****



**本 科 毕 业 论 文**

院 系 计算机科学与技术系

专 业 计算机科学与技术

题 目 基于tensorflow的深度学习肖像抠图算法

年 级 本科四年级 学 号 141220080

学生姓名 浦溪

指导教师 顾荣 职 称 讲师

提交日期 2018.5.15

**南京大学本科生毕业论文（设计、作品）中文摘要**

题目：基于tensorflow的深度学习肖像抠图算法

院系：计算机科学与技术系

专业：计算机科学与技术

本科生姓名：浦溪

指导教师（姓名、职称）：顾荣 讲师

摘要：深度学习由于其强大的表达能力和拟合能力，近几年在机器学习，模式识别等领域表现出优秀的性能。抠图属于语义分割的范畴，对于人物肖像的抠图算法，即将人物肖像和背景分割，涉及的主要难点是背景的复杂性以及肖像细节部分的分割。通过深度学习的方法进行人物肖像抠图能够取得远超过传统算法的准确率。我采用标注好的公开肖像数据集，首先用全卷积神经网络模型进行训练，然后进一步提出加入Resnet结构，对边缘区域进行模糊处理等改进模型，在IOU的性能度量上获得提高。

我采用Tensorflow作为训练框架，利用其并行化的特点以及pipeline数据流技术，大幅提高训练效率。由于深度学习所需要的大量数据，数据经常存储在不同的存储系统中，而且深度学习上层的应用框架多样化，我采用Alluxio为数据访问统一接口，使得不同的应用框架可以透明地读取多种存储系统中的数据，并且利用Alluxio的内存加速数据的访问速度。

关键词：深度学习；语义分割；抠图；全卷积神经网络；边缘模糊化; Tensorflow; Alluxio

**南京大学本科生毕业论文（设计、作品）英文摘要**

THESIS：Deep Portrait Segmentation methods based on tensorflow

DEPARTMENT：Computer Science and Technology

SPECIALIZATION: Computer Science

UNDERGRADUATE: Xi Pu

MENTOR: Rong Gu

ABSTRACT：Deep learning has been the state of art in the field of machine learning and pattern recognition because of its strong and robust learning ability. Portrait segmentation is a branch of semantic segmentation. It aims to generate pixel-level prediction as portrait(foreground) or background. Due to the complexity of various backgrounds and the detail area such as clothes and hair, it is challenging to achieve high accuracy in this task. Deep learning methods have far exceeded the performance of traditional methods in semantic segmentation tasks. So I choose the fully convolutional networks and train the model on the labelled portrait dataset. Then I apply the Resnet structure to improve the performance and use a soft boundary to make the model more boundary-sensitive.

I use Tensorflow as the framework and take advantage of its parallel characteristics and pipeline data stream to improve the efficiency of training and predicting. Due to the large volume of data in deep learning, we usually have to store these data in different storage systems. Alluxio transparently connects to existing storage systems and presents them as a single interface to users. Also, it provides local caching of frequently used data which makes deep learning training more cost effective and take less time.

KEY WORDS: deep learning; semantic segmentation; portrait segmentation; FCN; soft boundary; Tensorflow; Alluxio

目录

[1 引言 5](#_Toc513059206)

[1.1 研究问题及研究目的 5](#_Toc513059207)

[1.2 研究背景和相关工作 6](#_Toc513059208)

[1.2.1 语义分割 6](#_Toc513059209)

[1.2.2 肖像抠图 9](#_Toc513059210)

[2 流程和方法 10](#_Toc513059211)

[2.1 全卷积神经网络 10](#_Toc513059212)

[2.2 加入Resnet结构 12](#_Toc513059213)

[2.3 边缘模糊化处理 12](#_Toc513059214)

[2.4 数据增强 12](#_Toc513059215)

[3 实验结果 12](#_Toc513059216)

[3.1 性能指标IOU 12](#_Toc513059217)

[3.2 实验对比 12](#_Toc513059218)

[4 分布式存储和训练加速 12](#_Toc513059219)

[4.1 Tensorflow加速训练和预测 12](#_Toc513059220)

[4.2 Alluxio统一接口 12](#_Toc513059221)

[5 结论和后续工作 12](#_Toc513059222)

[5.1 实验中的不足 12](#_Toc513059223)

[5.2 后续的工作 12](#_Toc513059224)

[参考文献 13](#_Toc513059225)

# 引言

1.1 研究问题及研究目的

图像的语义分割是计算机视觉领域一个很基础的问题，语义分割的目的是将图片中属于不同类别的物体以像素点的精度标注出来。语义分割是很多计算机视觉任务的第一步要求，高精度的分割使得一些应用能够得到实际价值，例如自动驾驶中的物体识别，图像搜索引擎，医疗图像的精确识别，增强现实等等。与通常的语义分割任务相比，人物肖像的分割对精度有更高的需求，比如人物的边缘，包括衣服，头发等细小的区域，而且一般来说，人物的背景具有多样性和复杂性，将肖像和背景区分开有较大的难度。人物肖像的分割提取可以理解为，对一张肖像图进行像素级别的预测，对每一个像素点判断其属于前景(人物)还是背景。在人物肖像的分割中，对边缘区域的高精度识别是最重要的也是最具挑战性的问题，对边缘的准确的，平滑的标注可以带来更好的视觉效果和更方便的后续处理操作。

对人物肖像进行分割后，可以很容易的进行一些简单的应用，如图1.1所示[1]，(a)为输入图像，(b)为进行分割后的结果，白色为识别出的人物肖像前景，黑色为背景。图(c)-(e)为对分割后的结果进行不同的后续处理的图像，分别为风格化的人物肖像，深度背景图，卡通化的人物肖像。如图1.2所示，在完成肖像分割后，可以很容易地替换背景，将人物肖像和不同的背景相互组合。



图1.1

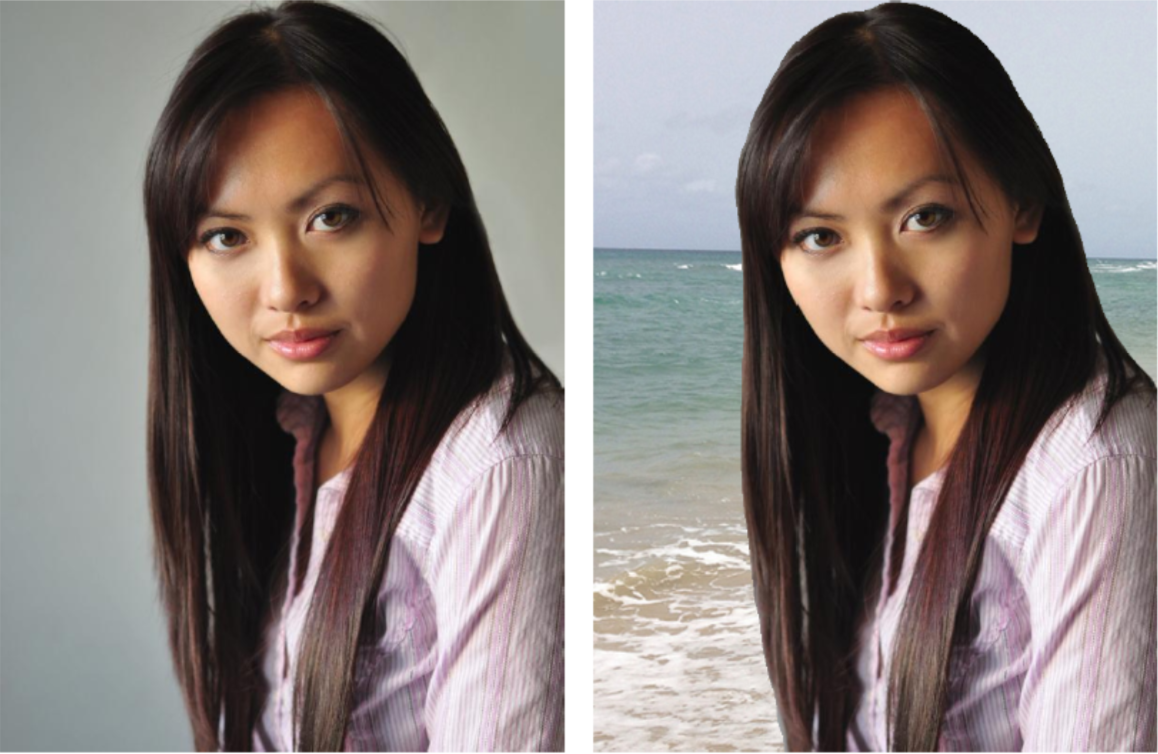


图1.2

1.2 研究背景和相关工作

1.2.1 语义分割

目前进行语义分割的主要方法是基于深度学习的模型，深度学习模型往往是作为特征提取的工具，可以不需要先验知识地提取图像训练的高维特征，然后选取分类器进行分类。常见的深度学习模型也经常被用作语义分割的一个模块进行使用，所以我先介绍几种经典的深度网络结构。

**AlexNet [2]**

AlexNet是2012年提出的深度卷积神经网络模型，在ILSVRC-2012数据集上获得 84.6%的top-5准确率。AlexNet的网络结构较为简单，包含五层卷积层，一层池化层，3层全连接层，使用ReLu作为激活函数。

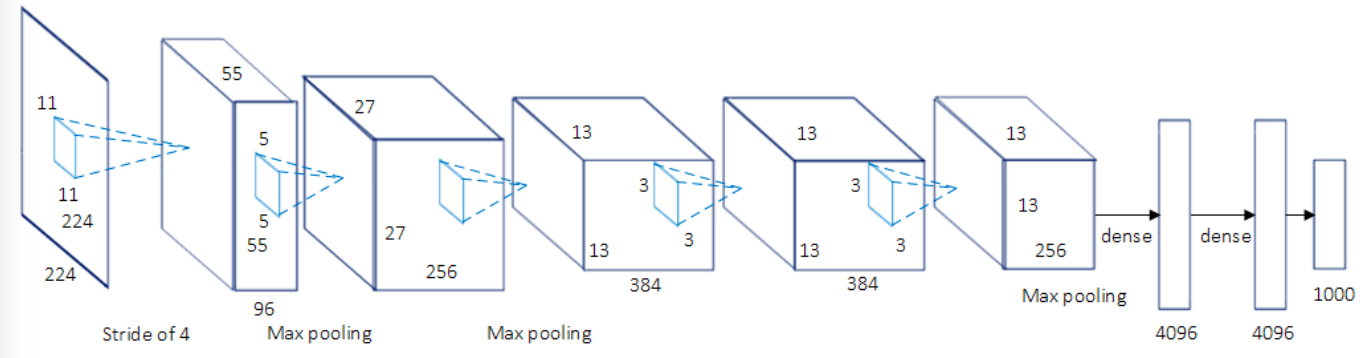


图1.3

**VGG [3]**

VGG将连续的几层卷积组成一个模块，每一个模块后连接池化层，VGG用更小的卷积核和更深的网络结构来提取特征，这样可以大幅减小参数数量，同时更容易拟合非线性函数。VGG-16包含16个卷积层，VGG-19包含19个卷积层。VGG-16在ILSVRC-2013数据集上获得了92.7%的top-5准确率。

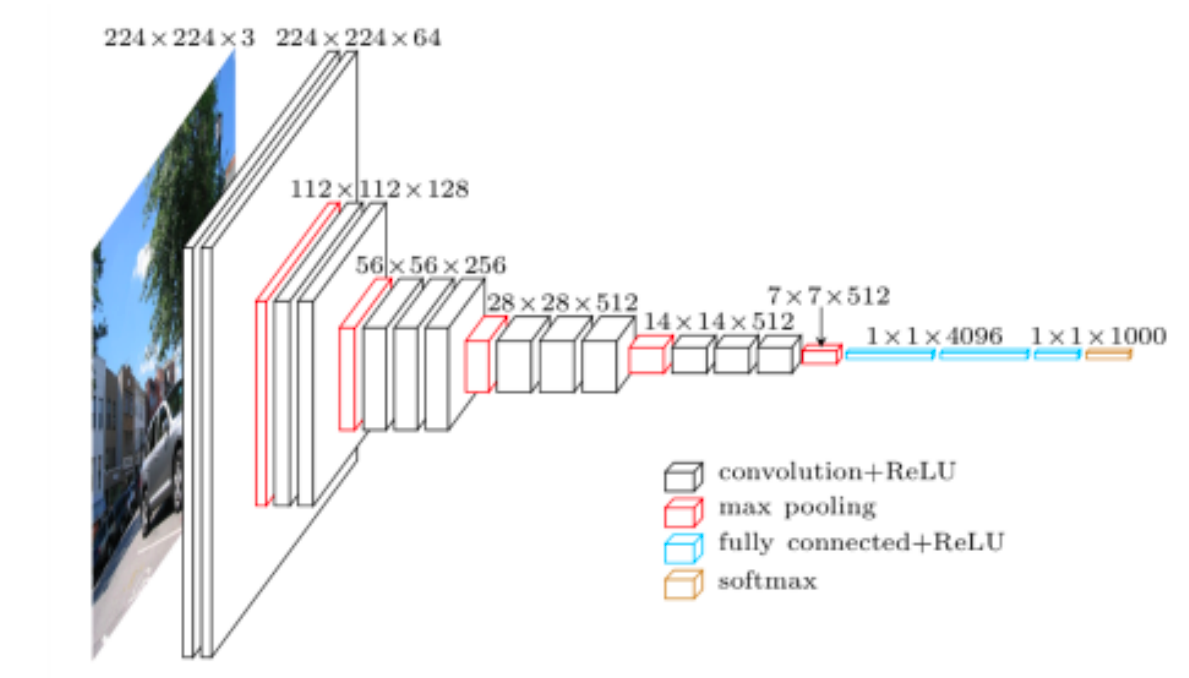


图1.4

**GoogleNet [4]**

GoogleNet是一个很复杂的深度网络模型，包含22层和一种感知模块的的部分。GoogleNet的主要特点是用不同的方式组合网络，包括一种网络中的网络(NiN)的结构，同时采用了1X1的卷积层来提取特征。GoogleNet在ILSVRC-2014的数据集上获得了93.3%的top-5准确率。

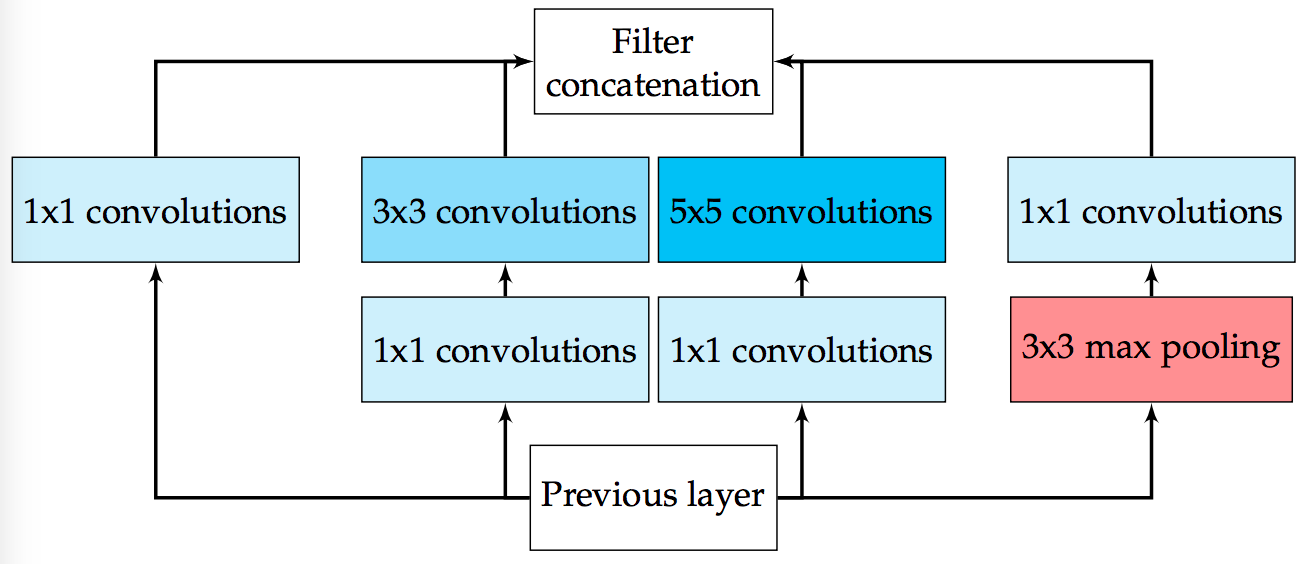


图1.5

**ResNet [5]**

ResNet的主要贡献是提出了残差模块(residual blocks)，在层与层之间引入了跳跃结构，使得模型只需要学习与上一层不同的特征，这种结构使得可以搭建非常深(152层)的网络，但是克服了原先存在的梯度消失问题。

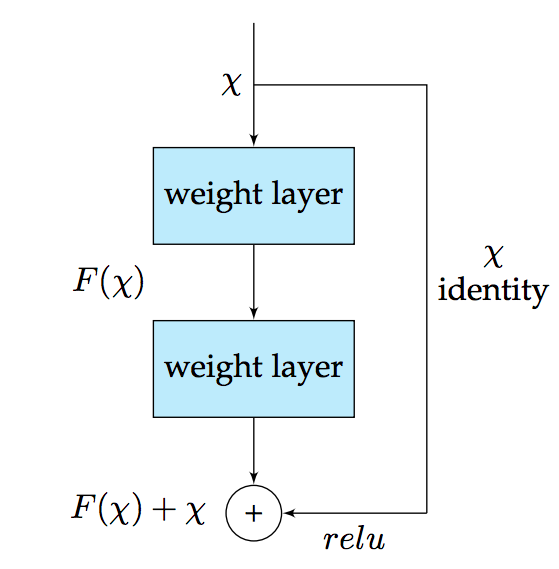


图1.6

这些网络结构都可以用来提取图像的高维特征，为了能够进行语义分割，我们需要实现像素点级别的预测，即对于输入图像中的每一个像素点，都需要判断它属于的类别。为了达到这一目的，引入了反卷积(deconvolution)的概念，如图1.7所示，反卷积可以理解为卷积操作的反向过程，将提取出的特征空间映射到与原图像大小相等的空间上。利用反卷积可以很容易地得到一个与原图像同样大小的对等映射，实现对应像素点的类别预测。



图1.7

**语义分割数据集**

下面介绍一些近年来公开的语义分割的数据集。

PASCAL Visual Object Classes ( VOC ) [6] 总共有21种类别，包含交通工具，房屋，动物，人等等，训练图片有1464张，验证集有1449张。

Microsoft Common Objects in Context ( COCO ) [7] 总共有80多种类别，82783张训练图片，40504张验证图片，80000张测试图片，该数据集是很多模型性能的测试用例。

Cityscapes [8] 是城市街道的图像数据集，主要用于自动驾驶的物体识别，总共有8大类和30小类，数据集包含5000张标注完整的图片和20000张粗糙标注的图片，该数据集是从50个城市的街景收集的。

1.2.2 肖像抠图

肖像抠图也属于语义分割，是一种二分类任务，将人物肖像和背景区分开。之前提到的数据集中也包含人物这一类别，但是图片数量有限，而且人物通常不是图片的主体部分，对人物的标注也很粗糙，对于衣服，头发等边缘部位没有很好的标注。在人物肖像抠图这一工作中，我们需要的是大量的人物正面照片，对细节部分要有精确的标注。Xiaoyong Shen 收集并标注了大量的人物肖像，如图1.8所示。他从Flickr上抓取了1800张人物的正面肖像图，用Photoshop进行了精确标注。数据集中包含了多种年龄，肤色，穿着，发型的人物，而且背景也具有多样性。本次实验我将采用该数据集用来训练和测试评估。



图1.8

# 流程和方法

2.1 全卷积神经网络

本次实验我采用的基本模型是基于Jonathan Long [9] 提出的全卷积神经网络 ( Fully Convolutional Networks)。

全卷积神经网络在分类网络(VggNet, GoogleNet)的基础上，将最后的全连接层替换为全卷积层，然后进行反卷积将特征映射为原图像大小。全卷积神经网络计算损失函数是根据网络输出的像素标签和真实标注值的差别，目标损失函数可以写为：

其中p是图像的下标，Xt(p)是以t为参数，神经网络计算出像素点p的标签值，l(p)是像素点p的真实值，损失函数e计算神经网络输出值和真实值之间的误差，loss为所有像素点的误差总和。

与图片分类的深度网络相似，全卷积神经网络也包含如下几层：

卷积层：利用卷积核将上一层的图片传至下一层，可以理解为每一层卷积都是在上一层的基础上提取更高维的特征。卷积核被训练成为有能力提取图片中的特征，如边角信息，或者更高层的语义信息。

激活函数：全卷积神经网络采用的是ReLu激活函数，函数为f(x) = max(0, x)。激活函数可以帮助网络学习非线性特征，同时ReLu可以有效地缓解梯度消失。

池化层：池化层有最大池化和平均池化，一般采取最大池化方法，即只保留一定范围内的最大值。池化可以有效减小网络规模，提取最重要特征的同时减小了过拟合风险。

反卷积层：反卷积核将上层特征映射到与输入图片同等大小的空间，反卷积核的参数也是直接利用神经网络训练得到，也有人尝试使用固定的反卷积核参数，但是在本实验中，让神经网络自主学习该参数表现更好。

损失函数层：这一层用来计算网络预测值和真实值的差异，在语义分割任务中，利用softmax函数和交叉熵来计算损失函数。

我采用随机梯度下降的方式来进行误差逆传播。

如图2.1所示，全卷积神经网络采用了一种跳跃结构 (skip structure) ，将粗糙的，高维的语义特征和精确的，低维的特征相结合，从而得到更好的预测效果。FCN-32s直接根据最后的池化层的特征进行反卷积，将池化层的图像放大了32倍，这样可能丢失很多低维的细节信息。FCN-16s则加入了上一层的池化层特征，将上一层的池化层放大16倍，与FCN-32s的图像取平均值，这样就可以保留部分细节信息。FCN-8s加入了更上一层的特征，将池化层放大8倍然后与之前的图像取均值。本次实验尝试过FCN-32s, FCN-16s, FCN-8s，FCN-8s能够取得更高的准确率。

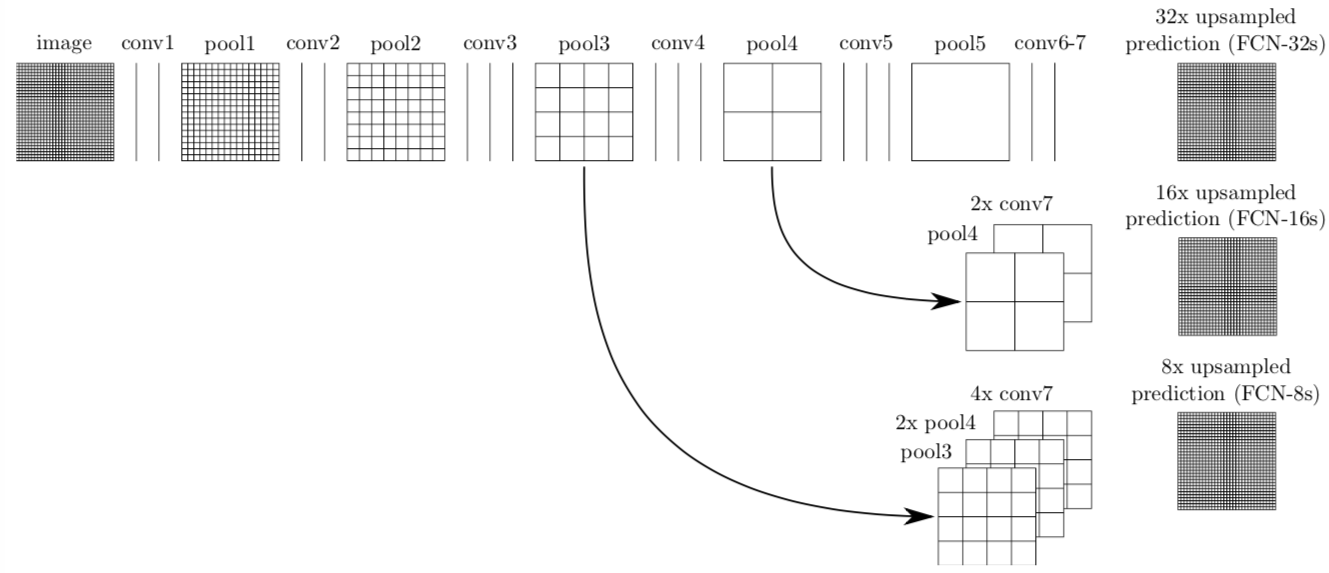


图2.1

2.2 加入Resnet结构

2.3 边缘模糊化处理

2.4 数据增强

由于深度学习在数据更多的情况下能够获得更好的性能，但是能够获取的数据有限，而且标注新的数据极为费时，所以采用数据增强的方式来增大数据容量，从而使得深度网络模型更具有鲁棒性。常见的数据增强方法有修改图片尺寸，按比例缩放，加噪，反转，旋转，对比度变换等。根据人物肖像的特点，我采取了旋转，伽马对比度变换，替换背景三种数据增强的方法。旋转角度分别为 -30，-15，15，30，伽马变换值分别为0.3，0.5，1.2，1.5，背景替换则随机从数据集中选取5张其他背景进行替换。经过数据增强，可以得到超过原来数据集十倍的数据。

# 实验结果

3.1 性能指标IOU

IOU为intersection over union，是度量语义分割任务的常见指标。假设在肖像分割任务中，人物肖像为正例，背景为负例，则IOU的计算方式如下:

IOU = (true positive) / (true positive + false positive +false negative)

即IOU的值等于真正例除以真正例，假正例，假反例的和。

由于IOU对于每一个像素点的权重都是相同的，即不管该像素点的位置，对其正确或错误的标注对IOU的影响是相同的。但是在人物肖像分割的任务中，不同位置的像素点有不同的重要性。例如对于背景中远离人物的像素点标注错误并不会影响分割效果，但是对于肖像和背景边缘的像素点，正确的分割就尤为重要。所以我采用另外一种性能指标，带权重的IOU，weight IOU，它赋予边缘区域更大的权重。在实验中，weight IOU在边缘10个像素点内赋予其他区域二倍的权重，这样对于边缘更敏感的模型可以获得更高的性能评分。

3.2 实验对比

# 分布式存储和训练加速

4.1 Tensorflow加速训练和预测

4.2 Alluxio统一接口

由于深度学习需要巨大容量的数据，这些数据很可能存放在不同的底层存储系统中，例如部分存放在

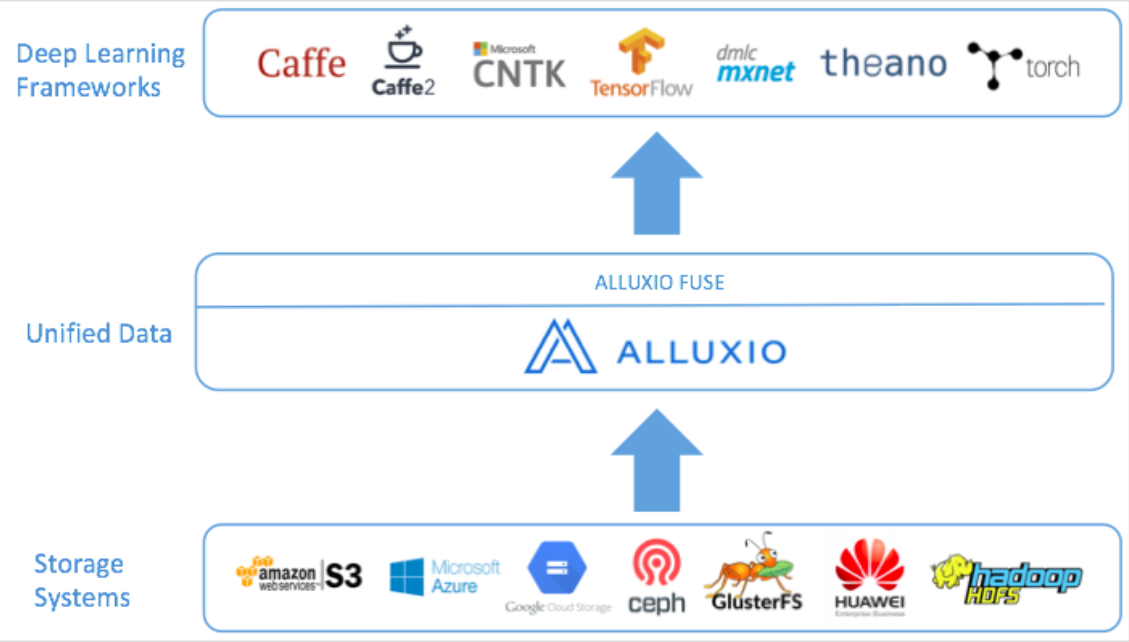


图4.1

# 结论和后续工作

5.1 实验中的不足

5.2 后续的工作

参考文献

[1] Shen, X. , Hertzmann, A. , Jia, J. , Paris, S. , Price, B. , Shechtman, E. and Sachs, I. (2016), Automatic Portrait Segmentation for Image Stylization. Computer Graphics Forum, 35: 93-102.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385

[6] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” International Journal of Computer Vision, vol. 111, no. 1, pp. 98–136, Jan. 2015

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in European Conference on Computer Vision. Springer, 2014, pp. 740–755.

[8] M. Cordts, M. Omran, S. Ramos, T. Scharwachter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset,” in CVPR Workshop on The Future of Datasets in Vision, 2015.

[9] Jonathan Long, Evan Shelhamer, Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

[] Shen X., Tao X., Gao H., Zhou C., Jia J. (2016) Deep Automatic Portrait Matting. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham

[] Xianzhi Du, Xiaolong Wang, Dawei Li, Jingwen Zhu, Serafettin Tasci, Cameron Upright, Stephen Walsh, Larry Davis. Boundary-sensitive Network for Portrait Segmentation. arXiv:1712.08675

[] A. Garcia-Garcia, S. Orts-Escolano, S.O. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez.

A Review on Deep Learning Techniques Applied to Semantic Segmentation. arXiv:1704.06857