

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2023)12-3855-15

论文引用格式: Xu S C and Zhu Z Q. 2023. Orthogonality-constrained multihead self-attention for scene text recognition. Journal of Image and Graphics, 28(12):3855-3869(徐仕成, 朱子奇. 2023. 正交约束多头自注意力的场景文本识别. 中国图象图形学报, 28(12):3855-3869)[DOI: 10.11834/jig.221049]

正交约束多头自注意力的场景文本识别

徐仕成, 朱子奇*

武汉科技大学计算机科学与技术学院, 武汉 430065

摘要: **目的** 场景文本识别(scene text recognition, STR)是计算机视觉中的一个热门研究领域。最近,基于多头自注意力机制的视觉Transformer(vision Transformer, ViT)模型被提出用于STR,以实现精度、速度和计算负载的平衡。然而,没有机制可以保证不同的自注意力头确实捕捉到多样性的特征,这将导致使用多头自注意力机制的ViT模型在多样性极强的场景文本识别任务中表现不佳。针对这个问题,提出了一种新颖的正交约束来显式增强多个自注意力头之间的多样性,提高多头自注意力对不同子空间信息的捕获能力,在保证速度和计算效率的同时进一步提高网络的精度。**方法** 首先提出了针对不同自注意力头上 Q (query)、 K (key)和 V (value)特征的正交约束,这可以使不同的自注意力头能够关注到不同的查询子空间、键子空间、值子空间的特征,关注不同子空间的特征可以显式地使不同的自注意力头捕捉到更具差异的特征。还提出了针对不同自注意力头上 Q 、 K 和 V 特征线性变换权重的正交约束,这将为 Q 、 K 和 V 特征的学习提供正交权重空间的解决方案,并在网络训练中带来隐式正则化的效果。**结果** 实验在7个数据集上与基准方法进行比较,在规则数据集Street View Text(SVT)上精度提高了0.5%;在不规则数据集CUTE80(CT)上精度提高了1.1%;在7个公共数据集上的整体精度提升了0.5%。**结论** 提出的即插即用的正交约束能够提高多头自注意力机制在STR任务中的特征捕获能力,使ViT模型在STR任务上的识别精度得到提高。本文代码已公开:<https://github.com/lexiaoyuan/XViTSTR>。

关键词: 场景文本识别(STR);视觉Transformer(ViT);多头自注意力;正交约束;计算机视觉

Orthogonality-constrained multihead self-attention for scene text recognition

Xu Shicheng, Zhu Ziqi*

School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China

Abstract: Objective Scene text recognition (STR) is a hot research field in computer vision that aims to recognize text information from natural scenes. STR is important in many tasks and applications, such as image search, robot navigation, license plate recognition, and automatic driving. Most of the early STR models usually comprise a rectification network and a recognition network, while recent STR models usually comprise a convolutional neural network (CNN)-based feature encoder and a Transformer-based decoder or a customized CNN module and Transformer encoder-decoder. These STR models usually have a complex model architecture, large computational load, and large memory consumption. A vision Transformer (ViT)-based STR model called ViTSTR maintains balance among accuracy, speed, and computational load. How-

收稿日期:2022-11-23;修回日期:2023-03-22;预印本日期:2023-03-29

* 通信作者:朱子奇 zhuzq@wust.edu.cn

基金项目:国家自然科学基金项目(61702382)

Supported by: National Natural Science Foundation of China (61702382)

ever, without data augmentation targeted for STR, ViTSTR requires improvement in its accuracy. One reason for its low accuracy is that the naive use of multihead self-attention in ViT does not guarantee that different attention heads capture distinct features, especially the diverse features in complex scene text images. To address this problem, this paper studies the application of orthogonality constraints in ViT based on ViTSTR and proposes novel orthogonality constraints for the multi-head self-attention mechanism in ViT, which explicitly encourages diversity among multiple self-attention heads, improves the ability of multihead self-attention to capture information in different subspaces, and further improves the accuracy of the network while ensuring speed and computational efficiency. **Method** The proposed orthogonality constraints comprise two parts, namely, the orthogonality constraints for the features of query (Q), key (K) and value (V) on different self-attention heads and the orthogonality constraints for the linear transformation weights of Q , K , and V on different self-attention heads. Q , K , and V play important roles in the self-attention mechanism as input features of the attention head. The orthogonality of features on different attention heads clearly encourages diversity among multiple attention head features. The orthogonality constraints for the Q , K , and V features allow different self-attention heads to focus on features in different query, key, and value subspaces, hence explicitly enabling different self-attention heads to capture distinct features and guide the ViT model in improving its performance in text recognition tasks for very diverse scenes. Specifically, after normalizing the Q , K , and V features of each head, the orthogonality of Q , K , and V features between different heads is calculated, and the corresponding orthogonality is added as the regularization term after the loss function. The orthogonality of the Q , K , and V features between different heads is penalized by back-propagation, which acts as a constraint on the orthogonality of the corresponding features. Adding orthogonality constraints to the linear transformation weights of the Q , K , and V features on different self-attention heads provides an orthogonal weight space in the learning process of these features, hence triggering implicit regularization in network training and fully utilizing the feature and weight spaces of multihead self-attention. A similar approach is used for the Q , K , and V weights to constrain the orthogonality of the Q , K , and V weight spaces by penalizing the orthogonality of the corresponding weights. The feature and weight orthogonality constraints can produce improvements when used individually or in combination. **Result** Experiment results show that compared to the benchmark method, the overall accuracy of the proposed method on the test dataset is improved by 0.447% when adding the feature orthogonality constraint. Meanwhile, when adding the weight orthogonality constraint, the overall accuracy of the proposed method is improved by 0.364%. When both feature and weight orthogonality constraints are added, the overall accuracy of the proposed method is improved by 0.513%. We then compare the orthogonality changes in different ablation experiments, including the orthogonality changes of the Q , K , and V features and weights among different self-attention heads. The proposed orthogonality constraint can lead to a significant improvement in the corresponding orthogonality. In addition, the feature orthogonality constraint favors the orthogonality of the weights, and the weight orthogonality constraint favors the orthogonality of the features, but these effects are small. We also produce attention maps of the model with added orthogonal constraints and the baseline model in the CUTE80(CT) dataset. These attention maps show that the model with added orthogonality constraints focuses on more information in the attention region compared to the baseline model, which is helpful for recognizing the correct results. We also compare the performance of our method and that of previous competitive methods on several popular benchmarks. Our proposed method shows improvements on both regular and irregular datasets. Compared with the baseline, the accuracy of the proposed method is improved by 0.5% on regular datasets IIIT5K-words (IIIT), street view text (SVT), and ICDAR2003 (IC03) (860), by 0.5% and 0.8% on the irregular datasets ICDAR2015 (IC15) (1811) and IC15 (2077), respectively, by 0.8% on SVT perspective (SVTP), and by 1.1% on CT. In sum, the proposed method shows an overall accuracy improvement of 0.5%. **Conclusion** This paper proposes a novel orthogonality constraint for the multihead self-attention mechanism that explicitly encourages this mechanism to capture diverse subspace information. The Q , K , and V feature orthogonality constraints are used to improve the ability of the multihead self-attention mechanism in capturing the feature space information of the input sequence, and the Q , K , and V weight orthogonality constraints are used to provide orthogonal weight space exploration for the learning of features. Experiment results validate the effectiveness of the proposed plug-and-play orthogonality constraints in STR tasks, especially in improving the accuracy of the ViT model in irregular text recognition tasks. The code is publicly available: <https://github.com/lexiaoyuan/XViTSTR>.

Key words: scene text recognition (STR); vision Transformer (ViT); multihead self-attention; orthogonality constrained; computer vision

0 引言

场景文本识别(scene text recognition, STR)是从自然场景图像中识别文本信息的广受欢迎的计算机视觉研究任务,在许多任务和应用中都具有重要的意义和价值,例如图像搜索(Tsai等,2011)、机器人导航(Schulz等,2015)、车牌识别(Wang等,2017)和自动驾驶(Bojarski等,2016)等。然而,自然场景中的图像充满了多样性(Zhang等,2019),如透视、弯曲和旋转等,尤其是任意形状或严重扭曲的文本图像在场景文本识别任务中仍然具有挑战性(Luo等,2019;Shi等,2019;Luo等,2020)。

大多数早期的STR模型都适合Baek等人(2019)提出的四阶段STR框架,通常包括4个阶段:图像预处理阶段、特征提取阶段、序列建模阶段和预测转录阶段(刘崇宇等,2021)。如ASTER(attentional scene text recognizer)(Shi等,2019)、SAFL(self-attention based scene text recognizer with focal loss)(Tran等,2020)和MORAN(multi-object rectified attention network)(Luo等,2019)等场景文本识别器,它们通常由一个矫正网络和一个文本识别网络组成。STR模型受到Transformer(Vaswani等,2017)架构的启发,如HGA-STR(holistic representation guided attention network for scene text recognition)(Yang等,2020)、Hamming OCR(optical character recognition)(Li等,2020)、NRTR(no-recurrence sequence-to-sequence text recognizer)(Sheng等,2019)和SATRN(self-attention text recognition network)(Lee等,2020)等场景文本识别网络。HGA-STR和Hamming OCR使用基于卷积神经网络(convolutional neural network, CNN)的编码器和Transformer的解码器共同构成场景文本识别网络,NRTR和SATRN使用一个定制的CNN模块来提取图像特征,用于Transformer编码器—解码器的文本识别。这些早期方法和近期基于Transformer架构的STR模型通常会带来复杂的模型结构以及较大的计算负载和内存消耗。

在STR的大量工作中,关注的重点是识别的准

确性,几乎没有强调速度和计算效率,而这些问题对于资源有限的移动机器来说同等重要,Atienza(2021)利用ViT(vision Transformer)(Dosovitskiy等,2021)的简单性和效率,提出了一个简单的单阶段的模型结构ViTSTR(ViT-based STR),能够在精度与速度、内存和计算负载的前沿,表明最佳的权衡。然而,如果不使用专门针对STR的数据增强,ViTSTR在准确性方面仍然有较大提升空间,导致准确性不足的一个原因就是,在ViT中简单地使用多头自注意力无法保证不同的注意力头能够捕捉到更具差异性的特征,特别是复杂场景文本图像中的多样性特征。

针对上述问题,本文基于ViTSTR研究了正交约束在ViT中的应用,针对ViT中的多头自注意力机制提出了一种新颖的正交约束,并在场景文本识别任务上证明了所提出的正交约束的有效性。与已有的多头注意力中的正交约束和正则化研究不同,本文所提出的正交约束强调“自”注意力。“自”注意力的“自”体现在输入序列的3种特征 Q (query)、 K (key)和 V (value)内进行注意力运算,即体现的是在“自”身的3种特征上进行注意力操作。

具体来说,本文所提出的正交约束由两部分组成,1)针对不同自注意力头中 Q 、 K 和 V 特征的正交约束;2)针对不同自注意力头中 Q 、 K 和 V 线性变换权重的正交约束。 Q 、 K 和 V 作为注意力头的输入特征,在自注意力机制中扮演重要角色。不同注意力头上的特征正交,可以显式增强多个注意力头特征之间的差异性。针对 Q 、 K 和 V 特征的正交约束可以使不同的自注意力头关注到不同的查询子空间、键子空间和值子空间的特征,这将显式地使不同的自注意力头捕捉到更具差异的特征,并引导ViT模型在极具多样性的场景文本识别任务中获得更高的性能。更进一步地对不同自注意力头上 Q 、 K 和 V 特征的线性变换权重添加正交约束,为 Q 、 K 和 V 特征的学习过程中权重空间的探索提供正交权重空间的解决方案,这将在网络训练中带来隐式正则化的效果,并更加充分地利用多头自注意力的特征空间和权重空间。在场景文本识别任务上的广泛实验表明,本文所提出的即插即用的特征正交约束和权重正交约束可以在单独使用时产生改进,也可以在组合使用

时产生更进一步的改进。

1 相关工作

为了平衡准确性、速度和计算效率,利用 ViT 的简单性和效率,一个单阶段且快速有效的场景文本识别模型 ViTSTR 被提出。ViTSTR 首先将输入的图片分割成图像块,然后将图像块的线性嵌入序列作为 Transformer 编码器的输入,一个可学习的图像块嵌入和每个嵌入的位置编码一起被添加。该网络被训练成端到端以预测一串字符。ViTSTR 的模型结构和 ViT 一样,只使用了 Transformer 的编码器,它们的唯一区别是预测头。本文工作是在 ViTSTR 的基础上进行的。

Li 等人(2018)引入了分歧正则化,以明确地增强多个注意力头之间的多样性,并在多个翻译任务(Bahdanau 等,2016;Szegedy 等,2015)上证明了所提方法的有效性。具体来说,包括 3 种类型的分歧正则化,分别促进与每个注意力头相关的子空间、所关注的位置和输出表示与其他头不同。本文工作同样是促进不同注意力头之间的多样性,不同的是本文针对多头自注意力模块的输入特征 Q 、 K 和 V 引入了正交约束,促进不同的自注意力头能够关注不同的查询子空间、键子空间和值子空间。

Lee 等人(2019)提出了一种正交约束的多头注意力机制,所提出的正则化技术能够提高关键词发现任务(Shan 等,2018;He 等,2017)中的检测性能。具体来说,所提出的正则化是由注意头之间的上下文约束和分数向量约束得到的,它们各自是正交的,惩罚了注意头之间的位置和上下文非正交性,促进从不同的子序列输出不同的表示,这与本文所做的工作是不同的。

Zhang 等人(2021)填补了 Transformer 中正交约束研究的空白,并在多个机器翻译任务(Bahdanau 等,2016;Szegedy 等,2015)和对话生成任务(Li 等,2017;Zhang 等,2018)上证明了在 Transformer 中使用正交性约束的有效性。具体来说,所提出的正交约束用于自我注意和位置前馈网络中的线性变换以及自我注意中的亲和矩阵。从线性变换的边界 Lipschitz 常数的角度出发,提出了自我注意和位置前馈网络中的线性变换的正交约束,以实现训练中的数值稳定性。根据亲和矩阵的有界扰动理论提出了自

我注意中的亲和矩阵的正交约束。本文所提出的正交约束中的权重正交约束和 Zhang 等人(2021)提出的线性变换的正交约束类似,不同的是本文只对自注意力的线性变换权重添加正交约束。

正交权重已在递归神经网络(recurrent neural network, RNN)(Dorobantu 等,2016;Mhammedi 等,2017;Vorontsov 等,2017;Arjovsky 等,2016)中得到广泛探索,以帮助避免梯度消失和爆炸问题(Dorobantu 等,2016)。在 CNN 中,正交权重也被认为可以稳定激活的逐层分布(Rodríguez 等,2017)并使优化更有效。Huang 等人(2018)证明了矩形正交矩阵可以稳定网络激活的分布并正则化前馈神经网络(feed-forward neural network, FNN)。Bansal 等人(2018)证明了在深度 CNN 中使用权重正交正则化会带来显著的精度提升。Xie 等人(2017)表明了正交性的约束下,学习权重的相关性被迫进入一致且相对较低的水平,并证明了正交约束的有效性。这些工作进一步引导本文探索权重的正交约束及其在基于 ViT 的场景文本识别模型中的应用。本文的主要工作集中在将正交约束用于基于 ViT 的场景文本识别模型中的多头自注意力机制。

2 本文方法

多头自注意力机制允许模型共同关注来自不同位置的不同表示子空间的信息。为了进一步保证多样性,本文提出了两种新颖的正交约束,来显式促进不同的头关注不同的子空间信息。具体来说,本文提出的正交约束包括两部分,1)针对不同自注意力头中 Q 、 K 和 V 特征的正交约束;2)针对不同自注意力头中 Q 、 K 和 V 线性变换权重的正交约束。特征正交约束和权重正交约束可以在单独使用时产生改进,也可以在组合使用时产生更进一步的改进。

2.1 多头自注意力

在 ViT 模型中,将输入序列投影为查询矩阵 Q 、键矩阵 K 和值矩阵 V ,它们是输入序列投影得到的不同的特征表示,自注意力机制则是将 Q 和一组 K - V 对映射到输出,然后计算 Q 与 K 之间的相似性作为注意力权重,并将注意力权重作用在 V 上,得到自注意力机制的输出。在 ViT 中通常使用多头自注意力作为一个单头自注意力的扩展。在多头自注意力中,在每个头上单独计算自注意力,然后将每个头上

的自注意力输出进行拼接,并将拼接结果做线性变换,得到与输入序列维度相同的输出作为多头自注意力的输出。这个过程可以表示如下:

对于一个输入序列 $\mathbf{z} \in \mathbf{R}^{N \times D}$, 一个标准的 QKV (query, key, value) 自注意力 (self-attention, SA) (Vaswani 等, 2017) 的计算方式可表示为

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z} \mathbf{U}_{\text{qkv}} \quad (1)$$

$$\mathbf{A} = f_{\text{softmax}} \left(\frac{\mathbf{q} \mathbf{k}^T}{\sqrt{d}} \right) \quad (2)$$

$$\text{SA}(\mathbf{z}) = \mathbf{A} \mathbf{v} \quad (3)$$

式中, $\mathbf{U}_{\text{qkv}} \in \mathbf{R}^{D \times 3D}$, $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbf{R}^{N \times D}$, $\mathbf{A} \in \mathbf{R}^{N \times N}$, $d = \frac{D}{h}$, h 是自注意力头的个数。

多头自注意力 (multihead self-attention, MSA) 是 SA 的扩展, 计算方式为

$$\text{MSA}(\mathbf{z}) = [\text{SA}_1(\mathbf{z}); \text{SA}_2(\mathbf{z}); \dots; \text{SA}_h(\mathbf{z})] \mathbf{U}_{\text{msa}} \quad (4)$$

式中, $\mathbf{U}_{\text{msa}} \in \mathbf{R}^{D \times D}$, 多头自注意力并行地计算每个头上的自注意力, 最后投影它们的联合输出。

2.2 基于 ViT 的 STR 模型

本文方法是对 ViTSTR 的改进, 在网络结构上与 ViTSTR 相同, 仅使用由 L 个编码器块堆叠的 Transformer 编码器。在编码器块内, 每个输入首先经过 LN (layer normalization) 层, 然后进入 MSA 层, 再经过 LN 层后进入 MLP (multi layer perceptron) 层。残差连接置于 LN 和 MSA/MLP 的输出之间。MLP 层由使用 GELU (Gaussian error linear unit) 激活函数的两个全连接层组成。

输入图像 $\mathbf{X} \in \mathbf{R}^{H \times W \times C}$ 调整为 n 个二维特征块 $\mathbf{X}_p^i \in \mathbf{R}^{P^2 \times C}$, $i = 1, 2, \dots, n$ 。图像为 $H \times W$, 特征块为 $P \times P$, C 是通道数, $n = HW/P^2$ 。

编码器的输入为

$$\mathbf{z}_0 = [\mathbf{X}_{\text{class}}; \mathbf{X}_p^1 \mathbf{E}; \mathbf{X}_p^2 \mathbf{E}; \dots; \mathbf{X}_p^n \mathbf{E}] + \mathbf{E}_{\text{pos}} \quad (5)$$

式中, $\mathbf{X}_{\text{class}} \in \mathbf{R}^{P^2 \times C}$ 是可学习的分类标记, $\mathbf{E} \in \mathbf{R}^{P^2 \times D}$, $\mathbf{E}_{\text{pos}} \in \mathbf{R}^{(n+1) \times D}$ 是可学习的位置编码。

MSA 层的输出为

$$\mathbf{z}'_l = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1} \quad (6)$$

式中, $l = 1, 2, \dots, L$, L 是编码器块的个数。

MLP 层的输出为

$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l \quad (7)$$

式中, $l = 1, 2, \dots, L$ 。

最后的预测结果输出是一个线性投影序列, 具

体为

$$\mathbf{y}_i = \text{Linear}(\mathbf{z}_L^i) \quad (8)$$

式中, $i = 1, 2, \dots, S$ 。 S 是识别文本的最大长度加上两个标记 [GO] 和 [s]。

2.3 Q、K 和 V 特征的正交约束

多头自注意力机制允许模型的不同表示子空间联合关注不同位置的信息 (Vaswani 等, 2017)。然而, 没有机制可以保证不同的注意力头确实捕捉到不同的特征。不同注意力头上的特征正交, 可以显式增强多个注意力头之间的差异性。 $\mathbf{Q}$ 、 $\mathbf{K}$ 和 $\mathbf{V}$ 作为注意力头的输入特征, 在自注意力机制中扮演重要角色。这些因素引导本文提出了针对 $\mathbf{Q}$ 、 $\mathbf{K}$ 和 $\mathbf{V}$ 特征的正交约束, 强制的正交约束可以使多头自注意力机制关注到不同的查询子空间、键子空间和值子空间, 关注到不同子空间信息可以显式地使不同的注意力头捕获到更具差异的特征。

对于一组输入序列有 $\mathbf{Z} \in \mathbf{R}^{B \times N \times D}$, B 代表 batch size 的值, 即一组输入的样本个数。有

$$[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = \mathbf{Z} \mathbf{U}_{\text{qkv}} \quad (9)$$

式中, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbf{R}^{B \times N \times D}$ 。在多头自注意力中, 将 $\mathbf{Q}$ 、 $\mathbf{K}$ 和 $\mathbf{V}$ 分别分散到不同的注意力头上, 则有 $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbf{R}^{B \times h \times N \times \frac{D}{h}}$ 。将每个头上的 $N \times \frac{D}{h}$ 维的特征展平为一个向量, 并归一化, 则有 $\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}} \in \mathbf{R}^{B \times h \times \frac{ND}{h}}$, 对特征 $\mathbf{Q}$ 的正交约束 FQ , 计算为

$$FQ = \sum_{l=1}^L \frac{1}{B} \sum_{i=1}^B \left\| \hat{\mathbf{Q}}_i \hat{\mathbf{Q}}_i^T - \mathbf{I} \right\|_F^2 \quad (10)$$

式中, L 是编码器块的个数, $\hat{\mathbf{Q}}_i \in \mathbf{R}^{h \times \frac{ND}{h}}$, $\mathbf{I} \in \mathbf{R}^{h \times h}$ 是一个单位矩阵, $\|\cdot\|_F$ 表示 Frobenius 范数。

ViT 中的 Transformer 编码器由 L 个编码器块组成, 本文对 L 个编码器块内的多头自注意力模块都计算特征正交约束, 并将计算结果进行累加, 作为最终的特征正交约束, FQ 作为正则项添加在损失函数后面。

遵循同样的方式, 对特征 $\mathbf{K}$ 和 $\mathbf{V}$ 的正交约束 FK 和 FV , 计算为

$$FK = \sum_{l=1}^L \frac{1}{B} \sum_{i=1}^B \left\| \hat{\mathbf{K}}_i \hat{\mathbf{K}}_i^T - \mathbf{I} \right\|_F^2 \quad (11)$$

$$FV = \sum_{l=1}^L \frac{1}{B} \sum_{i=1}^B \left\| \hat{\mathbf{V}}_i \hat{\mathbf{V}}_i^T - \mathbf{I} \right\|_F^2 \quad (12)$$

式中, $\hat{\mathbf{K}}_i, \hat{\mathbf{V}}_i \in \mathbf{R}^{h \times \frac{ND}{h}}$ 。

FQ 、 FK 和 FV 可以分别反映出不同自注意力头之间查询子空间、键子空间和值子空间特征的正交性,式(10)–(12)中减去单位矩阵 I 是为了消除同一个自注意力头和自己的特征是重叠的影响。

FK 和 FV 也同样作为正则项添加到损失函数后面, FQ 、 FK 和 FV 共同构成特征正交约束的正则项。

2.4 Q 、 K 和 V 权重的正交约束

将正交性引入权重空间同样具有重要意义,这不仅可以为 Q 、 K 和 V 特征的学习过程中权重空间的探索提供正交权重空间的解决方案,并在网络训练中带来隐式正则化的效果,也将更加充分地利用多头自注意力的特征空间和权重空间。因此,本文对 Q 、 K 和 V 特征的线性变换权重矩阵添加正交约束。以 Q 为例,使 Q 在不同自注意力头上的权重矩阵正交,这将为不同自注意力头上 Q 的权重学习探索正交投影空间。

用 U_q 、 U_k 和 U_v 分别表示 Q 、 K 和 V 特征的线性变换权重矩阵,即

$$[U_q, U_k, U_v] = U_{qkv} \quad (13)$$

式中, $U_q, U_k, U_v \in \mathbb{R}^{D \times D}$ 。在多头自注意力中,把权重矩阵分配到不同的自注意力头上,以 U_q 为例,可以表示为

$$[H_1^Q, H_2^Q, \dots, H_h^Q] = U_q \quad (14)$$

式中, $H_i^Q \in \mathbb{R}^{D \times \frac{D^2}{h}}$ 表示在第 i 个头上 Q 的权重矩阵, $i = 1, 2, \dots, h$ 。

将 H_i^Q 按列展平,得到 $\tilde{H}_i^Q \in \mathbb{R}^{1 \times \frac{D^2}{h}}$,将其按行拼接,得到

$$\tilde{U}_q = \begin{bmatrix} \tilde{H}_1^Q \\ \tilde{H}_2^Q \\ \vdots \\ \tilde{H}_h^Q \end{bmatrix} \quad (15)$$

式中, $\tilde{U}_q \in \mathbb{R}^{h \times \frac{D^2}{h}}$ 。对 $\tilde{U}_q$ 进行按行归一化,则有 $\hat{U}_q = \frac{\tilde{U}_q}{\|\tilde{U}_q\|_F}, \hat{U}_q \in \mathbb{R}^{h \times \frac{D^2}{h}}$ 。

Q 的权重矩阵的正交约束 LQ 的计算为

$$LQ = \sum_{l=1}^L \|\hat{U}_q \hat{U}_q^T - I\|_F^2 \quad (16)$$

式中, $I \in \mathbb{R}^{h \times h}$ 是一个单位矩阵。

与 FQ 一样,本文对 L 个编码器块内的多头自注意力模块都计算权重正交约束,并将计算结果进行

累加,作为最终的权重正交约束, LQ 作为正则项添加到损失函数后面。

遵循同样的方式,对 K 和 V 的权重矩阵的正交约束,计算为

$$LK = \sum_{l=1}^L \|\hat{U}_k \hat{U}_k^T - I\|_F^2 \quad (17)$$

$$LV = \sum_{l=1}^L \|\hat{U}_v \hat{U}_v^T - I\|_F^2 \quad (18)$$

式中, $\hat{U}_k, \hat{U}_v \in \mathbb{R}^{h \times \frac{D^2}{h}}$,与 $\hat{U}_q$ 相同,表示归一化后的矩阵。

LQ 、 LK 和 LV 可以分别反映出不同自注意力头之间查询权重子空间、键权重子空间和值权重子空间的正交性,式(16)–(18)中减去单位矩阵 I 是为了消除同一个自注意力头和自己的权重空间是重叠的影响。

LK 和 LV 也同样作为正则项添加到损失函数后面, LQ 、 LK 和 LV 共同构成权重正交约束的正则项。

最终的损失项可以表示为

$$J(\theta) = \arg \min_{\theta} \{L_{CE} + \alpha Foc + \beta Loc\} \quad (19)$$

$$Foc = \lambda_1 FQ + \lambda_2 FK + \lambda_3 FV \quad (20)$$

$$Loc = \mu_1 LQ + \mu_2 LK + \mu_3 LV \quad (21)$$

式中, L_{CE} 表示交叉熵损失,是ViTSTR中使用的损失,本文提出的正交约束作为正则项加在交叉熵损失后面。 Foc 表示特征的正交约束, Loc 表示线性变换权重的正交约束。 $\alpha, \beta, \lambda_1, \lambda_2, \lambda_3, \mu_1, \mu_2, \mu_3$ 为惩罚因子,是控制相应正则项的超参数。

3 实验

3.1 数据集介绍

Baek等人(2019)为大多数现有的STR方法引入了一个统一的框架,以及一致的数据集以便进行公平比较,本文也使用相同的数据集来进行训练和评估。

3.1.1 训练集

为大量真实的场景文本图像添加标记以用于STR模型的训练有着非常高的成本,因此,大多数STR模型都使用合成数据集作为真实数据的替代来进行训练,下面首先介绍STR论文中两个流行的合成数据集。

MJ(MJSynth)(Jaderberg等,2014)是一个包含

8.9 M幅图像,用于场景文本识别的合成数据集。

ST(SynthText)(Gupta等,2016)是另一个流行的合成数据集,它包含5.5 M幅裁剪图像以用于场景文本识别。

按照Baek等人(2019)的建议,本文采用了与Atienza(2021)相同的做法,在综合训练集MJ+ST上训练,在训练时,每个batch的训练集将有50%的图像来自MJ,50%的图像来自ST。图1显示了训练集MJ和ST中的部分图像。



图1 合成数据集的部分图像
Fig. 1 Samples from datasets with synthetic images
(a) MJ; (b) ST)

3.1.2 测试集

7个主要的真实自然场景下的STR数据集已广泛用于评估STR模型,这些数据集可以分为两类:规则数据集和不规则数据集。规则数据集包含水平排列的字符的文本图像,通常有着较小的弯曲度和形变量;不规则数据集则包含大量弯曲、垂直、透视和模糊以及严重变形的文本。

1) 规则数据集。IIIT(IIIT5K-words)(Mishra等,2012)包含3 000幅用于测试的图像,这些图像是从谷歌图像搜索中抓取的数据集,大多来自街道场景,如招牌、广告牌、路标、海报等。

SVT(street view text)(Wang等,2011)包含647幅用于测试的图像,这些图像收集自户外街道图像,有些是模糊或低分辨率的。

IC03(ICDAR2003)(Lucas等,2005)包含1 110幅用于ICDAR2003鲁棒阅读竞赛的图像,这些图像是从自然场景中采集的。在去除长度小于3个字符的单词后,剩下867幅图像。该数据集还有一个包含860幅图像的版本。

IC13(ICDAR2013)(Karatzas等,2013)是IC03的扩展,继承了IC03的大部分图像,是为ICDAR2013鲁棒阅读竞赛而创建的,包含1 015幅图像。该数据集还包含另一个857幅图像的版本。

2) 不规则数据集。IC15(ICDAR2015)(Karatzas等,2015)是为ICDAR2015鲁棒阅读竞赛而创建的,包含2 077幅图像,这些图像是由谷歌眼镜(Google Glasses)在佩戴者的自然运动状态下拍摄的,包含很多模糊、旋转和低分辨率的图像。同样,该数据集也有另一个1 811幅图像的版本(Bai等,2018; Cheng等,2017)。

SVTP(SVT perspective)(Phan等,2013)包含了645幅从谷歌街景中收集的图像,大多是在非正面视角下拍摄的带透视投影的图像。

CT(CUTE80)(Risnumawan等,2014)包含288幅从自然场景中收集的文本图像,其中大多是弯曲的弧形文本图像。

图2显示了7个测试集中的部分图像。图2(a)第1—4行分别来自规则数据集IIIT、SVT、IC03和IC13;图2(b)第1—3行分别来自不规则数据集

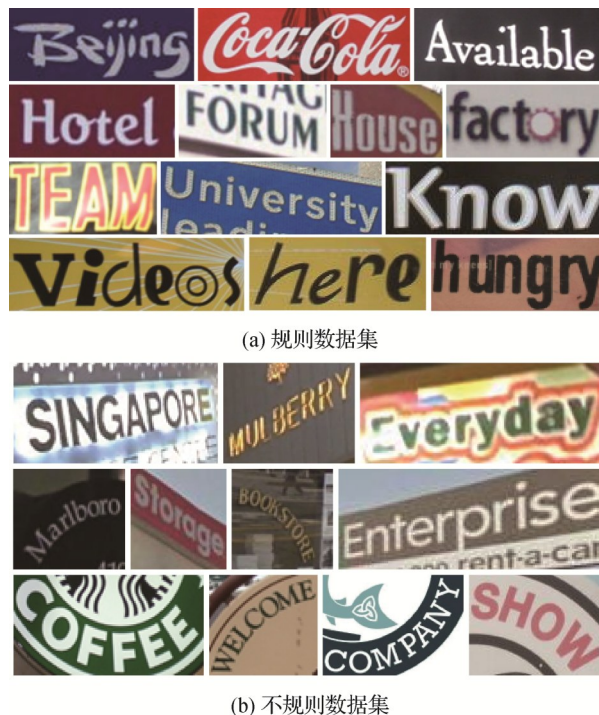


图2 真实场景数据集的部分规则和不规则图像
Fig. 2 Examples of regular and irregular real-world datasets
(a) regular datasets; (b) irregular datasets)

IC15、SVTP和CT。

3.2 实验设置

为了快速验证本文方法的有效性,首先在ViTSTR-Tiny (Atienza, 2021)的基础上进行实验,表1列出了本文推荐的训练配置。

与ViTSTR相同,使用DeiT (data-efficient image Transformers) (Touvron等, 2021)的预训练权重和He等人(2015)的方法进行初始化,每迭代2 000次后在Baek等人(2019)提供的验证集上进行验证,以选择在验证集上精度最高的模型用于在测试集上评估,

这一点与Atienza (2021)不同,本文和Baek等人(2019)保持一致。为了区别,在本文的推荐训练配置和评估方案选择上得到的模型命名为XViTSTR。

表2展示了XViTSTR不同版本的配置和精度对比。其中,精度是指对模型进行端到端训练后,选取在验证集上精度最高的模型,在测试集上测试得到的整体精度。与ViTSTR一样,本文识别的重点是96个拉丁字符,在评估前不旋转垂直文本,只对字母和数字进行评估,对于相同的英文字母,不区分大小写。

表1 训练配置
Table 1 Train conditions

配置项	值
训练集	50% MJ + 50% ST
迭代次数	30万次
batch size	192
优化器	Adam (Kingma 和 Ba, 2017) ($\beta_1 = 0.9, \beta_2 = 0.999$)
初始学习率	0.001
学习率计划	余弦退火学习率周期计划 (Loshchilov 和 Hutter, 2017)
梯度裁剪	5.0
Dropout (Hinton 等, 2012)	0.1
DropPath (Huang 等, 2016; Ghiasi 等, 2018)	0.1
输入图像尺寸/像素	224 × 224
图像通道数	1 (灰度图)
损失函数	cross-entropy (交叉熵)

表2 XViTSTR不同版本的配置和精度对比
Table 2 Comparison of the configuration and accuracy of different versions of XViTSTR

XViTSTR 版本	patch 尺寸	编码器块个数	embedding 尺寸	注意力头个数	序列长度	精度/%
	P	L	D	h	S	
Tiny	16	12	192	3	27	81.8
Small	16	12	384	6	27	83.6
Base	16	12	768	12	27	84.1

3.3 消融研究

表3展示了添加*Foc*和*Loc*在XViTSTR-Tiny模型上的精度对比结果。

为了验证特征正交约束和权重正交约束的有效性,首先对XViTSTR-Tiny版本按表1推荐的实验配置进行端到端训练,然后选取在验证集上精度最高

的模型在测试集上进行测试,得到在测试集上的整体精度为81.802%,并以此作为本文的基准。如表3中实验1所示,实验1仅使用交叉熵作为损失函数,其他实验除在损失函数中添加正交约束项外,都与实验1保持相同的实验条件。

为了验证特征正交约束*Foc*的有效性,在实验1

表 3 添加 *Foc* 和 *Loc* 在 XViTSTR-Tiny 模型上精度对比
Table 3 Add *Foc* and *Loc* for accuracy comparison on the XViTSTR-Tiny model

实验	<i>Foc</i>	<i>Loc</i>	精度/%
1	×	×	81.802
2	√	×	82.249
3	×	√	82.166
4	√	√	82.315

注:加粗字体表示各列最优结果,“√”表示添加,“×”表示不添加。

的基础上,仅在损失函数中添加 *Foc* 正交约束项,这时在实验 2 中的超参数 α 和 β 分别设置为 1 和 0, λ_1 , λ_2 , λ_3 分别设置为 1, 2, 1, μ_1 , μ_2 , μ_3 均设置为 0。

图 3 显示了实验 1 和实验 2 训练过程中不同自注意力头之间 *Q*、*K*、*V* 特征的正交性的变化对比图。实验中计算了实验 1—实验 4 中每次迭代时 *FQ*、*FK* 和 *FV* 的值,并取每 2 000 次迭代时的值。图 3 中实验 2 的蓝色实线和实验 4 的绿色实线在部分区域的

变化是重叠的。

图 3 中 *FQ*、*FK*、*FV* 的值越趋近于 0 就表示不同自注意力头之间对应特征空间的正交性越好。从图 3(a)可以看出,在不添加 *Foc* 时(实验 1),不同自注意力头之间 *Q* 特征的正交性在训练过程中整体处于 1.0 左右,这表明在不添加 *Foc* 时不同自注意力头之间的 *Q* 特征空间学习的特征存在一定冗余。在添加 *Foc* (实验 2)后,从图 3(a)中可以看出,强制的正交约束 *Foc* 可以使得不同自注意力头之间的 *Q* 特征空间的正交性迅速变好,并在整个训练过程中始终保持良好的正交性。从图 3(b)(c)中也可以得出相同的结论,这表明本文提出的正交约束 *Foc* 能够在训练中发挥作用,使不同的自注意力头能够关注到正交特征子空间,包括正交查询特征子空间、正交键特征子空间和正交值特征子空间。实验 2 所得到的模型在测试集上的表现也证明了本文所提出的特征正交约束 *Foc* 的有效性,如表 3 中的实验 2 所示,添加 *Foc* 的正交约束后精度相比实验 1 提高了 0.447%,达到 82.249%。

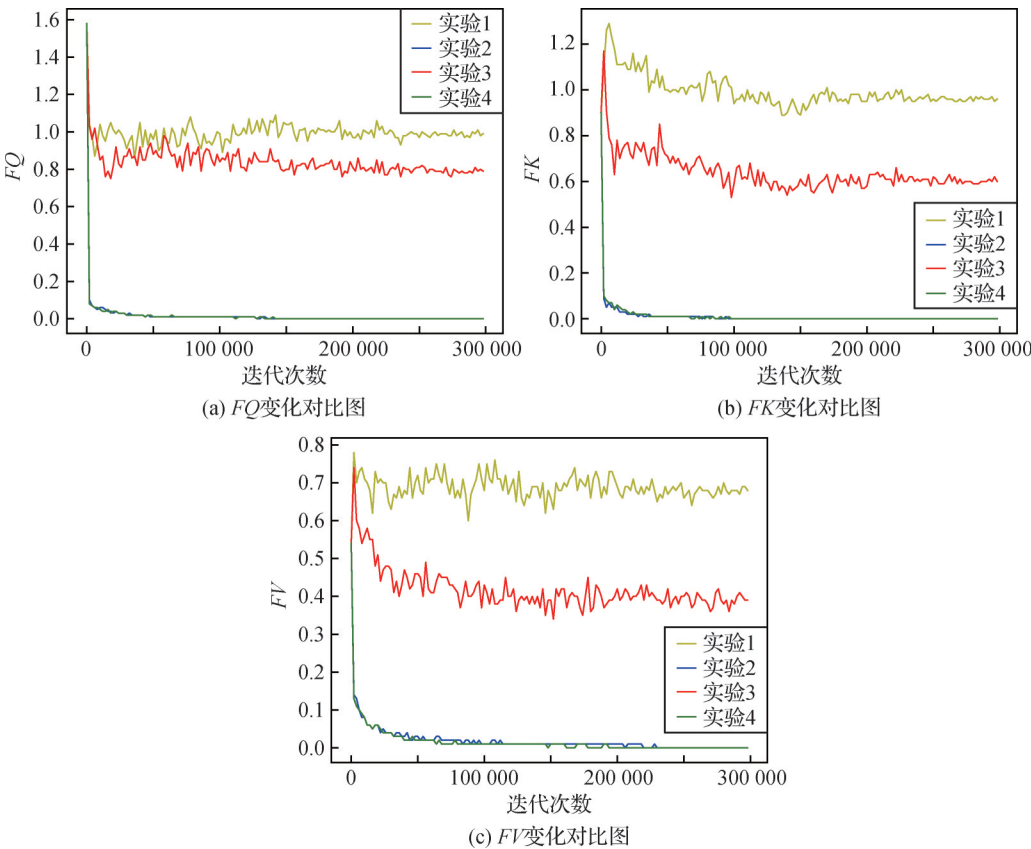


图 3 不同自注意力头之间 *Q*、*K*、*V* 特征正交性变化对比图

Fig. 3 Comparison of the orthogonality changes of *Q*, *K* and *V* features among different self-attention heads
((a) comparison of *FQ* changes; (b) comparison of *FK* changes; (c) comparison of *FV* changes)

为了验证权重正交约束 *Loc* 的有效性,在实验1的基础上,仅在损失函数中添加 *Loc* 正交约束项,这时在实验中的超参数 α 和 β 分别设置为 0 和 1, μ_1, μ_2, μ_3 分别设置为 10, 20, 10, $\lambda_1, \lambda_2, \lambda_3$ 均设置为 0。

图 4 显示了实验 1 和实验 3 训练过程中不同自注意力头之间 *Q*, *K*, *V* 权重的正交性的变化对比图。实验中计算了实验 1—实验 4 中每次迭代时 *LQ*、*LK* 和 *LV* 的值,并取每 2 000 次迭代时的值。图 4 中实验 3 的红色实线和实验 4 的绿色实线在部分区域的变化是重叠的。

图 4 中 *LQ*, *LK*, *LV* 的值越趋近于 0 就表示不同自注意力头之间对应权重空间的正交性越好。从图 4

可以看出,在不添加 *Loc* 时(实验 1),不同自注意力头之间 *Q*、*K*、*V* 权重的正交性在训练过程中整体趋于稳定。在添加 *Loc* 后(实验 3),相应的正交性迅速变好,并在之后的训练中能够保持正交性的稳定。这表明本文提出的正交约束 *Loc* 能够在训练中发挥作用,使不同的自注意力头能够关注到正交权重子空间,包括正交查询权重子空间、正交键权重子空间和正交值权重子空间。实验 3 所得到的模型在测试集上的表现也证明了本文所提出的权重正交约束 *Loc* 的有效性,如表 3 中的实验 3 所示,添加 *Loc* 的正交约束后精度相比实验 1 提高了 0.364%, 达到 82.166%。

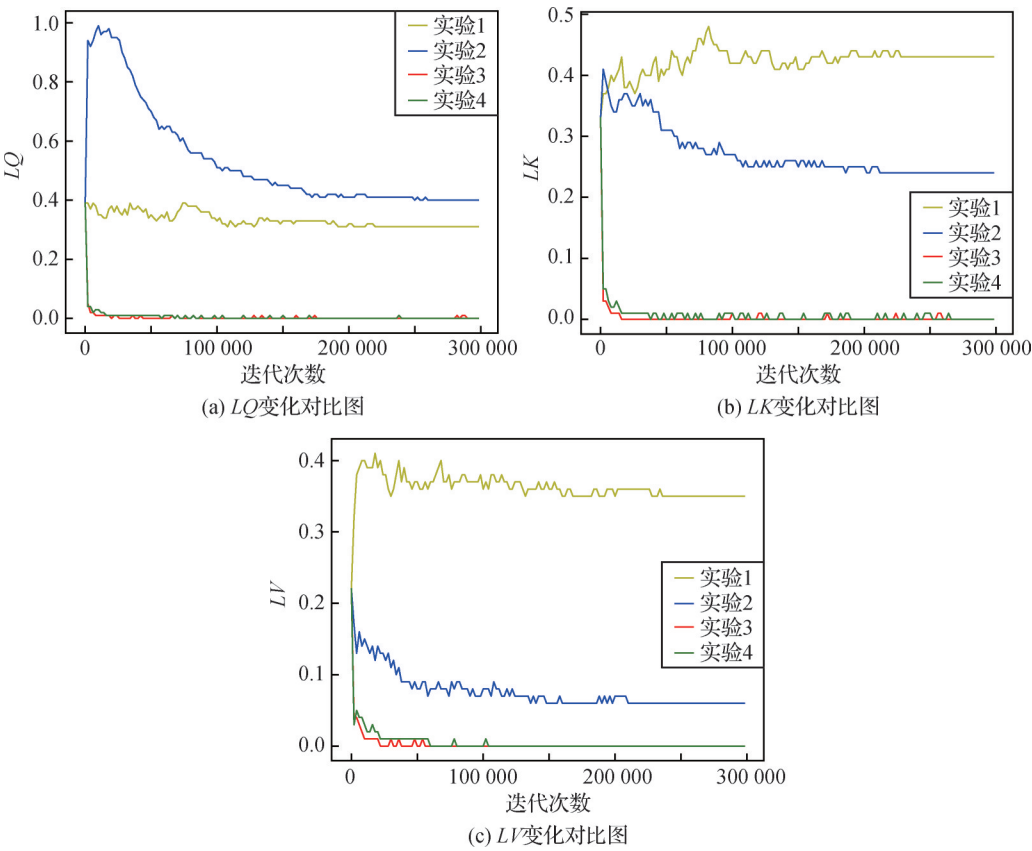


图 4 不同自注意力头之间 *Q*、*K*、*V* 权重正交性变化对比图

Fig. 4 Comparison of orthogonality changes of *Q*, *K* and *V* weights among different self-attention heads
((a) comparison of *LQ* changes; (b) comparison of *LK* changes; (c) comparison of *LV* changes)

事实上, *Foc* 和 *Loc* 不仅在单独使用时能够带来精度的提升,组合使用时也可以带来精度的进一步提升,如实验 4 所示,同时添加 *Foc* 和 *Loc* 时,超参数 α 和 β 都设置为 0.5, $\lambda_1, \lambda_2, \lambda_3$ 分别设置为 1, 2, 1, μ_1, μ_2, μ_3 分别设置为 10, 20, 10。从图 3 和图 4 中实验 4 的绿色实线变化都可以看出,这时, *FQ*, *FK*,

FV 和 *LQ*, *LK*, *LV* 的值都会迅速下降,并在训练过程中始终保持趋近于 0,说明这时 *Foc* 和 *Loc* 同时发挥了作用,使不同的自注意力头之间的 *Q*, *K*, *V* 特征空间和权重空间都保持正交。

值得关注的是,如图 4 所示,当只添加 *Foc* 时(实验 2),会导致 *LQ*, *LK*, *LV* 整体呈现下降趋势, *LQ* 先

增加到一个较大的值,然后稳步下降到接近不添加任何约束时(实验1) LQ 的值, LK 和 LV 的值整体比实验1要小且稳步下降。这表明,当添加特征的正交约束时,会对权重空间的正交性产生一定的影响,并使得相应的正交性整体上保持变好的趋势,特别是 LK 和 LV 的变化。

同样值得关注的是,如图3所示,当只添加 Loc (实验3)时,会导致 FQ , FK , FV 的值相比实验1在整体上呈现更明显的下降趋势。这表明,当添加权重的正交约束时,会使对应特征空间的正交性也变好,这也可以解释为什么添加权重的正交约束也可产生一定的改进,但产生的改进比直接添加特征的正交约束要小。

总的来说,单独添加 Foc 或 Loc 只会使特征空间或权重空间正交,而同时添加则会使特征空间和权重空间都能保持正交。如表3的实验4所示,同时添加 Foc 和 Loc 时的精度更高,相比实验1提高了0.513%,达到82.315%。

图5显示了由实验1和实验4得到的模型对CT数据集中部分图像的注意力可视化图。其中,第1列和第2列是由XViTSTR-Tiny模型得到的注意力图和识别结果,第3列和第4列是由XViTSTR-Tiny+ $Foc+Loc$ 模型得到的注意力图和识别结果。下划线字符是识别错误的字符。从图5可以看出,添加正交约束的模型在注意力区域上相比基准模型关注到了更多的信息,这对识别出正确结果来说是有帮助的。

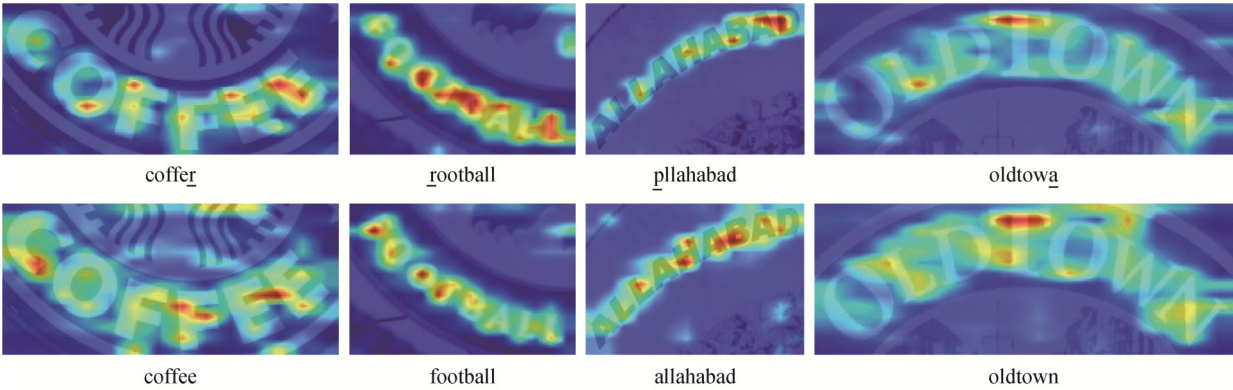


图5 不同模型对CT数据集中部分图像的注意力可视化

Fig. 5 Attention visualization of the images in the CT dataset part by different models

3.4 对比实验

本文还在几个流行的基准上与一些先进方法进行了比较,结果显示见表4。可以看到,本文方法相比基准,在规则和不规则数据集上均有提高,在规则

数据集IIIT、SVT、IC03(860)上都提升了0.5%,在不规则数据集IC15(1811)和IC15(2077)上分别提升了0.5%和0.6%,在SVTP上提升了0.8%,在CT上提升了1.1%,在所有测试集上的整体精度提升了0.5%。

表4 在公共测试集上不同模型的精度对比

Table 4 Accuracy comparison of different models on the public test dataset

模型											/%
	IIIT	SVT	IC03		IC13		IC15		SVTP	CT	
	3 000	647	860	867	857	1 015	1 811	2 077	645	288	平均精度
CRNN(Shi等,2017)	81.8	80.1	91.7	91.5	89.4	88.4	65.3	60.4	65.9	61.5	76.7
GCRNN(Wang和Hu,2017)	82.9	81.1	92.7	92.3	90.0	88.4	68.1	62.9	68.5	65.5	78.3
Rosetta(Borisjuk等,2018)	82.5	82.8	92.6	91.8	90.3	88.7	68.1	62.9	70.3	65.5	78.4
ViTSTR-Tiny(Atienza,2021)	83.7	83.2	92.8	92.5	90.8	89.3	72.0	66.4	74.5	65.0	80.3
XViTSTR-Tiny(基准)	84.8	85.3	93.8	93.4	90.9	90.3	73.9	68.0	77.7	72.9	81.8
XViTSTR-Tiny+Foc+Loc	85.3	85.8	94.3	93.7	91.4	90.7	74.4	68.6	78.5	74.0	82.3

注:加粗字体表示各列最优结果。

这说明本文方法能够提高多头自注意力机制对不同子空间信息的捕获能力,特别是在特征多样性极强的场景文本识别任务上。表5展示了流行方法和本文方法在精度、速度、参数量和计算负载上的对比。其中,参数量反映了内存的要求,每秒浮点运算次数

(floating point operations per second, FLOPS)估计了处理一幅图像所需的指令数量。可以看出,本文方法在速度和参数量接近或相同的情况下获得了有竞争力的表现。以上实验结果证明了本文方法的有效性。

表5 在2080Ti GPU上不同模型的精度、速度、参数量和计算负载(FLOPS)对比
Table 5 Comparison of accuracy, speed, number of parameters and computational load (FLOPS) of different models on 2080Ti GPU

模型	精度/%	速度/(ms/幅)	参数量/(1×10 ⁶)	FLOPS/(1×10 ⁹)
CRNN(Shi等,2017)	76.7	3.7	8.5	1.4
GCRNN(Wang和Hu,2017)	78.3	11.2	4.8	1.8
Rosetta(Borisyuk等,2018)	78.4	5.3	44.3	10.1
ViTSTR-Tiny(Atienza,2021)	80.3	9.3	5.4	1.3
XViTSTR-Tiny(基准)	81.8	14.1	5.4	2.1
XViTSTR-Tiny+ <i>Foc+Loc</i>	82.3	14.2	5.4	2.1

4 结 论

场景文本识别在许多任务和应用中都具有重要意义,使用多头自注意力机制的ViT模型用于场景文本识别可以实现速度和计算效率的平衡,然而面对多样性极强的场景文本识别任务,多头自注意力机制通常不能保证确实捕获到多样性的特征。本文提出了一种新颖的正交约束来显式增强多头自注意力机制捕获更具多样性的子空间信息。本文所提出的正交约束由两部分组成,即针对不同自注意力头之间 Q 、 K 、 V 特征的正交约束和针对不同自注意力头之间 Q 、 K 、 V 线性变换权重的正交约束。这将显式地使不同的自注意力头能够关注正交的查询子空间、键子空间和值子空间的特征,并为 Q 、 K 、 V 特征的学习提供正交权重空间的探索。实验证明本文所提出的即插即用的正交约束在STR任务上的有效性,特别是能够提高ViT模型在不规则文本识别任务上的精度。

本文工作不引入任何新的可学习参数,不影响模型的推理速度和计算效率,因为在推理测试时不需要计算正交约束,但会在训练过程中带来一定的计算消耗,在本文的实验中会带来7h左右的额外训练时间。另外,本文工作只探索了多头自注意力的输入特征 Q 、 K 、 V 的正交约束,对于多头自注意力中

的注意力特征和输出特征等其他特征的正交约束研究仍然是一个开放性的问题。未来的工作包括探索其他形式针对多头自注意力的正交约束,并将这一想法扩展到其他识别和分类任务中。

参考文献(References)

Arjovsky M, Shah A and Bengio Y. 2016. Unitary evolution recurrent neural networks [EB/OL]. [2022-09-22].
<https://arxiv.org/pdf/1511.06464.pdf>

Atienza R. 2021. Vision Transformer for fast and efficient scene text recognition//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 319-334 [DOI: 10.1007/978-3-030-86549-8_21]

Baek J, Kim G, Lee J, Park S, Han D, Yun S, Oh S J and Lee H. 2019. What is wrong with scene text recognition model comparisons? Dataset and model analysis//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 4714-4722 [DOI: 10.1109/ICCV.2019.00481]

Bahdanau D, Cho K and Bengio Y. 2016. Neural machine translation by jointly learning to align and translate [EB/OL]. [2022-08-29].
<https://arxiv.org/pdf/1409.0473.pdf>

Bai F, Cheng Z Z, Niu Y, Pu S L and Zhou S G. 2018. Edit probability for scene text recognition [EB/OL]. [2022-08-12].
<https://arxiv.org/pdf/1805.03384.pdf>

Bansal N, Chen X H and Wang Z Y. 2018. Can we gain more from orthogonality regularizations in training deep CNNs?//Proceedings

- of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 4266-4276
- Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel L D, Monfort M, Muller U, Zhang J K, Zhang X, Zhao J K and Zieba K. 2016. End to end learning for self-driving cars [EB/OL]. [2022-08-11]. <https://arxiv.org/pdf/1604.07316.pdf>
- Borisjuk F, Gordo A and Sivakumar V. 2018. Rosetta: large scale system for text detection and recognition in images//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London, UK: ACM: 71-79 [DOI: 10.1145/3219819.3219861]
- Cheng Z Z, Bai F, Xu Y L, Zheng G, Pu S L and Zhou S G. 2017. Focusing attention: towards accurate text recognition in natural images//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5086-5094 [DOI: 10.1109/ICCV.2017.543]
- Dorobantu V, Stromhaug P A and Renteria J. 2016. DizzyRNN: reparameterizing recurrent neural networks for norm-preserving back-propagation [EB/OL]. [2022-09-22]. <https://arxiv.org/pdf/1612.04035.pdf>
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houshy N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. [2022-04-07]. <https://arxiv.org/pdf/2010.11929.pdf>
- Ghiasi G, Lin T Y and Le Q V. 2018. DropBlock: a regularization method for convolutional networks [EB/OL]. [2022-08-12]. <https://arxiv.org/pdf/1810.12890.pdf>
- Gupta A, Vedaldi A and Zisserman A. 2016. Synthetic data for text localisation in natural images [EB/OL]. [2022-08-12]. <https://arxiv.org/pdf/1604.06646.pdf>
- He K M, Zhang X Y, Ren S Q and Sun J. 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1026-1034 [DOI: 10.1109/ICCV.2015.123]
- He Y Z, Prabhavalkar R, Rao K, Li W, Bakhtin A and McGraw I. 2017. Streaming small-footprint keyword spotting using sequence-to-sequence models [EB/OL]. [2022-08-29]. <https://arxiv.org/pdf/1710.09617.pdf>
- Hinton G E, Srivastava N, Krizhevsky A, Sutskever I and Salakhutdinov R R. 2012. Improving neural networks by preventing co-adaptation of feature detectors [EB/OL]. [2022-08-12]. <https://arxiv.org/pdf/1207.0580.pdf>
- Huang G, Sun Y, Liu Z, Sedra D and Weinberger K Q. 2016. Deep networks with stochastic depth [EB/OL]. [2022-08-12]. <https://arxiv.org/pdf/1603.09382.pdf>
- Huang L, Liu X L, Lang B, Yu A, Wang Y L and Li B. 2018. Orthogonal weight normalization: solution to optimization over multiple dependent stiefel manifolds in deep neural networks. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 3271-3278 [DOI: 10.1609/aaai.v32i1.11768]
- Jaderberg M, Simonyan K, Vedaldi A and Zisserman A. 2014. Synthetic data and artificial neural networks for natural scene text recognition [EB/OL]. [2022-08-12]. <https://arxiv.org/pdf/1406.2227.pdf>
- Karatzas D, Shafait F, Uchida S, Iwamura M, Bigorda L G i, Mestre S R, Mas J, Mota D F, Almazan J A and de las Heras L P. 2013. ICDAR 2013 robust reading competition//Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, USA: IEEE: 1484-1493 [DOI: 10.1109/ICDAR.2013.221]
- Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar V R, Lu S J, Shafait F, Uchida S and Valveny E. 2015. ICDAR 2015 competition on robust reading//Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR). Tunis, Tunisia: IEEE: 1156-1160 [DOI: 10.1109/ICDAR.2015.7333942]
- Kingma D P and Ba J L. 2017. Adam: a method for stochastic optimization [EB/OL]. [2022-08-12]. <https://arxiv.org/pdf/1412.6980.pdf>
- Lee J, Park S, Baek J, Oh S J, Kim S and Lee H. 2020. On recognizing texts of arbitrary shapes with 2D self-attention//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE: 2326-2335 [DOI: 10.1109/CVPRW50498.2020.00281]
- Lee M, Lee J, Jang H J, Kim B, Chang W and Hwang K. 2019. Orthogonality constrained multi-head attention for keyword spotting//Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Singapore, Singapore: IEEE: 86-92 [DOI: 10.1109/ASRU46091.2019.9003738]
- Li B C, Tang X, Qi X B, Chen Y H and Xiao R. 2020. Hamming OCR: a locality sensitive hashing neural network for scene text recognition [EB/OL]. [2022-09-26]. <https://arxiv.org/pdf/2009.10874.pdf>
- Li J, Tu Z P, Yang B S, Lyu M R and Zhang T. 2018. Multi-head attention with disagreement regularization//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL: 2897-2903 [DOI: 10.18653/v1/D18-1317]
- Li Y R, Su H, Shen X Y, Li W J, Cao Z Q and Niu S Z. 2017. DailyDialog: a manually labelled multi-turn dialogue dataset [EB/OL]. [2022-08-12]. <https://arxiv.org/pdf/1710.03957.pdf>
- Liu C Y, Chen X X, Luo C J, Jin L W, Xue Y and Liu Y L. 2021. Deep learning methods for scene text detection and recognition. Journal of Image and Graphics, 26(6): 1330-1367 (刘崇宇, 陈晓雪, 罗灿杰, 金连文, 薛洋, 刘禹良. 2021. 自然场景文本检测与识别的深度学习方法. 中国图象图形学报, 26(6): 1330-1367) [DOI: 10.11834/jig.210044]
- Loshchilov I and Hutter F. 2017. SGDR: stochastic gradient descent

- with warm restarts [EB/OL]. [2022-08-12].
<https://arxiv.org/pdf/1608.03983.pdf>
- Lucas S M, Panaretos A, Sosa L, Tang A, Wong S, Young R, Ashida K, Nagai H, Okamoto M, Yamamoto H, Miyao H, Zhu J M, Ou W W, Wolf C, Jolion J M, Todoran L, Worring M and Lin X F. 2005. ICDAR 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJ DAR)*, 7 (2): 105-122 [DOI: 10.1007/s10032-004-0134-3]
- Luo C J, Jin L W and Sun Z H. 2019. MORAN: a multi-object rectified attention network for scene text recognition [EB/OL]. [2022-08-11]. <https://arxiv.org/pdf/1901.03003.pdf>
- Luo C J, Zhu Y Z, Jin L W and Wang Y P. 2020. Learn to augment: joint data augmentation and network optimization for text recognition//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA: IEEE: 13743-13752 [DOI: 10.1109/CVPR42600.2020.01376]
- Mhammedi Z, Hellicar A, Rahman A and Bailey J. 2017. Efficient orthogonal parametrisation of recurrent neural networks using householder reflections [EB/OL]. [2022-09-22].
<https://arxiv.org/pdf/1612.00188.pdf>
- Mishra A, Alahari K and Jawahar C V. 2012. Scene text recognition using higher order language priors//*Proceedings of 2012 British Machine Vision Conference*. Surrey, UK: BMVA: 1-11 [DOI: 10.5244/C.26.127]
- Phan T Q, Shivakumara P, Tian S X and Tan C L. 2013. Recognizing text with perspective distortion in natural scenes//*Proceedings of 2013 IEEE International Conference on Computer Vision*. Sydney, Australia: IEEE: 569-576 [DOI: 10.1109/ICCV.2013.76]
- Risnumawan A, Shivakumara P, Chan C S and Tan C L. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41 (18): 8027-8048 [DOI: 10.1016/j.eswa.2014.07.008]
- Rodríguez P, González J, Cucurull G, Gonsalus J M and Roca X. 2017. Regularizing cnns with locally constrained decorrelations [EB/OL]. [2022-05-05]. <https://arxiv.org/pdf/1611.01967.pdf>
- Schulz R, Talbot B, Lam O, Dayoub F, Corke P, Upercroft B and Wyeth G. 2015. Robot navigation using human cues: a robot navigation system for symbolic goal-directed exploration//*Proceedings of 2015 IEEE International Conference on Robotics and Automation (ICRA)*. Seattle, USA: IEEE: 1100-1105 [DOI: 10.1109/ICRA.2015.7139313]
- Shan C H, Zhang J B, Wang Y J and Xie L. 2018. Attention-based end-to-end models for small-footprint keyword spotting [EB/OL]. [2022-08-29]. <https://arxiv.org/pdf/1803.10916.pdf>
- Sheng F F, Chen Z N and Xu B. 2019. NRTR: a no-recurrence sequence-to-sequence model for scene text recognition//*Proceedings of 2019 International Conference on Document Analysis and Recognition (ICDAR)*. Sydney, Australia: IEEE: 781-786 [DOI: 10.1109/ICDAR.2019.00130]
- Shi B G, Bai X and Yao C. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (11): 2298-2304 [DOI: 10.1109/TPAMI.2016.2646371]
- Shi B G, Yang M K, Wang X G, Lyu P, Yao C and Bai X. 2019. ASTER: an attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9): 2035-2048 [DOI: 10.1109/TPAMI.2018.2848939]
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2015. Rethinking the inception architecture for computer vision [EB/OL]. [2022-08-29]. <https://arxiv.org/pdf/1512.00567.pdf>
- Touvron H, Cord M, Douze M, Massa F, Sablayrolles A and Jégou H. 2021. Training data-efficient image Transformers and distillation through attention//*Proceedings of the 38th International Conference on Machine Learning*. PMLR: 10347-10357
- Tran B H, Le-Cong T, Nguyen H M, Anh Le D, Nguyen T H and Le Nguyen P. 2020. SAFL: a self-attention scene text recognizer with focal loss//*Proceedings of the 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Miami, USA: IEEE: 1440-1445 [DOI: 10.1109/ICMLA51294.2020.00223]
- Tsai S S, Chen H Z, Chen D, Schroth G, Grzeszczuk R and Girod B. 2011. Mobile visual search on printed documents using text and low bit-rate features//*Proceedings of the 18th IEEE International Conference on Image Processing*. Brussels, Belgium: IEEE: 2601-2604 [DOI: 10.1109/ICIP.2011.6116198]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Vorontsov E, Trabelsi C, Kadoury S and Pal C. 2017. On orthogonality and learning recurrent networks with long term dependencies [EB/OL]. <https://arxiv.org/pdf/1702.00071.pdf>
- Wang J F and Hu X L. 2017. Gated recurrent convolution neural network for OCR//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, USA: Curran Associates Inc.: 334-343
- Wang K, Babenko B and Belongie S. 2011. End-to-end scene text recognition//*Proceedings of 2011 International Conference on Computer Vision*. Barcelona, Spain: IEEE: 1457-1464 [DOI: 10.1109/ICCV.2011.6126402]
- Wang X L, Man Z P, You M Y and Shen C H. 2017. Adversarial generation of training examples: applications to moving vehicle license plate recognition [EB/OL]. [2022-08-11].
<https://arxiv.org/pdf/1707.03124.pdf>
- Xie D, Xiong J and Pu S L. 2017. All you need is beyond a good init: exploring better solution for training extremely deep convolutional

- neural networks with orthonormality and modulation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 5075-5084 [DOI: 10.1109/CVPR.2017.539]
- Yang L, Wang P, Li H, Li Z and Zhang Y N. 2020. A holistic representation guided attention network for scene text recognition. *Neurocomputing*, 414: 67-75 [DOI: 10.1016/j.neucom.2020.07.010]
- Zhang A, Chan A, Tay Y, Fu J, Wang S H, Zhang S, Shao H J, Yao S C and Lee R K W. 2021. On orthogonality constraints for Transformers//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: ACL: 375-382 [DOI: 10.18653/v1/2021.acl-short.48]
- Zhang S Z, Dinan E, Urbanek J, Szlam A, Kiela D and Weston J. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: ACL: 2204-2213 [DOI: 10.18653/v1/P18-1205]
- Zhang Y P, Nie S, Liu W J, Xu X, Zhang D X and Shen H T. 2019. Sequence-to-sequence domain adaptation network for robust text image recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 2735-2744 [DOI: 10.1109/CVPR.2019.00285]

作者简介

徐仕成,男,硕士研究生,主要研究方向为场景文本识别和计算机视觉。E-mail: lexiaoyuan@foxmail.com

朱子奇,通信作者,男,副教授,主要研究方向为计算机视觉、机器学习、防伪检测和信息安全。E-mail: zhuzq@wust.edu.cn