When CNNs Meet Vision Transformer: A Joint Framework for Remote Sensing Scene Classification

徐梓豪

摘要

场景分类是遥感图像解释的不可或缺的部分，目前已经探索了各种基于卷积神经网络 (CNN) 的方法来提高分类准确性。尽管它们在高分辨率遥感图像 (HRRS) 上已经显示出了良好的分类性能，但提取的特征的区分能力仍然有限。在这篇论文中，提出了一个结合 CNN 和 ViT 的高性能联合框架 (CTNet) 来进一步提高HRRS场景分类的特征的区分能力。CTNet 方法包括两个模块，包括 ViT 的流 (T-stream) 和CNN的流 (C-stream)。对于T-stream，将扁平化的图像块输入到预训练的 ViT 模型中，以挖掘 HRRS 图像中的语义特征。为了与 T-stream 互补，将预训练的CNN转移到C-stream中提取局部结构特征。然后，将语义特征和结构特征连接起来预测未知样本的标签。最后使用了一个联合损失函数来优化联合模型并增加类内聚合。

关键词：HRRS，CNN，ViT，CTNet

1 引言

由于大量遥感数据的产生，近年来许多图像解释任务都得到了快速发展，如图像分类、图像检索和语义分割[1]-[3]。其中，高分辨率遥感图像(HRRS)场景分类也成为了一个活跃的研究话题，其目的是提炼出语义级别的信息，并为给定的图像分配一个类标签[4]-[6]。

由于HRRS图像中复杂的物体结构和空间布局，提取有区分性的特征是提高分类准确度的最重要步骤[7]。对于场景分类任务，手工特征方法和基于深度学习的方法是两个主流的特征提取方法。前者主要侧重于设计单一的手工特征（如颜色特征和纹理特征）或融合多特征以提高分类结果[8]-[10]。然而，这些方法在表示复杂场景的固有属性上存在困难，因此，描述 HRRS 场景的能力是有限的。

2 相关工作

鉴于深度学习的快速发展，且被引入到 HRRS 场景分类领域后取得了良好的分类效果[11]-[13]。基于深度学习的方法可以分为以下三个分支：全训练方法、预训练模型方法和微调方法。

2.1 全训练方法

全训练方法对网络结构和参数的设计没有任何限制，这使得设计的模型与遥感数据更为相关。Bi等人[14]独立构建了一个多实例密集连接 ConvNet (MIDC-Net)，并考虑了低和中等特征用于航空场景分类。对于全训练方法，需要大量的标注数据从零开始训练设计的模型。尽管HRRS图像的数量很多，但标注的样本相对较少，这不利于全训练模型的收敛。

2.2 预训练模型方法

为了解决标注样本不足的问题，先前在自然图像数据集（例如，ImageNet）上预训

练的模型被转移用于HRRS场景分类。一些现有的研究将预训练模型视为特征提取器，然后重新编码这些特征以提高分类性能。Cheng 等人[15]提出了一种称为卷积特征包（BoCF）的特征编码方法，使用卷积神经网络（CNNs）提取的深度特征作为分类的视觉词。Hu等人[16]利用来自全连接（FC）层和最后一个卷积层的特征向量形成不同的图像表示。Li等人[17]致力于通过特征编码方法整合多层特征，以充分挖掘预训练模型的潜力。与全训练方法相比，基于预训练模型的方法只需要有限的标注样本和时间消耗。然而，由于自然图像和遥感图像之间存在巨大的差异，直接从预训练模型中提取的特征在描述 HRRS 场景时存在困难。

2.3 微调方法

与基于预训练模型的方法相比，对遥感数据进行微调预训练模型被认为是提高特征区分能力的更好策略。Chaib等人[18]通过判别相关性分析（DCA）细化了预训练VGG-Net的FC层的特征。Xu等人[19]开发了一个双流特征聚合深度神经网络(TFADNN)，其中包括有区分性特征的流和普通特征的流，用于提取有区分性的场景表示。Fang等人[20]利用循环卷积模块（CCM）有效地融合了频率和空间域产生的特征。尽管基于CNN的方法已经显著地提高了分类性能，但与HRRS图像的空间布局相关的关键长距离依赖性被忽略了。

2.4 Trasoformer

近年来，Transformer 在许多领域显示了学习序列之间长距离信息的潜力，如自然语言处理(NLP)和高光谱图像分类[21]。此外，Dosovitskiy等人[22]将 Transoformer 应用于计算机视觉领域，他们提出的 Vision Transoformer（ViT）在图像分类上表现出色。ViT 模型已经被用于场景分类，以探索 HRRS 图像的特征。Bazi等人[23]转移了经过数据增强和网络剪枝的预训练 ViT 模型，以提高分类性能。尽管 ViT 模型可以利用长距离语义表示，但在扁平化的过程中，结构信息被破坏了。

3 本文方法

3.1 本文方法概述

为了进一步提高特征的区分能力，本文提出了 CTNet，其框架如图1所示。如图1所示，CTNet 由两个不同的流组成，分别提取语义特征和结构特征。此外，提出了一个联合损失函数，用于优化端到端的模型。

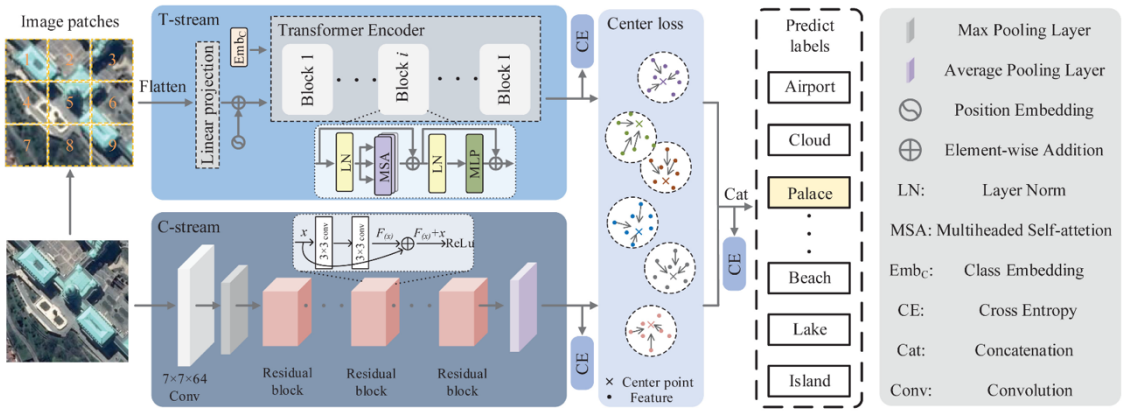

Fig. 1. Framework of the proposed CTNet method.

图 1.方法示意图

在 T-stream 中，一个输入的 HRRS 图像 I_m 被重构为一个扁平化图像块序列 $S_p \in R^{N\times(P_s^2\cdot I_c)}$，其中 $N = I_h \times I_w / P_s^2$ 表示块的数量，P_s 表示每个图像块的大小。然后，图像块序列通过线性投影进一步处理，将其映射到 D 维度的可训练嵌入矩阵 $E \in R^{(P_s^2\cdot I_c)\times D}$。为了分类的方便，嵌入的块必须与可学习的类嵌入 E_{class} 连接，其状态作为输出的 T-stream 中的最终表示 X_M^0。此外，为了记录扁平化块的位置并保持它们的位置信息，位置嵌入 $E_{pos} \in R^{(N+1)\times D}$ 被引入并添加到块的表示中。生成的ViT输入与类标记和位置嵌入 X_0的结合表达如下：

$$X_0 = \left[E_{class}; S_p^1 E; S_p^2 E; \ldots; S_p^N E\right] + E_{pos}$$

紧接着的模块名为Transformer编码器，是对高分辨率遥感（HRRS）图像进行分类的最重要步骤。当输入表示 X_0形成后，M 个块被用来探索 HRRS 图像的语义特征。如图1所示，每个块主要包含两个成分，包括交替的多头自注意力（MSA）层和多层感知机（MLP）。第 m 个块的中间变量 X'_m 和输出 X_m 的计算公式如下：

$$X'_m = MAS\left(LN(X_{m-1})\right) + X_{m-1} (m = 1, \ldots, M)$$
$$X_m = MLP(LN(X'_m)) + X'_m (m = 1, \ldots, M)$$

值得注意的是，采用了残差连接来传递 HRRS 图像的信息，且层归一化（LN）在两个结构之前都已经被使用。T-stream的最终输出 F_T 可以通过下列公式获得：

$$F_T = LN(X_m^0)$$

作为一种广泛使用的卷积神经网络（CNNs），ResNet34被选为例子来说明C-stream（结构特征流）的原理。ResNet34的最后一个全连接（FC）层被移除，其余层作为特征提取器来执行。ResNet34的主要组件是残差块，它由带有快捷连接的卷积层堆叠而成，残差块的输出 y 通过修正线性单元（ReLU）表达如下：

$$y = ReLU(\mathcal{F}(x, \{w_i\}) + x)$$

其中 x 代表每个残差块的输入，$\mathcal{F}(x, \{w_i\})$ 是可学习的残差映射。因此，对于每张高分辨率遥感（HRRS）图像 $I_m = R^{H\times W\times C}$，C-stream的输出表示 F_C 被获取来捕获结构特征。

为了更好地优化 CTNet，提出了一个联合损失函数，它包含三个交叉熵损失 L_{CE} 和两个中心损失 L_C[24]，设计用于在训练阶段计算反馈误差。联合公式定义如下：

$$L_j = L_{CE}^{CT} + L_{CE}^{T} + L_{CE}^{C} + \lambda L_T^C + \lambda L_C^C$$

其中 L_{CE}^{CT} ，L_{CE}^{T}和 L_{CE}^{C} 分别提高联合模型，T-stream 和 C-stream 的分类准确率。L_T 和 L_C 有助于促进每个流中类内特征的聚合，而 λ 代表权重系数。L_C 的操作计算如下：

$$L_C = \frac{1}{2}\sum_{i=1}^{S_b} \left\|F_i - C_j\right\|^2$$

其中 F_i 表示每个流的第 i 个特征，S_b 表示小批量的大小，C_j 是第 j 个类的中心。

4 复现细节

4.1 与已有开源代码对比

该论文并并没有给出源代码，所以本次复现的代码除了Vision Transformer 的主体

部分使用了Github上的源码，其余部分均为自己实现。由于GPU的不同，原文 batch 的大小为 24，复现时使用的 batch 为 35。

4.2 实验环境搭建

本次的实验所使用的是 pytorch，虚拟环境为conda，GPU为 RTX 4090。

5 实验结果分析

本次实验使用的数据集是来自西北工业大学的 RESISC45 数据集，该数据集由45个类别构成，每个类别有700张图像，图像的大小为256*256。实验结果显示损失函数大约会在 30 个 epoch 左右收敛，总计花费一个小时多一点，在测试集上的准确率可以接近 93%，与论文给出的结果相差了两个百分点。

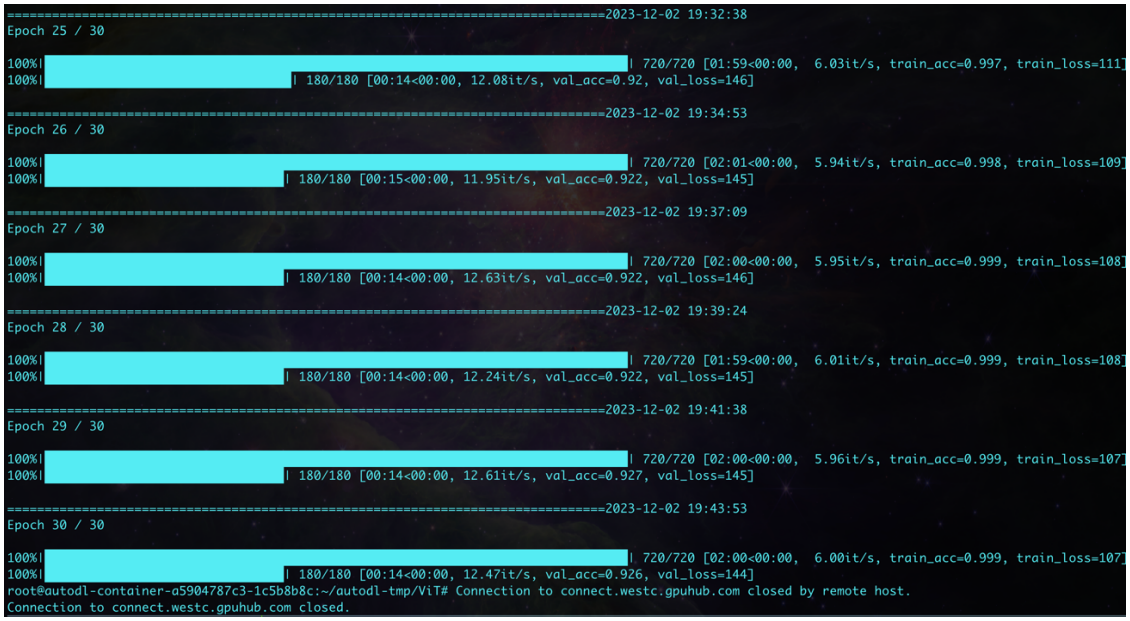

图 2. 实验结果示意

6 总结与展望

在这篇论文中，提出了一个结合 CNN 和 ViT 的高性能联合框架（CTNet）来进一步提高HRRS场景分类的特征的区分能力。CTNet 方法包括两个模块，包括 ViT 的流（T-stream）和CNN的流（C-stream）。并且使用了一个联合损失函数来优化联合模型并增加类内聚合。复现的实验结果与原文给出的结果还相差了两个百分点，因此还存在许多不足的地方需要改进和优化。

参考文献

[1] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100-111, Jan. 2020.

[2] F. Luo, L. Zhang, B. Du, and L. Zhang, "Dimensionality reduction with enhanced hybrid-graph discriminant learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5336-5353, Aug. 2020.

[3] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A bench- mark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 197-209, Nov. 2018.

[4] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classifi- cation: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865-1883, Oct. 2017.

[5] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965-3981, Jul. 2017.

[6] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1894-1898, Nov. 2020.

[7] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916-6928, Sep. 2019.

[8] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Inf. Sci.*, vol. 348, pp. 209-226, Jun. 2016.

[9] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual- words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747-751, Jun. 2016.

[10] H. Huang and K. Xu, "Combing triple-part features of convolutional neural networks for scene classification in remote sensing," *Remote Sens.*, vol. 11, no. 14, p. 1687, Jul. 2019.

[11] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.

[12] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8905–8918, Dec. 2020.

[13] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[14] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple- instance densely-connected ConvNet for aerial scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4911–4926, Mar. 2020.

[15] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

[16] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.

[17] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[18] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.

[19] K. Xu, H. Huang, P. Deng, and G. Shi, "Two-stream feature aggregation deep neural network for scene classification of remote sensing images," *Inf. Sci.*, vol. 539, pp. 250–268, Oct. 2020.

[20] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Robust space–frequency joint representation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7492–7502, Oct. 2019.

[21] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspec- tral image classification," *Remote Sens.*, vol. 13, no. 3, p. 498, Jan. 2021.

[22] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–21.

[23] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, p. 516, Feb. 2021.