ICPR MTWI 2018 挑战赛三比赛报告

队名：nelslip(iflytek&ustc)

团队成员：张建树，朱意星，杜俊，戴礼荣，陈明军，

吴嘉嘉，胡金水

一、 算法介绍

因为该挑战赛需要识别系统从一整张图中识别出所有存在的文本行，所以我们的策略是首先使用检测器将整图中所有可能的文本行检测出来，再在检测后的文本行上采用文本行识别器识别。

关于检测器的实现，我们的使用的方法基于 FPN [1]，并且我们使用了 PANet [2] 的方法扩展了网络，由于文本行有可能是任意形状的四边形，针对文本行的特殊性，我们使用了 SLPR [3] ，具体而言我们利用水平和垂直的滑动线条去等距地拟合文本行的边缘，这样可以解决任意四边形框四个顶点标注的歧义性，之后我们使用文本行的轮廓去生成旋转矩形，事实上类似于 Cascade R-CNN [4]，我们使用了两次 R-CNN，第一步我们使用 RPN 生成水平矩形框，第二步我们使用 SLPR 生成的轮廓得到旋转矩形框，最后再次拟合四个顶点坐标。

关于文本行识别器的实现，我们使用的基本算法框架是基于注意力机制的编解码模型，其中编码器是一个多层卷积神经网络，其用来从原始网络图片中提取高维视觉特征。解码器是一个以 GRU 为基本单元的单向循环神经网络，其用来将编码器提取出的高维视觉特征转换为字符串，且每次只解出一个字符。我们选择使用 DenseNet[5]框架作为我们的卷积神经网络编码器。采用 DenseNet 作为编码器的基于注意力机制的编解码模型的具体实现在文献[6]中有详细的介绍。为了让我们的模型更好的适应文本行识别问题，我们在卷积神经网络编码器之后又拼接了两层双向 GRU 层，以此实现对文本行的时序信息建模。本次挑战赛我们算法最重要的核心点在于使用了部件分析网络[7]来处理文本行识别。我们选用部件分析网络的原因在于本次文本行识别任务中，汉字的文本行识别占主导，且汉字类别分布极不均匀，而部件分析网络是通过分析汉字内部的组成部件及部件之间的空间结构来识别汉字，这种做法相比于将一整个汉字作为一个类别来识别有两大优势：其一，模型最终的分类类别数大大减少，4000 多类的汉字类别被简化到 500 多类的汉字部件类别；其二，这种依赖分析汉字部件的识别方式使得模型具备了识别集外汉字和低频汉字的能力。我们首先将每个汉字根据其固有的部件结构拆解成 IDS(Ideographic Description Sequences，可在维基百科检索到)字符串，而部件分析网络的任务则是通过注意力机制来分析汉字内部的部件组成和空间结构，最终解出汉字的 IDS 字符串，解出 IDS 字符串后再重新匹配成完整汉字。在文本行识别任务中，我们在每个汉字 IDS 字符串之后添加一个 end-of-word(eow)的标志来分隔每个汉字，以此实现基于汉字部件的文本行识别。此外，本次比赛的网络图片背景十分复杂，不少图片在 RGB 三通道下很难识别，即便是人眼也很难识别，因此我们还

提取了原始图片的 HSV 三通道，与 RGB 通道一起送入卷积神经网络训练。最后，由于本次文本行识别任务还包含了旋转的文本行识别，所以我们在网络训练过程中添加了文本行旋转的数据增强。

二、 模型结构

我们增加了一些策略来使得我们的检测器获得更好的性能，包括数据增强（旋转，缩放），多尺度测试，我们使用了在 ImageNet 上预训练的 ResNeXt-101 (32*8d) [8]作为网络的基础，并且只使用了单一模型。

我们文本行识别器的编码器由 DenseNet 和双向 GRU 拼接而成，DenseNet 参考[5]中的网络配置，我们使用的 DenseNet 总共有 171 层，增长率为 24，使用了瓶颈层和压缩层；紧接 DenseNet 之后的双向 GRU 层每层含有 512 个节点数。解码器由一个单向的 GRU 组成，该层含有 256 个节点数，且该解码器含有一个带全覆盖机制的注意力模型[9]。我们随机初始化训练了 12 个模型，其中 3 个模型利用了 HSV 三通道信息，3 个模型添加了文本行旋转作为数据增强。在测试的时候，我们在 beam search 的每一步融合了这 12 个模型。

三、 开发环境

我们使用了 caffe2 作为实验平台训练我们的检测器模型，所有关于检测的实验都是在 1080Ti GPU 上进行的。我们训练文本行识别模型的所有实验是在 Theano 0.10.0 上实现，英伟达 Tesla 系列 P40GPU，显卡大小为 24G。

四、 数据集

本次比赛官方提供了 1 万张图片，我们将这 1 万张图片分出 9 千张作为训练集，另外 1 千张作为开发集。训练检测器的过程中我们没有使用额外数据。训练文本行识别器时，我们从 9 千张图片截取出了 128,210 文本行用于训练模型，从 1 千张图片中截取出了 15,288 文本行用于开发验证。除此以为，我们还额外收集了 25 万的自然场景下的文本行并且人工标注了这些文本行用于训练。这额外的 25 万文本行含有 18 万的中文文本行和 7 万的英文文本行。

五、 团队介绍

本次参赛团队是由中国科学技术大学语音及语言信息处理国家工程实验室和科大讯飞合作组成，其中张建树，朱意星，杜俊，戴礼荣为中国科学技术大学语音及语言信息处理国家工程实验室成员，陈明军，吴嘉嘉，胡金水为科大讯飞成员。我们希望通过参加本次比赛让外界关注到我们在 OCR 领域和检测领域的技术。

本团队的队长为张建树，其邮箱地址为 xysszjs@mail.ustc.edu.cn，手机号码为 15856910468，现住址为中国安徽省合肥市蜀山区黄山路 443 号中国科学技术大学西校区 8 号宿舍楼 640 室。

六、 参考文献

[1] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in CVPR, vol. 1, no. 2, 2017, p. 4.
[2] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” arXiv preprint arXiv:1803.01534, 2018.

[3] Y. Zhu and J. Du, “Sliding line point regression for shape robust scene text detection,” arXiv preprint arXiv:1801.09969, 2018.
[4] Z. Cai and N. Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” arXiv preprint arXiv:1712.00726, 2017.
[5] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 1, no. 2, 2017, p. 3.
[6] J. Zhang, J. Du, and L. Dai, “Multi-scale attention with dense encoder for handwritten mathematical expression recognition,” arXiv preprint arXiv:1801.03530, 2018.
[7] J. Zhang, Y. Zhu, J. Du, and L. Dai, “RAN: Radical analysis networks for zero-shot learning of Chinese characters,” arXiv preprint arXiv:1711.01889, 2017.
[8] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017, pp. 5987–5995.
[9] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai, “Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition,” Pattern Recognition, vol. 71, pp. 196–206, 2017.