下实现了与令牌剪枝相当的准确率。我们的工作论证了令牌合并作为免训练加速首选构建模块的价值，并通过改进推动准确率-效率权衡的帕累托前沿。

A.3. 令牌离群值

为改进令牌合并技术，我们从近期发现的令牌离群问题入手。该问题出现在用于视觉与语言任务的大型 Transformer 模型中。令牌离群值最初在激活量化研究中受到关注，

表 3。不同令牌压缩技术的比较。我们的延迟空间合并（DSM）框架充分发挥了免训练加速的优势。

	预训练	微调	免训练
DynamicViT（Rao 等人，2021）	✗	✓	✗
SPViT（Kong 等人，2021）	✗	✗	✗
A-ViT（Yin 等人，2022）	✗	✓	✗
E-ViT（Liang 等人，2022）	✓	✓	✗
ATS（Fayyaz 等人，2022）	✗	✗	✗
ToMe（Bolya 等人，2023年）	✓	✗	✓†
DSM（我们的方法）	✗	✗	✓

†容易出现精度下降。

化Xiao 等人，2022；Dettmers 等人，2022；Lin 等人，2023年；Heo 等人，2023年）。同样，在有监督和无监督视觉 Transformer 中都发现了令牌离群值（Bondarenko 等人，2023年；Darcet 等人，2023年），这些令牌被描述为低信息量的背景令牌，能够汇聚全局信息（其功能类似于 [CLS]令牌）。

令牌离群值的成因可以追溯到注意力中的 Softmax 函数；注意力权重之和必须为 1（Miller，2023年；Xiao 等人，2023年）。当注意力头不希望更新残差流时，它会通过高度关注一个低信息量令牌来执行“空操作”（Bondarenko 等人，2023年）。在这项工作中，我们证实了 Softmax 函数所导致的“注意力汇聚”确实存在，并以此为基础探究视觉 Transformer 中独特的注意力行为。

A.4. 免训练加速

对于视觉 Transformer，用于分类任务的大多数模型都是 ViT 和 DeiT 模型系列中的小型（或微型）变体。以往基于训练的令牌缩减技术主要针对小型模型进行研究，因为大型模型的训练成本很高（Rao 等人，2021；Liang 等人，2022）。考虑到训练和超参数调优的成本，总计算量对许多研究人员和从业者而言可能难以承受（Steiner 等人，2021）。鉴于性能最佳的模型压缩成本过高，我们提出一**种用于压缩大型视觉 Transformer 的免训练框架**。我们将方法的速度视为与最终模型性能同等重要的评价指标，并将解决方案限定为免训练。我们的工作旨在回答以下问题：如何在无需昂贵训练的情况下压缩视觉 Transformer，从而实现高准确率的推理模型？

B. 详细方法论

DSM超参数。我们将延迟参数 D设为收敛注意力开始出现的转折点注意力行为开始显现。也就是说，对于深度为 12 的 DeiT-S 模型，我们选择 $D = 2$，因为收敛注意力在第二个块中出现（参见图2。我们在图8和图9中展示了更多网络，其中网络深度的 1/6 处通常是注意力行为从发散转为收敛的转折点。

局部合并参数 T可以设定为窗口大小 w 和缩减率 r 的函数。这是因为，随着窗口大小逐渐增大，局部合并会自然退化为全局合并。随着令牌逐步合并，序列长度会变得小于窗口大小。因此，增大窗口大小会产生部分局部合并，并平滑过渡到全局合并。

C. 更多实验

C.1. 支持合并的理由

在深入探讨 DSM 评估之前，我们首先从更根本的层面阐述为何对于免训练的视觉Transformer加速，令牌合并比令牌剪枝更适合作为构建模块。现成的视觉 Transformer模型通常是在密集令牌分布（不丢弃令牌）下预训练的。因此，训练好的模型“期望”看到的不仅是与任务相关的令牌，也包括背景令牌等相关性较低的令牌。信息量较低的令牌也可以起到正则化作用，因此，假定不太重要的令牌可以被移除而不会降低预测性能，是有风险的。因此，在免训练设置下，令牌剪枝可能并非最佳设计选择。

标准令牌压缩的启发式消融。 与表 4所示一致，我们通过比较基于各种重要性标准的剪枝与合并，实证支持令牌合并的合理性。对于剪枝，移除 L2范数最低的令牌；对于合并，合并余弦相似度分数最高的令牌。我们观察到，注意力块 X 的输出是一个出人意料的优秀剪枝启发式方法，但效果比合并方案低 2%。 K 嵌入标准在合并方面表现最佳。它对令牌的表征效果最好，甚至优于 X，尽管后者每个令牌的嵌入维度更大。这可能是过参数化嵌入导致的：更多通道可能带来噪声。由于 K 经过多头注意力的通道数较少，其紧凑的表示能够解决这一问题。

DSM 的启发式消融。我们对框架中表现最佳的相似度启发式进行了消融。如表5所示， K 是最佳选择。随机选择的准确率最低，而采用其他任何启发式都会导致更差的准确率-吞吐量权衡。

表 4。**剪枝与合并** 比较：使用 ViT-L 和 $r = 7$。在免训练设置下，合并能更有效地保留准确率。X 位于注意力块内部。

	剪枝		合并	
特征	ACC	IM/S	ACC	IM/S
随机	2.96	131.3	61.89	136.5
X	**81.58**	129.1	83.41	125.7
K	71.86	**130.7**	**83.51**	**132.7**
Q	73.65	130.3	83.25	131.3
V	78.9	130.3	83.44	132.3

表5.**设计选择**。在 ViT-L 上应用 DSM 并采用$r = 18$时，K 嵌入能够实现最佳的准确率-吞吐量权衡。

	随机	X	K	Q	V
ACC	59.6	82.7	**82.9**	82.4	82.6
IM/S	589.1	552.2	**591.5**	589.2	589.5

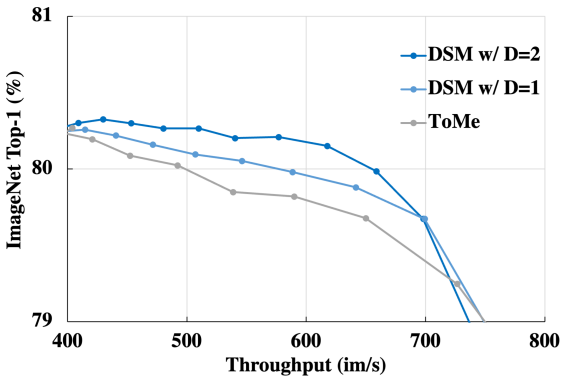


图5。**尖锐度最小化模型**在 ViT-B 上使用 SAM优化器训练后，更适合压缩。借助延迟合并（记作 D），吞吐量可提升 1.6× 。

C.2. 主要结果

我们还评估了 DSM 在各种非基于 Transformer 的视觉架构中，相对于基于训练的加速方法的效果。由于卷积具有权重共享的特性，且峰值内存占用较低（因为不使用二次复杂度的自注意力），卷积神经网络通常能更有效地利用参数。如表6所示，我们观察到，采用 DSM 的 DeiT-S 在准确率与计算量的权衡方面表现相当，可媲美通过神经架构搜索得到的 EfficientNet 等成本高得多的方法（Tan & Le，2019年）。

此外，我们还针对更大的模型（ViT-H）和更小的模型（DeiT-Ti）进行了额外实验。表7 分别比较了 DSM 与 ToMe 在 ViT-H 和 DeiT-T 上的表现。

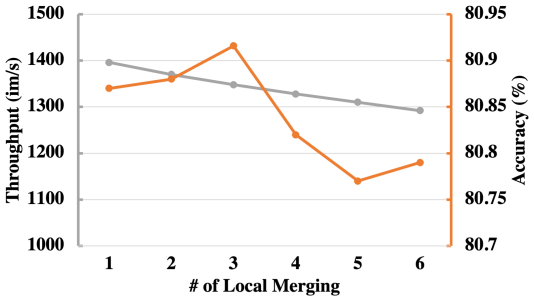


图6。局部合并对吞吐量和准确率的影响。我们观察到，局部合并的使用程度越高，由于额外的数据移动，吞吐量越低。不过，对于 $L = 3$，通常存在一个最佳平衡点；这是我们采用递增窗口技术找到的默认设置。

表6。与基于卷积的视觉架构的比较。

模型	Top-1	加速（↑）
DeiT-S	79.8	1×
EfficientNet-B2（Tan& Le，2019年）	80.1	1.33×
EfficientNet-B3（Tan& Le，2019年）	81.6	0.78×
ResNet-152（He 等人，2016年）	78.3	0.56×
RegNetY-4GF（Radosavovic 等人，2020年）	80.0	1.23×
采用 DSM 的 DeiT-S（$r16$）	79.6	1.5×
采用 DSM 的 DeiT-S（$r18$）	79.4	1.6×

表7。ViT-H 和 DeiT-T 在 ImageNet-1k 上的比较

		Δ 准确率（%）	吞吐量（图像/秒）	# 乘加运算次数（G）
ViT-H	ToMe	-0.2	50.18	145.84
	DSM	-0.2	**56.90**	**129.64**
	ToMe	-0.6	64.01	113.90
	DSM	-0.6	**72.60**	**101.79**
	ToMe	-0.8	70.38	103.36
	DSM	-0.8	**79.30**	**92.59**
DeiT-T	ToMe	-0.1	2457	1.18
	DSM	-0.1	**2722**	**1.09**
	ToMe	-0.5	3020	0.93
	DSM	-0.5	**3257**	**0.86**
	ToMe	-2.0	3881	0.69
	DSM	-2.0	**4001**	**0.71**

D. 更多可视化

在图7中，我们展示了属于最终合并令牌的输入令牌。对于 ToMe，我们使用 $r = 24$ ，而对于我们的框架，我们使用 $r = 28$和 $D = 4$ 以及 $w = 8$ 。请注意，由于分辨率更高，参数有所不同。为匹配最终令牌数量，我们的框架不会合并最后一个块。可以看到，在第二张图中，马尔济斯犬的脸被我们连续地合并为单个令牌，而 ToMe 则将鼻子分离出来。第一张照片中哈士奇的身体以及第三张照片中间的人群也有同样的情况；我们的框架倾向于合并更多连续的令牌。



图 7。定性比较使用 ViT-L_{384} 模型的 DSM 与 ToMe。合并后的令牌具有相同的边框和填充颜色。DSM 合并更多相邻的图像素块，这些图像块在语义上相似，从而产生更易解释且保留原始特征的结果。

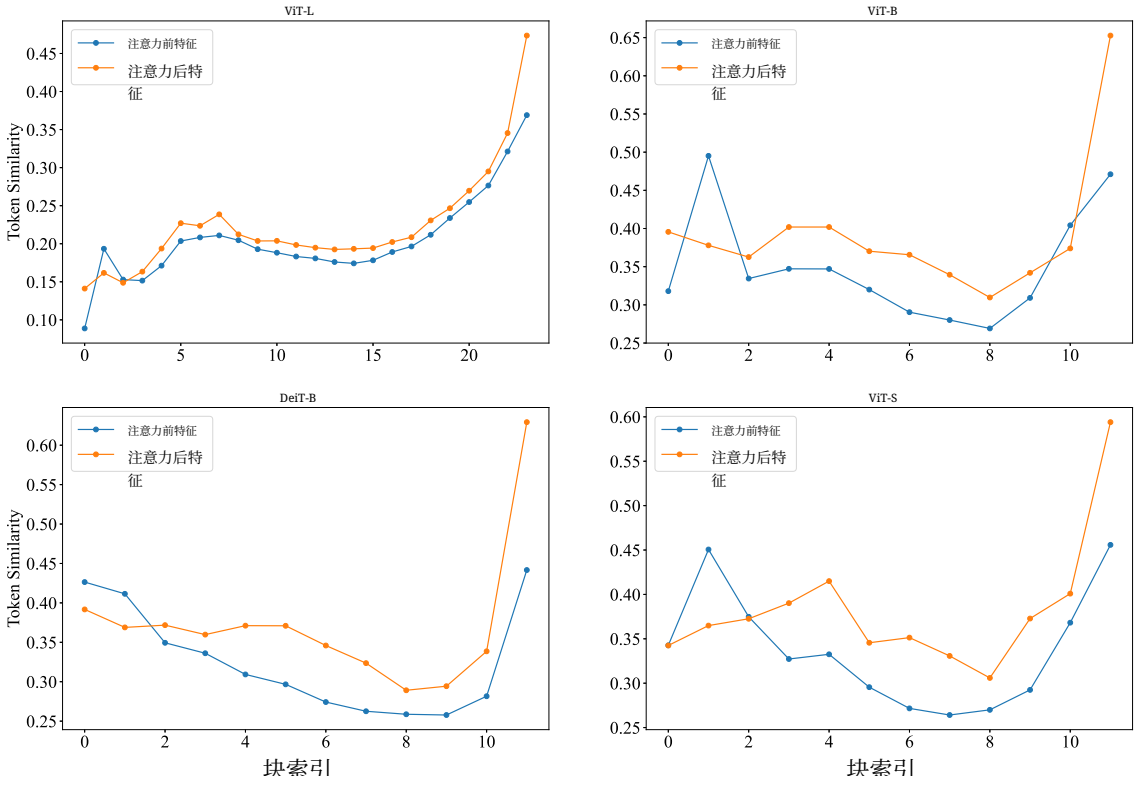


图 8. **延迟收敛注意力** 现象在各种预训练视觉变换器中均有观察到。注意力块会在某个阈值层之后持续使令牌变得更加相似，该阈值层约位于网络的六分之一处。

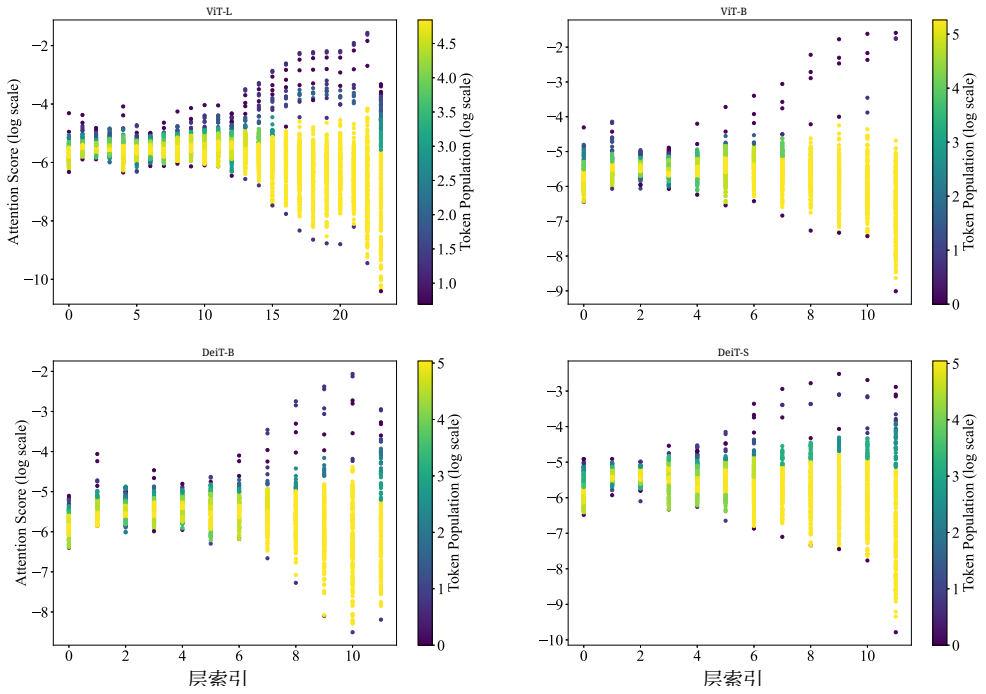


图 *9.* 不同视觉Transformer架构中观察到的离群标记。

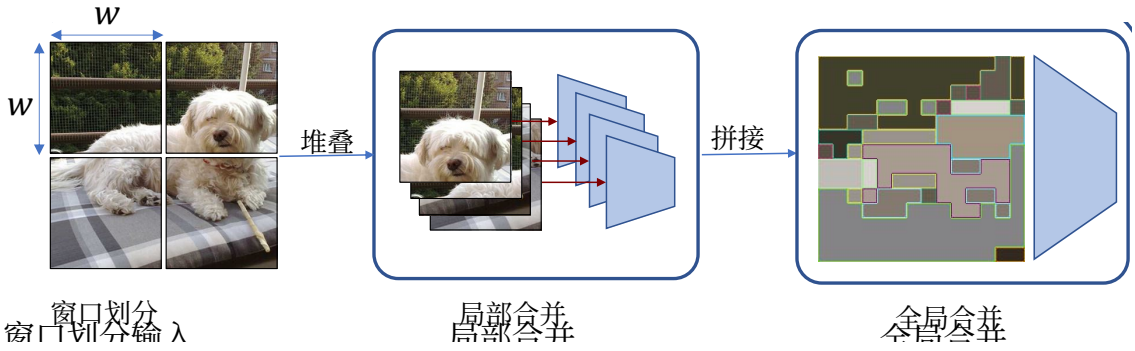


图 *10.* 通过视觉输入示意了结合窗口划分的局部合并。