Tranformer 在视觉领域内的发展综述

姓名：习钟　　学号：20212108032　　日期：2021 年 12 月 27 日

摘要

Transformer 是一种主要基于自注意力机制的深度神经网络，最早应用于自然语言处理领域。由于其强大的表示能力，研究人员正在寻找将 Transformer 应用于计算机视觉任务。在各种视觉基准测试中，基于 Transformer 的模型的性能与其他类型的模型相似或更好，比如在视觉领域已经取得成功的卷积神经网络和循环神经网络。鉴于其高性能和较少的视觉特征归纳偏置需求，Transformer 越来越受到计算机视觉社区的关注。在本文中，整理了有关于 Transformer 的相关视觉领域模型，并通过将模型分类到不同的任务并分析它们的优缺点。

关键字： Transformer，自监督，计算机视觉

Abstract

Transformer, first applied to the field of natural language processing, is a type of deep neural network mainly based on the self-attention mechanism. Thanks to its strong representation capabilities, researchers are looking at ways to apply transformer to computer vision tasks. In a variety of visual benchmarks, transformer-based models perform similar to or better than other types of networks such as convolutional and recurrent networks. Given its high performance and less need for vision-specific inductive bias, transformer is receiving more and more attention from the computer vision community. In this paper, we review these vision transformer models by categorizing them in different tasks and analyzing their advantages and disadvantages.

Keywords: Transformer, Self-attention, Computer Vision

深度神经网络 (DNN) 已成为当今人工智能 (AI) 系统中的基础设施。不同类型的任务通常涉及不同类型的网络，例如，多层感知（MLP）或全连接（FC）网络是神经网络的经典类型，它由堆叠在一起的多个线性层和非线性激活函数组成 [1]。卷积神经网络 (CNNs) 引入卷积层和池化层处理平移不变数据，例如图像 [2],[3]。循环神经网络（Recurrent Neural Network, RNN）是一类以序列（sequence）数据为输入，在序列的演进方向进行递归（recursion）且所有节点（循环单元）按链式连接的递归神经网络（recursive neural network）[4],[5]。Transformer 是一个新型神经网络。它主要利用 self-attention 机制 [6],[7] 提取内在特征 [8] 并在人工智能应用中展现出了能推广使用的巨大潜力。Transformer 在视觉领域里的发展随着 ViT 这篇论文的出现爆发出了巨大的潜力以及热度，ViT 完全证明了把 Transformer 用于图像领域是完全可行，此后 Transformer 在视觉领域的相关研究层出不穷，本文将以研究 Transformer 在视觉领域的应用为主旨，统计整理各方数据与论文来分析 Transformer 在视觉领域里的一些进步。

1　Transformer 介绍

Transformer 最早应用于自然语言处理 (NLP) 任务，并在自然语言处理上取得了巨大的进步 [8],[9],[10]。Vaswani 等人 [8] 第一个提出了基于注意力机制的 Transformer，用于机器翻译和词义解析任务。Devlin 等人提出了一种名为 BERT（简称 BERT）的新型语言表征模型（改进为双向编码器表示的 Transformer 的），利用了未标记文本对 Transformer 进行预训练每个单词的上下文（它是双向的）。BERT 在 11 NLP 任务上获得了最好成绩。受到 Transformer 在 NLP 领域的重大成功的启发，研究人员最近应用了 Transformer 到计算机视觉 (CV) 任务。在视觉应用中，CNNs 在视觉领域里独占鳌头 [11],[12]，但现在 Transformer 有取代 CNN 的潜力。陈等人 [13] 训练了一系列 Transformer 来自动回归预测像素，在图像分

类任务上取得与 CNN 差不多的成绩。另一种是最近由 Dosovitskiy 等人提出的 ViT（Visual Transformer），它将完整的 Transformer 直接应用于图像 patch 序列 [14]。并在多个图像识别任务上取得了最好的成绩。除了基本的图像分类，transformer 已经用于解决各种其他计算机视觉问题，包括对象检测 [15],[16]、语义分割、图像处理和视频理解。由于其卓越的性能，越来越多的研究人员提出基于 Transformer 的视觉处理模型，用于改进现有的视觉处理任务。

Transformer 最早应用于自然语言领域（NLP）中处理机器翻译任务。如图 1 所示，它由编码器模块和解码器模块组成具有多个编码器/解码器，每个编码器和解码器由一个自注意力层和一个前馈神经网络组成，而每个解码器还包含一个编码器-解码器注意力层。在 transformer 可以用来翻译句子之前，句子中的每个单词都需要嵌入到具有 D 维度的向量中。

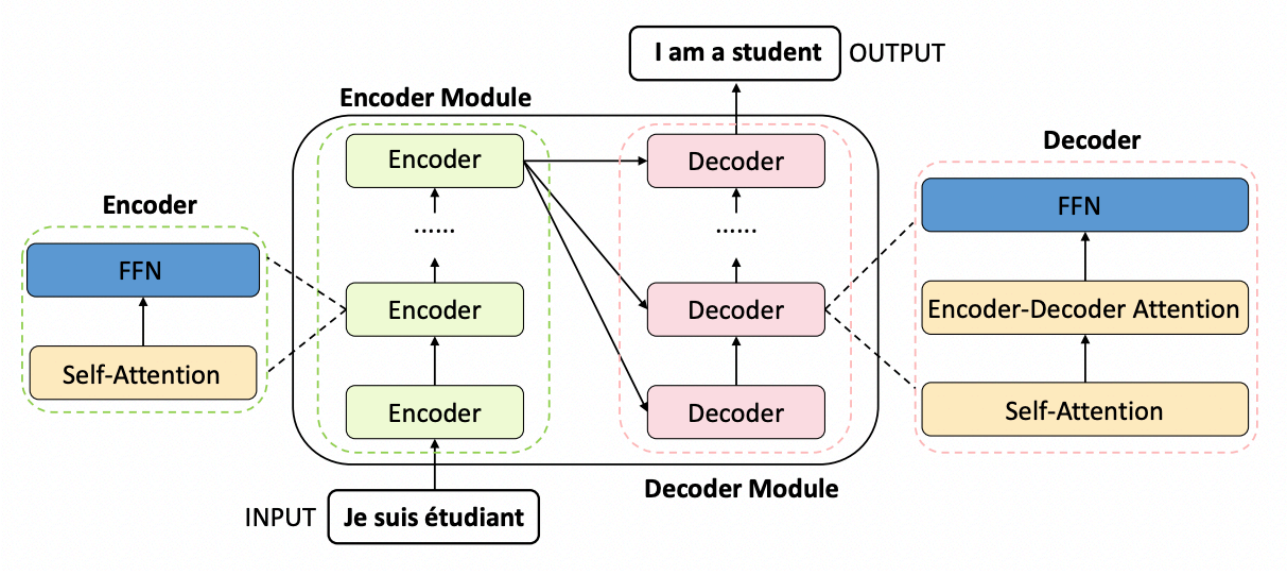

图 1: Transformer 翻译句子流程图

在 Transformer 出现之前，RNNs（例如 LSTM [17]）作为广泛使用的模型之一在处理自然语言时增加了注意力机制。然而 RNN 需要从前一个有序处理的信息流隐藏状态到下一个，这排除了在训练期间使用加速和并行化的可能性，因此阻碍了 RNN 处理更长序列或构建潜力更大的模型的可能性。2017 年，Vaswani 等人提出的 Transformer 作为一种完全基于多节点构建的新型编码器-解码器架构自注意力机制和前馈神经网络。它的目的是解决端到端的（seq-to-seq ）自然语言处理任务（例如机器翻译），很轻松地通过获取全局依赖来实现。Transformer 迎来了巨大的成功，同时表明只利用注意力机制也可以达到非常好的效果，丝毫不输于 RNN。同时 Transformer 的网络结构更适合大规模并行计算，从而可以在更大的数据集上进行训练，例如用于大规模自然语言处理的预训练模型 (PTM)。

当然除了用于大型语料库上训练 PTM 任务，基于 transformer 的模型也被应用于 NLP 的许多其他相关领域以及多模态任务。

2 Transformer 在图像领域里的进步

受到 Transformer 在 NLP 领域成功经验的启发，一些研究人员探索了相似的模型是否可以有效地学习处理图像。相比于自然文本，图像涉及更多的维度、噪声和冗余状态，因此被认为更难利用 Transformer 的自注意力机制。不同于 CNNs，Transformer 可以用 backbone 网络作为图像分类的网络。吴等人 [18] 采用了 ResNet 作为常用的基准并使用视觉 Transformer 来代替最后的卷积。实际上，他们应用卷积层来提取低级特征，然后将这些特征输入视觉 Transformer 进行训练。对于视觉 Transformer，他们使用标记器将像素分组为少量的图像标记，每个代表图像中的语义概念。这些图像标记直接用于图像分类，Transformer 对标记之间的关系进行建模。如下图 2 所示, 模型可以分为完完全全使用 Transformer 来处理，结合 CNN 的 Transformer 来处理，同时监督学习，自监督学习也在考虑结合 Transformer 来处理。

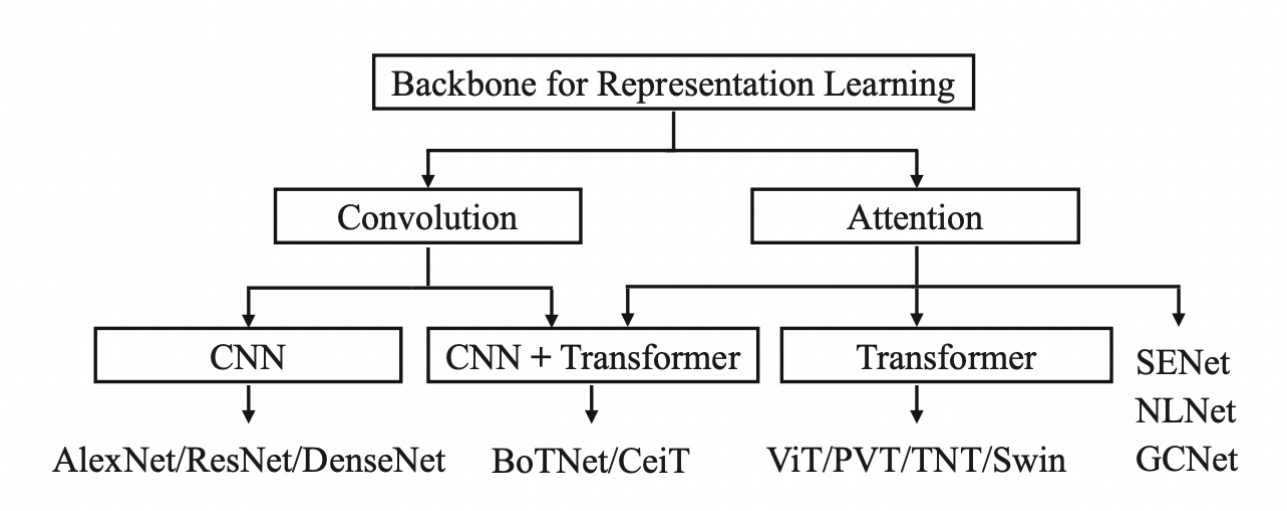

图 2: Transformer 的使用方向分类

2.1 完完全全的 Transformer：ViT

Dosovitskiy 等人 [14] 最近提出的 Vision Transformer (ViT)，这是一个在图像上表现良好的纯 Transformer，可以直接应用于图像序列的分类任务。如图 3 所示，ViT 尽可能地遵循 Transformer 的原始设计。

在 NLP 中，自注意力是两两相互的，但是将图片切割了之后，很显然位置信息就变了，所以图像的位置信息我们通过 Linear Projection of Flattened Patches 这个来进行全连接生成 token，具体操作如下: 我们给 Patch 加上 Position Embedding，即加上一个位置编码信息，一旦加上之后，这个 token 就包含了这个图片原本有的图像信息又包含了这个图像块现在有的位置信息。只要得到每一个 patch 的 token，那么就和 NLP 中的处理方式一样了，只需要将所有的 token 输入 Transformer Encoder，就会得到相应的很多输出。当然有这么多的输出 token，用哪个输出 token 去做分类呢? 可以通过加入一个特殊字符的方式，即通过加上 Extra learnable embedding 字符 (class embedding)，它也是一个 position embedding，其位置信息永远是 0。这样的话，所有的 token 都在相互交互计算，而且这个该 class embedding 能够做到从别的 embedding 中去学到有用的信息，所以只需要根据这一个 class embedding 去做最后的判断既

可以获得结果。图中的 MLP Head 是一个通用的分类头，最后用交叉熵函数去做模型的训练。在 ViT 中用的 Transformer Encoder 流程和 Transformer 中的也是一样的，即 Embedding Patch 进入 Layer Norm 以后，进行一系列的 Mutil-Head Attention 之后，再进入 Layer Norm，最后通过 MLP，就生成了一个 Transformer Block，叠加 L 次之后就形成了一个 Transformer Encoder。

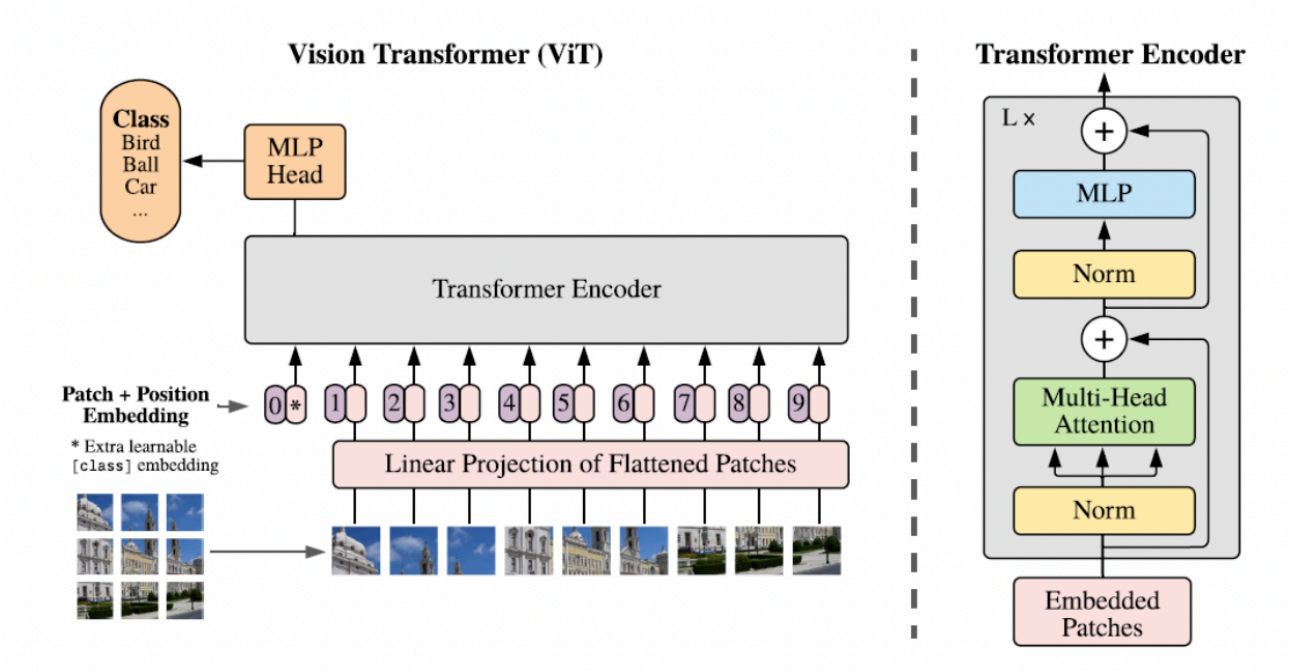

图 3: ViT 的处理流程

ViT 在中等规模的数据集上训练时产生的结果例如 ImageNet，达到的准确率低于同等大小的 ResNet，这是因为 Transformer 缺乏一些 CNN 固有的归纳偏差——例如特定方差和局部性——它们在在数据量不足的情况下训练时不能很好地概括整个图像。但是在大型数据集上（1400 万到 300 百万图像）训练模型超过了归纳偏差的模型。当预训练在足够大的规模时，Transformer 在数据点较少的任务上取得了优异的成绩。例如，在 JFT-300M 数据集上，ViT 接近甚至超过目前其他相关网络在多个图像识别基准上的最好表现，在 ImageNet 上达到了 88.36% 的准确率，CIFAR-10 上的 99.50% 的准确率、CIFAR-100 上的 94.55% 的准确率。

2.2 带卷积的 Transformer

Transformer 已能成功应用于各种视觉任务，但即便它们能够捕捉远距离 patch 输入中的依赖关系，Transformer 和现有的 CNN 之间性能仍然存在差距。一个主要原因可能是缺乏提取局部信息的能力。尽管现在已经有增强局部性的 ViT 变体模型，但是结合了带卷积神经网络的 Transformer 可以说是一种更直接明了的方式，直接将局部特征引入常规 Transformer 中。

CPVT [19] 提出了一种条件位置信息编码 (CPE) 方案，它以当地邻域为条件输入 token 并适应任意输入大小，利用卷积进行精细级别的特征编码。CvT [20]、CeiT [21]、LocalViT [22] 和 CMT [23] 分析了潜在的缺点，将卷积与 Transformer 组合在一起时，每个 Transformer 中的前馈网络（FFN）是促进了相

邻 token 之间卷积层的关联。

除此之外，一些研究人员已经证明基于 Transformer 的模型可能更难有拟合数据 [24]、[25] 的优势，换句话说，它们对优化器、超参数和训练调度策略这些因素更敏感。Visformer 用两种不同的训练设置解释了 Transformer 和 CNNs 之间的不同。结果显示相比于 Transformer，CNNs 的训练时间更短，其数据增强仅包含随机裁剪和水平翻转。

2.3 自监督学习

众所周知，深度学习训练需要大量的数据，然而在实际的下游任务训练中并没有很多的标注样本数据，通常只有几千甚至几百个数据，但是依然能训练出效果很好的模型，原因就是使用了“有监督预训练 + 下游任务微调”这样的范式。

以 ImageNet 为例，模型首先是在有 120 万标注数据的 ImageNet 分类数据集进行预训练，得到预训练模型，此后，下游任务则是基于预训练模型进行微调，通常的下游任务包括语义分割、目标检测、细粒度识别等等。相比不使用预训练模型，使用预训练模型的下游任务在模型性能上有很大的提升。

何等人提出 MoCo 在 7 个下游任务中，利用自监督预训练首次超越了有监督预训练的效果。这很可能意味着人工智能自监督或无监督时代的到来，这不但可以利用几乎无限的训练数据而无需标注，更重要的是，从认知的角度看，“自监督预训练 + 下游任务微调”这样的训练范式也与人类的学习方式更加接近。

iGPT [26] 结合了 Transformer 来做图像的自监督学习，主要思路包括一个预训练阶段，然后是一个微调阶段。在预训练阶段，借鉴了自回归和 BERT 的目标策略思想，为了实现像素预测，一系列的 Transformer 结构取代了在 NLP 中使用的 token。预训练被认为是在组合使用时有利的初始化或正则化器。同时在微调阶段，添加了一个小分类头到模型，这有助于生成一个小分类目标并适应所有权重值。

3 实验结果及分析

正如图 4 所示，对比代表性的 CNN 网络和在图像上使用 Transformer 模型，表中实验数据是在 V100 GPU 以及 Pytorch 的环境基础上所得到的结果。Vit 可以在训练数据集规模足够大的情况下在分类方面实现好于 resnet 的效果，且使用更少的训练资源 (相对)，不过在数据集较小的情况下效果较差，这是因为 Transformer 相较 CNN 来说，无法学习到一些有用的归纳偏置 (inductive biase)，即一些先验知识，（1）locality，CNN 假设相邻的区域会有相邻的特征，桌子和椅子大概率会在一起，考得越近的东西相关性就会越强（2）translation equivariance，平移等变性，CNN 的卷积核像一个 template 模板，同样的物体无论移动到哪里，遇到了相同的卷积核，它的输出一致。但是在结合了 CNN 的 ViT 上，可以发现效果相对来说是最好的，这意味着在结合 Transformer 和传统图像处理的方式里，仍然有着很多可以借鉴思考的地方。Transformer 的通用性一定会在图像里面取得更多的实践结果。

4 结论及未来发展

Transformer 的所有结构，包括多头自注意力、多层感知器、快速连接、层归一化、位置编码和网络拓扑等放到了视觉领域仍然有着很大的作用。通过实验结果可以看到结合 CNN 和 Transformer 获得了在图像识别最好的表现，表明通过本地连接和全局连接它们可以相互补充，再对 backbone 网的进一步研究一定可以提出更多的想法和实践。比如相比于 ViT 中的切割图像方式，人类视觉系统以一种完全不同的

Model	Params (M)	FLOPs (B)	Throughput (image/s)	Top-1 (%)
CNN				
ResNet-50 [89], [260]	25.6	4.1	1226	79.1
ResNet-101 [89], [260]	44.7	7.9	753	79.9
ResNet-152 [89], [260]	60.2	11.5	526	80.8
EfficientNet-B0 [213]	5.3	0.39	2694	77.1
EfficientNet-B1 [213]	7.8	0.70	1662	79.1
EfficientNet-B2 [213]	9.2	1.0	1255	80.1
EfficientNet-B3 [213]	12	1.8	732	81.6
EfficientNet-B4 [213]	19	4.2	349	82.9
Pure Transformer				
DeiT-Ti [55], [219]	5	1.3	2536	72.2
DeiT-S [55], [219]	22	4.6	940	79.8
DeiT-B [55], [219]	86	17.6	292	81.8
T2T-ViT-14 [260]	21.5	5.2	764	81.5
T2T-ViT-19 [260]	39.2	8.9	464	81.9
T2T-ViT-24 [260]	64.1	14.1	312	82.3
PVT-Small [232]	24.5	3.8	820	79.8
PVT-Medium [232]	44.2	6.7	526	81.2
PVT-Large [232]	61.4	9.8	367	81.7
TNT-S [85]	23.8	5.2	428	81.5
TNT-B [85]	65.6	14.1	246	82.9
CPVT-S [44]	23	4.6	930	80.5
CPVT-S-GAP [44]	23	4.6	942	81.5
CPVT-B [44]	88	17.6	285	82.3
Swin-T [148]	29	4.5	755	81.3
Swin-S [148]	50	8.7	437	83.0
Swin-B [148]	88	15.4	278	83.3
CNN + Transformer				
Twins-SVT-S [43]	24	2.9	1059	81.7
Twins-SVT-B [43]	56	8.6	469	83.2
Twins-SVT-L [43]	99.2	15.1	288	83.7
Shuffle-T [105]	29	4.6	791	82.5
Shuffle-S [105]	50	8.9	450	83.5
Shuffle-B [105]	88	15.6	279	84.0

图 4: 相关实验及其结果统计

方式组织视觉信息，而不是一次不加区别地处理整个场景。它循序渐进地、选择性地将注意力集中在视觉空间的有趣部分，而忽略不感兴趣的部分。例如，Dosovitski 等人观察到相对位置编码与绝对位置编码相比没有带来任何增益，语言建模采用原始相对位置编码，输入数据为一维单词序列。但对于视觉任务，输入通常是 2D 图像或视频序列，其中像素具有高度空间结构。目前尚不清楚：从一维到二维的扩展是否适用于视觉模型；方向信息在视觉任务中是否重要？这些都会得到解决。

参考文献

[1] F. Rosenblatt. The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory, 1957.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NeurIPS, pages 1097–1105, 2012.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[7] A. Parikh, O. Tackstr ¨ om, D. Das, and J. Uszkoreit. A decomposable ¨ attention model for natural language inference. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2249–2255, 2016.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. NeurIPS, 30:5998–6008, 2017

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACLHLT (1), 2019.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[12] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards realtime object detection with region proposal networks. In NeurIPS, 2015.

[13] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In International Conference on Machine Learning, pages 1691–1703. PMLR, 2020.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.

[16] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. In ICLR , 2021.

[17] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[18] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, M. Tomizuka, K. Keutzer, and P. Vajda. Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677, 2020.

[19] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen. Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882, 2021.

[20] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. arXiv preprint arXiv:2103.15808, 2021.

[21] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu. Incorporating convolution designs into visual transformers. arXiv preprint arXiv:2103.11816, 2021.

[22] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool. Localvit: Bringing locality to vision transformers. arXiv preprint arXiv:2104.05707, 2021.

[23] J. Guo, K. Han, H. Wu, C. Xu, Y. Tang, C. Xu, and Y. Wang. Cmt: Convolutional neural networks meet vision transformers. arXiv preprint arXiv:2107.06263, 2021.

[24] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian. Visformer: The vision-friendly transformer. arXiv preprint arXiv:2104.12533, 2021.

[25] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollar, and R. Gir- ´ shick. Early convolutions help transformers see better. arXiv preprint arXiv:2106.14881, 2021.

[26] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In International Conference on Machine Learning, pages 1691–1703. PMLR, 2020.