基于 TextCaps 的视觉辅助识别应用

李涵鑫 李丹
四川大学锦城学院 计算机与软件学院 四川 成都 610000

【摘 要】中国目前存在着近乎两千万的视障人士，虽然中国的盲道长度居于世界第一，但盲道使用率却不甚理想。这其中，盲道的占用和导航设备的笨重成为了视障人士的主要障碍，因此在计算机视觉领域有一些先进的模型来解决这些路面识别的问题，但是大多依赖于笨重或贵重的仪器，因此本文采用一种轻量级的图像识别模型（TextCaps）来进行对于视障人士的路面情况识别框架。TextCaps 是基于胶囊网络的一种基于超小型数据集的模型，其不仅保留了胶囊网络中对于空间特征的抓取，并且有效的解决了图像质量问题和图像数据集过小的问题。本文利用了 TextCaps 的空间能力，对于数据集较小的不同情况的路面进行分类，其中包括不同的灯光、障碍物和室内，室外的情况，能够有效的抓取其中的空间特征。在低精度 RGB 图像中，TextCaps 和胶囊网络的分类精准率分别是 66% 和 51%，而在传统 CNN 架构中，例如 VGG16，最高达到了 84%。

【关键词】胶囊；图片识别；视觉辅助

1 介绍

对于大多数模型来说，图像识别需要大量被标记的数据集以及相当强大的硬件计算能力，随之而来的就是实际设备的笨重和贵重，这使得应用在硬件上变得十分困难。尽管在计算机视觉领域图像识别已经有了很好的结果，但是我们希望能消耗更少的计算资源而得到相同的结果。

一个实体通过大量的卷积和池化操作的过程会忽略图片的空间特征，而只会着重图片的具体特征。许多模型仅仅只是提取到图片的单独特征就判断其是属于哪一个分类，这并不以利于许多依赖空间特征的任务。在视觉辅助识别应用中，我们需要模型着重于提取物体的空间特征才能够得到一个合理的预测结果，因此我们采用了胶囊网络的方法，通过借助向量的维度来存储更多有关于图像本身的空间特征。为了解决数据集过小的问题，我们采取了 TextCaps 模型，其主要是借助胶囊网络中的实例化参数的特点来达到增强数据的作用。我们的方法是基于 Hinton 提出的 CapsNets 模型。而 TextCaps 对于 CapsNets 做出了稍微的修改，其不仅利用反卷积网络代替了全连接解码器网络，而且对实例化参数添加可控的随机噪声，以便于更好的显示图片特征。

本文的主要工作在于将这项模型应用到 RGB 图像当中，实现了在较少的数据集下对于道路情况的图像识别，并且对于不同数据集下的 TextCaps 模型进行了对比分析，比较了在不同模型下对于道路图像识别的优劣，并进行了对比分析。

2 模型介绍

我们采取了 CapsNets 来达到更好的识别效果，在 2.1 节会详细介绍这种模型的架构。为了获取到更多数据集，我们采用了数据增强技术，在 2.2 节我们会详细介绍这种网络的结构。

2.1 基于胶囊网络的图像识别

为了更好的进行针对视障人士的道路障碍识别，我们采用了胶囊网络的网络架构，其主要是由胶囊网络和解码器网络组成。

2.1.1 胶囊网络

胶囊网络的前三层是卷积层，分别是步幅为 1 的 64 个 3 × 3 的内核；步幅为 1，128 个 3 × 3 的内核；步幅为 2，256 个 3 × 3 的内核。然后是主胶囊层，其由 32 个 8 维的胶囊组成，其中胶囊是一组向量神经元。每个胶囊是八个卷积层的堆叠在一起，每个卷积层的维度是 2 × 2，步长为 2 的 9 × 9 的内核。该组向量通过动态路由算法进入字符胶囊层。

路由协议算法首先将低级别到高级别的胶囊的预测结果置为 0，然后进入循环迭代更新，通过 softmax 将 bi 转换成概率 ci，然后第 l+1 层所有的向量加权求和，低层所有向量的共识输出，接着将数据归一到 [0,1]，表示低层到高层胶囊的概率，最后新的 bij 等于 Ujli 和 vj 的积，前者代表的是当前向量对 vnj 的个人预测，后者代表的是所有低层向量对于 vnj 的共识预测。

字符胶囊层是一个全连接层，其输出的 16 维向量即该实体的实例化参数。

2.1.2 解码器网络

解码器网络包含一个全连接层，以及五个反卷积层，其输入是实例化参数，然后输出重建图像，全连接层和前四个反卷积层的激活函数 ReLU 激活函数，最后一个反卷积层采用 sigmoid 激活函数。为了更好的解决 RGB 图像下的图像重建，本文做了一点改动，即采用了五个 3 维的反卷积层来实现重建一个 RGB 图像。

2.2 基于 TextCaps 模型的图像识别

我们采用了基于胶囊网络的 TextCaps 模型架构，不仅保留了胶囊网络的特征而且提供了一种数据增强技术。

数据增强技术方法的提出来自胶囊网络中的实例化参数概念。

我们用原数据集训练 CapsNet，得到一个其对应的实例化参数，将实例化参数进行遮盖，仅仅保留其最大概率类别的实例化参数，将其进入解码器进行训练从而得到输出的重建图像集。对重建图像集进行锐化操作，将原数据集和锐化后的重建数据集进行组合，利用这个新的数据集可以训练得到一个新的解码器。我们采用扰动算法对于实例化参数和新得到的数据集进行实例化参数扰动，最终输出一个新的扰动后的数据集 Ipre，将扰动后的数据集和原训练数据集组合就形成了新的数据集。

扰动算法首先计算每个实例化参数的方差，对方差进行降序排列，取得方差最大的两个实例化参数，并且扫描出最大和最小的实例化参数，得到其均值点，作为最大噪声。计算所有最大噪声点，得到平均最大噪声点。如果实例化参数大于 0，那么就增加平均最大噪声和最大噪声之间的最小值，否则就减去平均最大噪声点和最大噪声之间的最小值。最后将实例化参数输入到重训练后的解码器，输出扰动后的图像。

3 数据集

3.1 视障人士道路数据集（blind）

在视障人士的日常使用中，智能手机置于胸前，定时拍下面前的道路图像，我们将道路图像分成了“清晰图像”和“非清晰图像”两类。其中包含了 342 张图片数据，调整样本为 28×28 像素。其中 171 张训练集图片和 171 张测试集图片，该数据集包含了屋内和屋外的情况，包括了干和湿的地板，日光和人造光，以及不同类型的障碍物，例如楼梯、树木、洞、动物、交通锥等，如图 1 所示。

图 1 视障人士道路数据集（部分）

3.2 CT-COVID 数据集

对于 COVID 肺部图像 CT 数据集，我们将其分成了患有肺炎和未患有肺炎两类。其中包含了 746 张图片数据，我们将其调整为 28×28 的 jpg 图片，其中 395 张训练图像以及 351 张测试图像，该数据集包含了肺炎患者和非肺炎患者肺部 CT 的各种情况。如图 2 所示。

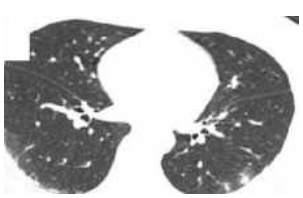

图 2（1）：肺炎患者肺部 CT

图 2（2）：非肺炎患者肺部 CT

4 实验结果和分析

为了评估视障人士道路情况数据集在 TextCaps 模型上的性能，我们加入了多种传统 CNN 网络架构对其进行平均，将结果进行对比分析，在 4.1 节我们会对其进行具体描述和分析。TextCaps 的优秀性能作用于手写数据集，但是在 RGB 图像上的性能还是未可知的，因此我们在第 4.2 节利用两种 RGB 数据集对于 TextCaps 进行性能评估。

4.1 视觉辅助图片分类结果分析

如表 1 所示，在 17 种 CNN 架构中，我们能够发现 VGG16 和 VGG19 的表现明显优于其他的传统 CNN 架构，分别最高达到了 83.29% 以及 81.36%，而其他的模型在图像上的表现仅仅只在于 50% 上下，和人类的直觉判断概率相似，因此如果不经过某种改变的话，传统的模型将不能胜任这种分类任务。

在 TextCaps 中，我们可以清晰地看到准确率达到了 66.6%，明显高于除了 VGG16 和 VGG119 的 CNN 模型，并且优于人类的直觉判断，因此，TextCaps 在空间特征能力的获取上明显优于其他大部分的 CNN 架构，尤其还是在耗费计算资源如此小的模型中，依然能够有着优异的性能，但是我们也能看到 TextCaps 性能远远低于 VGG16 和 VGG19，我们分析由于图像的像素过小，本身的空间信息便不足以支撑更好的模型预测，但是高精

准度的 RGB 图像所依赖的更大的资源消耗，因此我们倾向于能够将这个模型更好的应用在 RGB 图像空间中。

表 1：11 种模型关于视障人士道路情况数据的准确度

结构	精确度	结构	精确度
Xception	49.43%	InceptionV3	51.18%
VGG16	83.39%	InceptionResNetV2	51.18%
VGG19	81.36%	MobileNet	51.48%
ResNet50	51.17%	DenseNet121	51.18%
ResNet101	51.18%	DenseNet169	51.45%
ResNet152	48.05%	DenseNet201	48.34%
ResNet50V2	51.19%	NASNetMobile	51.78%
ResNet101V2	51.18%	MobileNetV2	50.90%
ResNet152V2	51.18%	TextCaps	66.60%

4.2 TextCaps 在 RGB 空间性能分析

如表 1 所示，TextCaps 在视障人士道路情况数据集的表现虽然精准度不高，但是对于大部分 CNN 模型的表现高出一截，但是一个数据集并不能完全表现 TextCaps 模型在 RGB 空间的表现，由于该模型会受到数据集的大小和空间信息的影响，因此我们选择新型冠状肺炎患者肺部 CT 数据集来对 TextCaps 的 RGB 性能表现进行对比。

如表 2 所示，视障人士道路情况数据集的准确度达到了 66.6%，而 CT–COVID 的准确度仅达到了 52.99%。TextCaps 在两个数据集的表现都不是非常优异，其性能表现仅仅只是高于一般 CNN 模型。因此，虽然 TextCaps 模型确实能够抓取到图像的空间特征，但是由于图像像素信息太少，仅仅是在 28 × 28 的图片集上，模型不能够很好的还原出重建图像，两种数据集的重建图像如图 3，因此如果能够利用更多的计算资源在更高精度上的 RGB 图像进行分析和重建，我们或许能够得到更好的结果。

如表 2 所示，TextCaps 在 blind 数据集上的精确度相比于在 CT–COVID 数据集的精确度高出 13.7%，从图 3 中我们可以从两种数据集的重建图像可以清晰的看出，blind 数据集的重建图像精度明显高于 CT–COVID 数据集的精度，也就是说在重建成果上的不同导致了两者在测试准确度上的差异。根据原始图像数据图 1 和图 2 可以分析得到，blind 数据集中的空间特征更为明显，具有房屋环境，灯光昏暗、大小，路面平整，障碍物间隔、类型等多种空间特色，而 CT–COVID 数据集的图像只是出现了其肺部的图像，而无论是肺炎患者和非肺炎患者，他们的肺部 CT 图像仅仅是在肺部上有着微小的差异，并不存在明显的空间特征的差异。因此我们可以认为 TextCaps 在 RGB 图像上对空间特征强烈的图像有着更好的表现。

表 2：基于 TextCaps 的两种数据集的准确度

数据集	精确度
Blind	66.60%
CT–COVID	52.99%

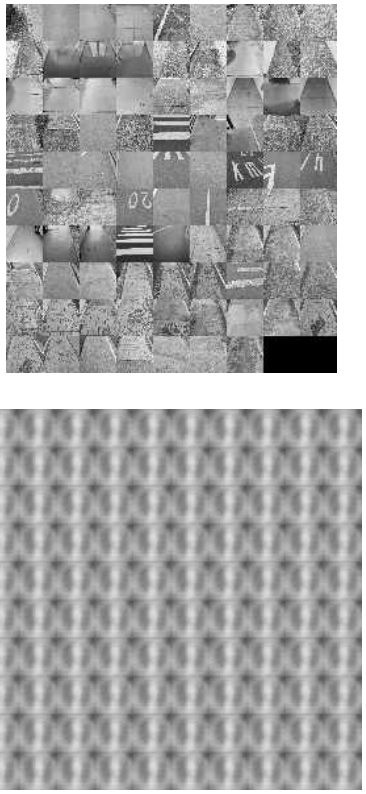

图 3：两种数据集的重建图像

5 结论

本文我们主要应用视障人士道路情况数据集在 TextCaps 模型，以此解决再视障人士再日常生活中的导航问题。基于此模型，我们能够发现数据集在模型的准确度达到了 66.6%，这个准确度虽然不算很高，但是也由于其他大部分的卷积神经网络模型，而此模型能够大大节省计算机资源，并且可以很好的应用在智能设备中。我们希望以此能够帮助视障人士进行导航。

我们进一步分析了 TextCaps 模型在于 RGB 空间上的性能表现。虽然囿于硬件的计算能力不足，我们仅仅在低精度下的 RGB 图像上进行了 TextCaps 在不同数据集上的分析对比，但是我们可以清晰的发现，TextCaps 在 Blind 和 CT–COVID 数据集上的性能表现分别是 66.6% 和 52.99%，这表明 TextCaps 在 RGB 图像上能够很好的抓取其中的空间特征，因此我们希望能够在高精度的 RGB 图像上利用大量的计算资源进行分析，以此能够获得更高的精确度。

【参考文献】

[1] BREVE, Fabricio Aparecido; FISCHER, Carlos Norberto. Visually Impaired Aid using Convolutional Neural Networks, Transfer Learning, and Particle Competition and Cooperation In: 2020 International Joint Conference on Neural Networks (IJCNN 2020), 2020, Glasgow, UK. Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN 2020), 2020.

[2] Jayasundara V , Jayasekara S , Jayasekara H , et al. TextCaps: Handwritten Character Recognition With Very Small Datasets[C]// 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019.

[3] Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: NIPS, Long Beach, CA (2017) 3856-3866.

[4] Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming auto-encoders. In: ICANN, Berlin, Heidelberg (2011) 44-51.

[5] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 1800 - 1807.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." Computational and Biological Learning Society, 2015, pp. 1-14.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 770 - 778.

[8] "Identity mappings in deep residual networks," in Computer Vision - ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630 - 645.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 2818-2826.

[10] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Thirty-first AAAI conference on artificial intelligence, 2017.

[11] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[12] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 4700 - 4708.

[13] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018, pp. 8697-8710.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018, pp. 4510-4520.