层级化的矢量量化变分自编码器 (Hierarchical-VQVAE)

郑问迪*
2018012654
计94
zwd18@mails.tsinghua.edu.cn

王哲凡
2019011200
计93
wzf19@mails.tsinghua.edu.cn

周Yan
2019011301
计93
zhouyan19@mails.tsinghua.edu.cn

1 简介

自从ResNet[1]问世以来，计算机视觉领域的顶尖模型对比中，一直充斥着ResNet的身影；那些可以不断刷新经典图像数据集各项指标的模型，也多以ResNet作为其参考的主干模型。但相比于继续执着深挖ResNet及其衍生模型的潜能，越来越多的计算机视觉领域研究者开始致力于寻找一个新的主干模型。

近年来，Transformer[2]在自然语言处理领域大放异彩，其结合了自注意力机制与残差性等于一体的网络结构，让深度学习相关研究人员在面对文本序列时，不再只有使用循环神经网络这一种选择。在Transformer 编码器基础上的预训练语言表征模型，BERT[3]，更是在自然语言处理领域掀起了大规模预训练语言模型的热潮。

出于Transformer和BERT在自然语言处理领域取得的成功，在过去两年间，许多研究者开始致力于将这些成功的实践引至计算机视觉领域。其中，较为成功的工作包括Vision Transformer[4]、Swin Transformer[5]和BEIT[6]等，这些模型、方法的出世也预示着Transformer在计算机视觉领域的前景。

而在视觉领域以Vision Transformer为主干模型的工作之中，以BEIT，Peco为代表的一些工作采取了首先将图片切分成序列的方式，并且采用的序列化方式又大多为DALL-E模型[7]中的离散变分自编码器（dVAE），或者又可称为矢量量化变分自编码器（VQ-VAE）[8]。矢量量化变分自编码器在编码的过程中，会首先将图像中的每一小块，对应到一个离散的编码表上，这也与自然语言处理中的序列化过程不谋而合。在这些以序列化为核心的计算机视觉领域工作中，对于图像的切分、序列化十分关键。除此之外，图像的序列化本身也可应用于图像压缩等场景，具备比较大的实用价值。

分词向量化作为深度学习训练最基本与最成功的思想之一，最初在自然语言处理领域有非常广泛的应用。包括BPE[9]，WordPiece[10]与ULM[11]等的分词方法已经是语言模型训练中基本达成共识的预处理方法。而与之对应地，在当前阶段，图像分词方法上的工作依然处于相对稀缺的状态，可探究的空间仍然较大。一个好的图像切分序列化表示方法不仅有利于图像任务下的下游训练任务，也将有利于文本表示和图像表示间的模态对齐，使得DALL-E，CogView[12]，Nuwa[13]等的图文多模态的单模型联合训练范式成为可能。

然而，在现实中各式各样的应用场景下，图像切分序列化的质量以及灵活程度均还有比较大的提高空间。在当前已有提出的模型下，往往对于一个全新的下游任务需求亦或是不同大小的图像切割方式，我们就只能重新训练一个全新的编码表来表示。

*组长

在我们的工作中，为了缓解这样的问题，我们提出了一个层级化的矢量量化变分自编码器的构建方案，其保留了朴素矢量量化变分自编码器的优势，并通过池化等方式，提供了将编码得到的长序列（高分辨率信息）转换为短序列（低分辨率信息）的可能。在联合训练的前提下，我们设计的各层自编码器成功取得了与单层单一训练时几乎持平的表现。除此之外，我们还成功实现了层级间的共享编码本（codebook）以降低联合训练的不必要开销，让不同层级的矢量量化变分自编码器可以在一个一致的编码体系下进行表示。

通过这种层级化联合训练的设计，BEIT 等视觉预训练框架可以有更灵活的序列长度（分辨率）选择。同时，得益于编码本的共享，模型也拥有了单次训练即可适配不同场景的可能性。同样，在尝试与其他类似模型的比较时，一个共享的编码本和同样的后续网络结构，也有助于比较公平性的保证。

综上所述，我们的工作相比于过去而言，有以下贡献:

- 满足了现有应用场景下，对于多级别、不同分辨率或不同序列长度的图像序列化需求。
- 满足了多级自编码器的编码本共享，使得不同模型在一些任务的具体比较上更加公平，对于相关预训练模型具体的复杂下游任务应用也更为简洁、方便。
- 在对于多级矢量量化变分自编码器的联合训练条件下，各级的图像重建效果，与朴素版本单独训练一级的效果基本持平。

2 相关工作

近年来，基于注意力机制的Transformer模型在自然语言处理领域取得巨大成功，使得其作为主干模型向计算机视觉领域发生迁移。其中，以BEIT[6]，Peco[14]为首的部分工作基于对图片进行分词的预处理进行下游任务训练，取得了较好的效果。这些工作展现出图像分词作为预处理工作的潜力与重要性。对于图文多模态领域，图像分词的重要性表现得更为突出，Dalle，CogView，Nuwa等图文视频多模态生成模型分别对文本、图像、视频等输入数据进行分词以达成模态对齐的效果，使得模态对齐的联合自监督训练成为可能。

作为图像分词任务的重要方法，矢量量化变分自编码器VQ-VAE[8]将隐变量上的图像表示通过矢量量化的方法映射于编码本的固定条目，从而达到分词映射的效果，是上述工作均采用的基本图像分词方法。本质上说，这也是一种对于临近隐变量向量进行聚类从而统一进行表示的做法。在这一角度上，iGPT[15]提供了一种更为细粒度的聚类做法，对像素级别的图像表示单元进行分词表示。对于图片基于离散编码的分词表示，能够进行更为便捷的自回归图像生成等后续任务的训练，也是VQ-VAE与iGPT上述两个工作所包含的内容。

VQ-VAE-2[16]在VQ-VAE的基础上更进一步，为了提高表示序列重建图片的质量，提出了层级化的矢量量化变分自编码器结构，使用多层级的序列化表示进行图像的唯一重建，期待不同层级的序列分词能够对图像不同层次的信息进行编码。

我们提出的层级化矢量量化变分自编码器模型的思路来源于VQ-VAE与VQ-VAE-2提出的层级化结构，但与过去的层级化结构有所不同：(1)) 我们的模型目标寻求图像多层级表示下多层级各自独立重建图片的能力。(2) 我们的模型在层级化的条件下成功实现了可共享的编码本机制，使得多层表示含义类似，更易进行多层级间的对齐。

3 方法

我们的工作大体可以分为矢量量化变分自编码器的主体部分和多层自编码器部分。

3.1 矢量量化变分自编码器（VQ-VAE）

VQ-VAE与VAE[17][18]有着显著的区别。首先，我们需要定义一个隐式向量空间$E \in \mathbb{R}^{N\times D} = \{e_i^T\}_{i=1}^N$ 作为编码表，其中N 表示隐向量的个数，D 则表示隐向量的维度，而$e_i \in \mathbb{R}^D$ 则为具体的单个编码向量。

对于输入x，首先我们将其作用一个参数为θ 的编码器结构，得到$z_\theta(x)$，我们希望在编码表中找到其对应的最接近的编码向量:

$$z_{\theta,E}(x) = q_E(z_\theta(x)) = \arg\min_{e\in E} \|e - z_\theta(x)\|_2 \tag{1}$$

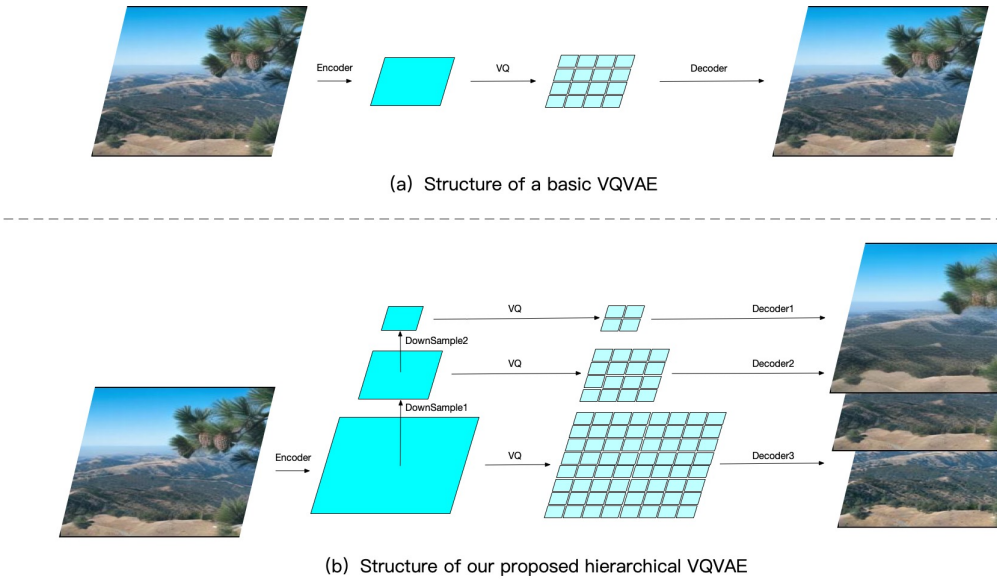

Figure 1: 单层矢量量化变分自编码器与层级化矢量量化变分自编码器结构对比

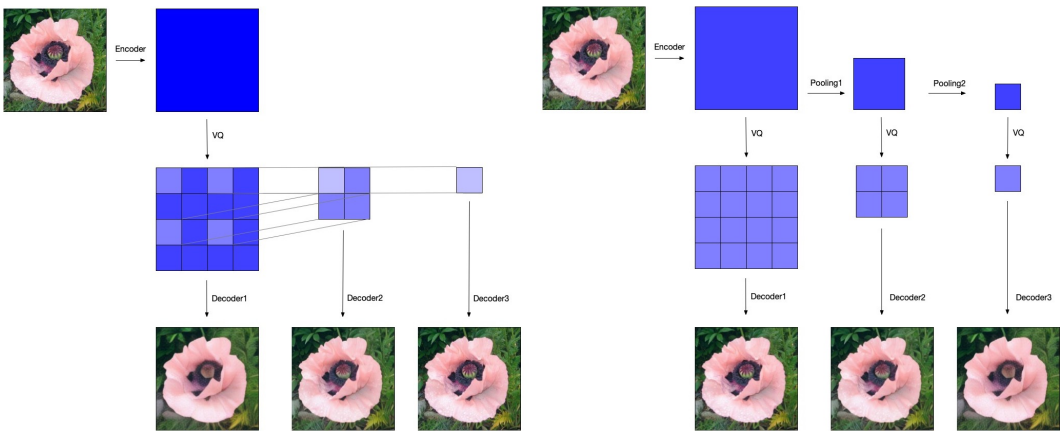

Figure 2: 两种改进策略(1)子序列选取(Subset Sampling); (2)残差池化(Pooling)

公式 1 中并没有从$z_{\theta,E}(x)$ 到θ 的梯度传播，但鉴于其形式与直连相似，我们可以通过梯度拷贝的方式，将$z_{\theta,E}(x)$ 的梯度$\nabla_{z_{\theta,E}}L$ 复制给$z_\theta(x)$，使$\nabla_{z_\theta}L = \nabla_{z_{\theta,E}}L$，进而传导给$\theta$，这可以通过以下方式实现:

$$\hat{z}_{\theta,E}(x) = \text{sg}\left[z_{\theta,E}(x) - z_\theta(x)\right] + z_\theta(x) \tag{2}$$

其中sg 算符表示stop gradient 即停止梯度传播，也即将其作用的对象视为不可更新的常数。

对于解码器，我们定义$p_\phi(x'|z)$ 为其输出的分布，则只须将其作用于离散化后的编码器输出即可得到$p_\phi(x'|\hat{z}_{\theta,E}(x))$。

最终的损失函数被设置为:

$$L = \log p_\phi(x|\hat{z}_{\theta,E}(x)) + \|\text{sg}[z_\theta(x)] - z_{\theta,E}(x)\|_2^2 + \beta\|z_\theta(x) - \text{sg}[z_{\theta,E}(x)]\|_2^2 \tag{3}$$

其中sg 算子与公式 2 所表达含义相同。

公式 3 的第一项表示自编码器的重建损失，这与朴素的自编码器基本一致，只是添加了从解码器输入到编码器输出的梯度传播通道，同时训练编码器参数θ 和解码器参数ϕ。第二项可称为编码本损失，其应用了矢量量化的思想，意图将编码本中的编码向量向其最近

的$z_\theta(x)$ 靠近，单独训练编码本参数E。而第三项则为了防止编码器输出在不同的编码向量之间跳跃而设计，其中β 为可调节的超参数，在实验中固定为$\beta = 1$，专用于训练编码器参数θ。

3.2 VQ-VAE 对图像输入的处理

假设图片的输入为$X \in [0,1]^{H\times W\times C}$，在作用一个一般化的编码器后得到$z_\theta(X) \in \mathbb{R}^{H'\times W'\times D} = [z_{ij}]_{H'\times W'}$，其中$z_{ij} \in \mathbb{R}^D$。

上文描述的对于输入的离散化，实际会被作用于每一个z_{ij} 上，即$z_{\theta,E}(X) = [q_E(z_{ij})]_{H'\times W'}$，再进一步得到拷贝梯度后的编码器输出$\hat{z}_{\theta,E}(X)$，并将其用于解码器的图像重建，具体过程可以参见Figure 1 中图(a)。

为了方便后续层次化的展开，此处假设H', W' 均为2 的幂次（实验中均设置为64）。

3.3 层次化的矢量量化变分自编码器（hierarchical VQ-VAE）

为了保持实验设定的简洁性与一致性，我们此处假设h 级序列长度分别为$H'\times W', H'/2\times W'/2, \cdots, H'/2^{h-1}\times W'/2^{h-1}$。

我们首先利用朴素的VQ-VAE 编码器作用输入图片上得到编码器输出$z_{\theta,1}(X) = z_\theta(X) \in \mathbb{R}^{H'\times W'\times D}$。为了满足VQ-VAE 同时对多种序列长度的支持，我们对编码器的结果进行降采样后再应用解码器:

$$\begin{aligned} z_{\theta,i+1}(x) &= \text{DownSample}_i(z_{\theta,i}(x)), i = 1,2,\cdots,h-1 \\ z_{\theta,E,i}(x) &= \arg\min_{e\in E} \|e - z_{\theta,i}(x)\|_2, i = 1,2,\cdots,h \\ \hat{z}_{\theta,E,i}(x) &= \text{sg}\left[z_{\theta,E,i}(x) - z_{\theta,i}(x)\right] + z_{\theta,i}(x), i = 1,2,\cdots,h \end{aligned} \tag{4}$$

其中$z_{\theta,i}(x), z_{\theta,E,i}(x), \hat{z}_{\theta,E,i}(x) \in \mathbb{R}^{H/2^{i-1}\times W'/2^{i-1}}$。

公式 4 中，对于子序列选取策略，每个DownSample将对输入直接分级采样，如图 2(1)所示。而对于池化策略来说，如 2(2)，每个DownSample 表示一个降采样块，其可以将输入$x' \in \mathbb{R}^{2h\times 2w\times c}$ 转换为输出$y \in \mathbb{R}^{h\times w\times c}$。具体而言，每个降采样块由一个基本的$3\times 3$ 作为卷积核、步长为2 的卷积神经网络，和一个卷积核大小为2×2、步长为2 的平均池化层组成:

$$y = \text{DownSample}(x') = \text{Conv}(x') + \text{AvgPool}(x') \tag{5}$$

通过公式 4、公式 5 可以看到，通过在矢量量化操作前，提前进行降采样得到不同规模的编码器输出，后续各层级VQ-VAE 得以使用共享的编码本E。

而后续各级重建时，我们再各自使用解码器$p_{\phi,i}(X'|Z)$ 解码，以计算最终的损失函数:

$$\begin{aligned} L = &\sum_{i=1}^{h} \log p_{\phi,i}(X|\hat{z}_{\theta,E,i}(X)) + \sum_{i=1}^{h} \|\text{sg}[z_{\theta,i}(X)] - z_{\theta,E,i}(X)\|_2^2 \\ &+ \beta \sum_{i=1}^{h} \|z_{\theta,i}(X) - \text{sg}[z_{\theta,E,i}(X)]\|_2^2 \end{aligned} \tag{6}$$

具体的模型过程以及和朴素VQ-VAE 的区别可以参见Figure 1 中图(b)。

4 实验

4.1 实验设置

本次实验使用ImageNet ILSVRC 2012[2]的训练集作为实验的训练数据集，在评估测试重建损失时，则使用由ImageNet测试集采样进行计算。

[2]https://image-net.org/challenges/LSVRC/2012/index.php

Table 1: 1000张测试集图片单层重建平均损失对比。CogView项为CogView中使用的VQ-VAE测试数据，Single为单层VQ-VAE模型，Sampling与Pooling分别表示子序列选取和残差池化两种策略下的层级化VQ-VAE。AvgPool及ConvPool表示残差池化层消融实验的对比模型。

	256to16		256to32		256to64	
模型	l1	perceptual	l1	perceptual	l1	perceptual
CogView	/	/	0.1020	0.2192	/	/
Single	**0.1528**	**0.2160**	**0.0954**	**0.0992**	**0.0512**	**0.0325**
Sampling	0.1687	0.2365	0.1110	0.1273	0.0767	0.0672
Pooling	**0.1539**	**0.2256**	**0.0989**	**0.1107**	**0.0543**	**0.0363**
AvgPool	0.1789	0.2723	0.1181	0.1478	0.0756	0.0610
ConvPool	0.1604	0.2322	0.1006	0.1148	0.0552	0.0389

Table 2: 不同模型编码10000张测试集图片对编码本使用率对比。模型各项表示同图 1。其中，CogView中训练使用的编码本大小为8192，其余模型使用的编码本大小为2×10^4。对于层级化的矢量量化变分自编码器，占用条数的统计为多层同时占用编码本条目的交集。

模型	CogView	Sampling	Pooling	AvgPool	ConvPool
占用条数	8188	20000	19905	3106	1

对于实验所训练的矢量量化变分自编码器，在尝试对比的结构设置之外，我们尽量维持其余结构的一致性。训练的编码本大小设置为2×10^4，条目的隐变量维数为256。对于层级化的变分自编码器，出于内存限制以及其一个编码器对应多个解码器的结构，设定编码器下采样层中残差块数为2，解码器上采样层则不设残差块，形成高密度编码结合轻量级解码的结构。对于单层的变分自编码器，编码器与解码器的单层均设2个残差块以维持编码与解码过程的平衡。

对于训练的重建评估指标，我们采用一维损失与预训练感知模型损失[19]混合的指标，训练时使用的二者权重均为1，编码本上的损失权重也设为1。

对于实验中每一种不同的模型结构设置，我们在8张A100 GPU上进行步数为1×10^5的训练，观察损失的变化曲线，同时取步数为1×10^5时刻的模型检查点作为测试评估版本。单步训练的批量大小为$5 \times 8 = 40$，初始学习率为1.0×10^{-4}，使用AnnealingLR技术进行学习率规划。

4.2 实验结果

单层重建对比 对于同时训练多层重建的矢量量化变分自编码器，与仅训练单层的矢量量化变分自编码器进行对比。比较中包含的层级化策略有子序列选取与残差池化，相关介绍可见方法部分所述。如图 3中所示，尽管三者训练都能得到模型收敛的结果，但子序列选取下的层级化模型，每一层的重建效果都略差于单层训练的情况。而采用残差池化的方法的层级化模型在每一层上的重建效果，从训练曲线上可见，能够与单层训练的效果相当。又见表 1，对比单层重建、子序列选取与残差池化的多层重建基于测试图片集的评估，依然发现残差池化的多层矢量量化变分自编码器能够取得与单层训练基本相当的效果，整体实验结论与损失曲线显示保持一致。

层级化策略对比 我们在多层重建损失合并的情况下对不同的层级化策略进行进一步比较，训练损失变化曲线见图 4(a)。可以看到，比起以池化或子选取的简单方法，以ResnetBlock结合卷积下采样层的复杂层级化结构反而不能得到更优的重建质量，而与上一部分的比较结论相一致的是基于残差池化的层级化矢量量化变分自编码器在所有层级化策略中能够得到最优的图像解码质量。参见表 1统计的层级化方法的总计测试损失，能够得到与比较训练损失曲线相一致的结论，说明实验结果没有由过拟合带来的比较偏差。

池化层消融实验 对于结合均值池化与卷积下采样的残差池化部分，我们进行消融实验来进一步验证这一设计的效果。我们分别使用均值池化层与卷积下采样层取代原有的残差池化层进行对比。如图 4(b)所示，在采取残差池化的层级化策略的情况下，能取得最佳的重建

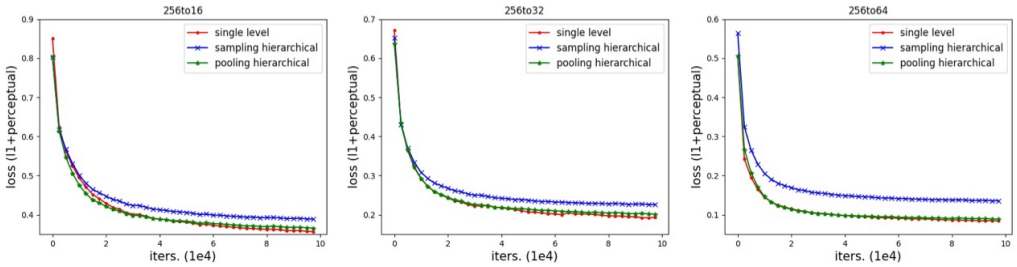

Figure 3: 256-16/256-32/256-64三级重建的训练损失曲线对比

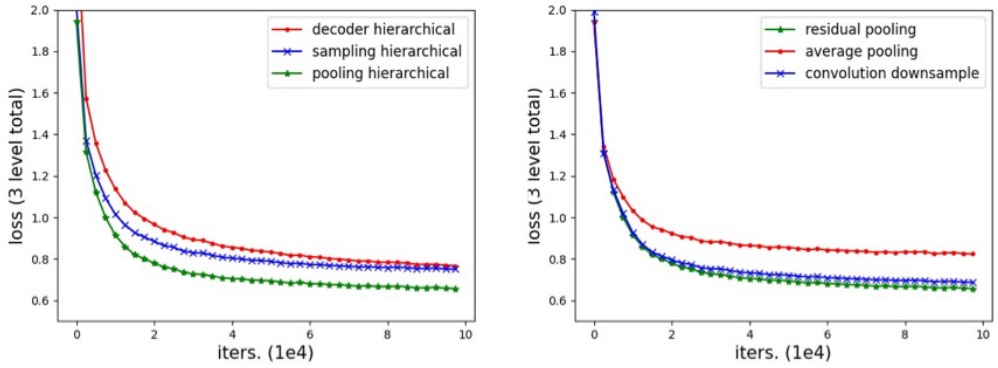

Figure 4: （a）不同层级化策略训练图像重建损失曲线对比（b）残差池化层消融实验的训练损失曲线对比

效果，而与之对比的均值池化层和卷积下采样的拆分的重建损失均高于二者结合的残差池化策略。又如表 **??**中所见，将均值池化与卷积下采样结合的残差池化策略使得图像编码对于编码本的使用与共享达到较高水平，而仅有卷积下采样层的模型尽管能有类似或略逊的重建效果，层级之间依然无法实现编码本的共享。这一实验结果印证了残差池化结构设计目标的成功——即在能够基本维持隐变量表示分布的前提条件下拟合重建效果更佳的编码器结构。

解码图像质量分析 为通过更为直观的方式对比上述矢量量化变分自编码器设计策略，我们查看不同模型编码真实图片并进行解码重建的结果。图片依然取自ImageNet ILSVRC的测试集。从展示样本对比中可见，层级化训练得到的各级图像重建效果能够与单层重建相当甚至更优。

5 总结

综上所述，我们提出了一个层级化的矢量量化变分自编码器(Hierarchical VQVAE)，并通过采样、池化等方式，提供了将编码得到的长序列转换为短序列的可能。我们还成功实现了层级间的共享编码本（codebook）以降低联合训练的不必要开销，让不同层级的矢量量化变分自编码器可以在一个一致的编码体系下进行表示。这种层级化的矢量量化变分自编码器的各层重建效果与单层同等大小几乎持平，故可以认为我们在没有降低编码器表示能力的前提下，降低了训练开销，并在一次训练中获得了不同尺度的序列。

我们的工作有助于在现实中各式各样的应用场景下，提升图像切分序列化的质量以及灵活程度，让不同模型在一些任务的具体比较上更加公平，从而为更广泛的下游任务应用提供可能。此外，共享编码本这一特点，在图片的清晰度提升、缩略图生成这样的任务上也有应用空间。

Figure 5: 解码图像样本展示，256 × 256大小的图片被编码至16 × 16、32×及64

参考文献

References

[1] He, K., X. Zhang, S. Ren, et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 2016.

[2] Vaswani, A., N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008. 2017.

[3] Devlin, J., M.-W. Chang, K. Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Dosovitskiy, A., L. Beyer, A. Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Liu, Z., Y. Lin, Y. Cao, et al. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[6] Bao, H., L. Dong, F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[7] Ramesh, A., M. Pavlov, G. Goh, et al. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[8] Oord, A. v. d., O. Vinyals, K. Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[9] Sennrich, R., B. Haddow, A. Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[10] Schuster, M., K. Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

[11] Kudo, T. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*, 2018.

[12] Ding, M., Z. Yang, W. Hong, et al. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021.

[13] Wu, C., J. Liang, L. Ji, et al. N\" uwa: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021.

[14] Dong, X., J. Bao, T. Zhang, et al. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

[15] Chen, M., A. Radford, R. Child, et al. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[16] Razavi, A., A. van den Oord, O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876. 2019.

[17] Kingma, D. P., M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Doersch, C. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[19] Zhang, R., P. Isola, A. A. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595. 2018.