广义Dice重叠作为高度不平衡分割的深度学习损失函数

Carole H. Sudre[1,2], Wenqi Li[1], Tom Vercauteren[1], Sebastien Ourselin[1,2], and M. Jorge Cardoso1、2

[1]英国伦敦大学学院CMIC平移成像组，NW1 2HE
[2]伦敦大学学院神经病学研究所痴呆研究中心，英国伦敦WC1N 3BG

摘要近年来，深度学习被证明是图像分析的有力工具，现在被广泛用于分割2D和3D医学图像。深度学习分割框架不仅依赖于网络架构的选择，还依赖于损失函数的选择。当分割过程针对罕见的观测值时，候选标签之间很可能发生严重的类别不平衡，从而导致性能次优。为了缓解这一问题，人们提出了加权交叉熵函数、灵敏度函数或Dice损失函数等策略。在这项工作中，我们研究了这些损失函数的行为，以及它们在2D和3D分割任务中存在不同标签不平衡率时对学习率调整的敏感性。我们还建议使用广义骰子重叠的类重新平衡特性(分割评估的已知度量)作为不平衡任务的鲁棒和准确的深度学习损失函数。

1 介绍

在医学图像分析中，一个常见的任务是检测、分割和表征病理区域的能力，这些区域只代表全图像中非常小的一部分。例如，多发性硬化症或老龄化人群中的脑瘤或白质病变就是这样。众所周知，这样的不平衡问题会导致建立良好的、生成的和区分的分割框架的不稳定。深度学习框架已经成功地应用于2D生物数据的分割，最近被扩展到3D问题[10]。近年来出现了处理类别不平衡的多策略设计(例如特定器官、病理……)。在这些策略中，一些将努力重点放在减少不平衡上，通过选择正在分析的训练样本，冒着降低训练中的变异性的风险[3,5]，而另一些则推导出了更合适和鲁棒的损失函数[1,8,9]。在这项工作中，我们研究了之前发布的三个损失函数在2D和3D中不同的多类分割问题中的训练行为，同时评估了它们对学习率和样本率的鲁棒性。我们还建议使用广义骰子重叠的类再平衡特性作为平衡和非平衡数据的新损失函数。

2 方法

2.1 不平衡数据的损失函数

本工作中比较的损失函数之所以被选中，是因为它们具有解决类别不平衡问题的潜力。所有损失函数都在二元分类(前景vs.背景)公式下进行了分析，因为它代表了允许对类别不平衡进行量化的最简单的设置。请注意，将其中一些损失函数表述为1类问题将在一定程度上缓解不平衡问题，但结果不会轻易推广到多个类。设R为具有体素值r_n的参考前景分割(金标准)，P为N个图像元素上前景标签的预测概率图p_n，背景类概率为1−P。

加权交叉熵(Weighted cross-entropy, WCE):加权交叉熵在文献[9]中已被广泛使用。WCE的两类形式可表示为

$$\mathrm{WCE} = -\frac{1}{N}\sum_{n=1}^{N} w r_n \log(p_n) + (1 - r_n)\log(1 - p_n),$$

其中w是归属于前景类的权重，这里定义为$w = \frac{N - P_n}{P_n}$。加权交叉熵可以简单地扩展到不止两类。

Dice loss (DL) Dice得分系数(DSC)是一种广泛用于评估分割性能的重叠度量，当金标准或ground truth可用时。Milletari等人提出的[8]作为损失函数，骰子损失的2类变体，记为DL_2，可以表示为

$$\mathrm{DL}_2 = 1 - \frac{\sum_{n=1}^{N} p_n r_n + \epsilon}{\sum_{n=1}^{N} p_n + r_n + \epsilon} - \frac{\sum_{n=1}^{N}(1 - p_n)(1 - r_n) + \epsilon}{\sum_{n=1}^{N} 2 - p_n - r_n + \epsilon}$$

这里使用术语是为了通过避免除以0的数值问题来保证损失函数的稳定性，即R和P为空。

敏感性-特异性(SS):在评估分割结果时，敏感性和特异性是两个被高度重视的特征。Brosch等人描述了将这些评估转化为损失函数的过程。[1]

$$\mathrm{SS} = \lambda \frac{\sum_{n=1}^{N}(r_n - p_n)^2 r_n}{\sum_{n=1}^{N} r_n + \epsilon} + (1 - \lambda)\frac{\sum_{n=1}^{N}(r_n - p_n)^2(1 - r_n)}{\sum_{n=1}^{N}(1 - r_n) + \epsilon}.$$

参数λ(在敏感性和特异性之间的平衡中加权)被设置为0.05，如[1]所建议的那样。当其中一个集合为空时，需要再次使用术语来处理除0的情况。

广义骰子损失(Generalized Dice Loss, GDL): Crum等[2]提出了广义骰子分数(Generalized Dice Score, GDS)作为一种用单个分数评估多个类分割的方法，但尚未在判别模型训练中使用。我们建议使用GDL作为训练深度卷积神经网络的损失函数。它采取的形式是:

$$\mathrm{GDL} = 1 - 2\frac{\sum_{l=1}^{2} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{2} w_l \sum_n r_{ln} + p_{ln}},$$

其中wl用于为不同的label$_p$set属性提供不变性。在下文中，我们采用wl =1 /(Nn=1 rln) 2时的符号GDLv。如上所述

在[2]中，当选择GDL$_v$加权时，每个标签的贡献被其体积的倒数修正，从而降低了众所周知的区域大小和Dice分数之间的相关性。在随机梯度下降训练方面，在二分类问题中，关于p_i的梯度为:

$$\frac{\partial \mathrm{GDL}}{\partial p_i} = -2\frac{(w_1^2 - w_2^2)\left[\sum_{n=1}^{N} p_n r_n - r_i \sum_{n=1}^{N}(p_n + r_n)\right] + N w_2 (w_1 + w_2)(1 - 2r_i)}{\left[(w_1 - w_2)\sum_{n=1}^{N}(p_n + r_n) + 2N w_2\right]^2}$$

注意，这个梯度可以轻松地扩展到两个以上的类别。

2.2 深度学习框架

为了广泛研究不同网络架构中的损失函数，由于其最先进的性能，四个先前发布的网络被选择为用于分割的代表性网络，并使用Tensorflow重新实现。

2D网络:使用两个为2D图像设计的网络来评估损失函数的行为:UNet[9]和twpathcnn[3]。UNet架构呈现u形模式，其中降压是一系列两个卷积，然后是下采样层，而升压由一系列两个卷积和上采样组成。每个尺度下的下采样和上采样路径之间都建立了连接。为肿瘤分割而设计的twpathcnn[3]，在全卷积二维设置中使用，通常假设在切片厚度较大的情况下，三维分割问题可以用二维网络近似。这个网络涉及两个网络的并行训练——一个局部子网络和一个全局子网络。前者由两个卷积层组成，核大小为7-^{2}和5-^{2}，max-out正则化分别与大小为4-^{2}和2-^{2}的max-pooling层交织;而后者网络由一个核大小为13-^{2}的卷积层和一个大小为2-^{2}的最大池化层组成。然后，在最终的全连接层之前，将局部网络和全局网络的特征串联起来，从而对输入图像的中心位置进行分类。

3D网络:在3D环境中使用DeepMedic架构[4]和HighResNet网络[6]。DeepMedic由两个网络并行训练组成:一个网络考虑图像的全分辨率，另一个网络考虑图像的下采样版本。在应用两个全连接层从而产生最终分割之前，将得到的特征进行串联。HighResNet是一个紧凑的端到端网络，通过一组连续的卷积块和残余连接将图像体映射到体素分割。为了在多个尺度上合并图像特征，卷积核被膨胀为2倍或4倍。输入体积的空间分辨率在整个网络中保持不变。

3 实验与结果

3.1 实验

我们选择的两个分割任务突出损失函数对脑病理的影响:第一个任务处理肿瘤分割，肿瘤位置往往未知，大小差异很大，第二个任务包括与年龄相关的白质高强度分割，病变可以呈现各种形状、位置和大小。

为了评估每个损失函数的训练行为，对两个网络分别测试了不同的样本和学习率。学习率(LR)被选择为对数间隔，并设置为10^{-3}，10^{-4}和10^{-5}。对于每个网络，使用三种patch大小(small:S, moderate:M, large:L)，根据网络的设计产生不同的有效视场来训练模型。根据patch大小使用不同的批处理大小。每个网络的初始和有效补丁大小、批量大小和由此产生的不平衡被收集在表1中。为了确保所有损失函数的合理行为，如果训练块包含至少一个前景元素，则选择训练块。更大的patch大小通常代表更不平衡的训练集。这些网络在没有训练数据增强的情况下进行训练，以确保训练行为之间具有更多的可比性。补丁中的不平衡根据网络和上下文的差异很大，在最坏情况下达到3D补丁的中位数0.2%

2D网络被应用于BRATS[7]，这是一个神经肿瘤数据集，其中分割任务在这里定位图像中的背景(健康组织)和前景(病理组织，这里是肿瘤)。

表1。四个网络的patch大小和采样率的比较。

	乌内特			TwoPathCNN			DeepMedic			HighResNet		
批量大小	5	3	1	5	3	1	5	3	1	5	3	1
Initial Patch Size有效补丁大小	56 16	64 24	88 48	51 19	63 31	85 53	51 3	63 15	87 39	51 15	63 27	85 49
不平衡率	0.52	0.33	0.15	0.29	0.25	0.16	0.20	0.01	0.002	0.02	OOI	0.003

表2。UNet和TwoPathCNN在2D环境下200次DSC迭代的比较。结果采用中位数格式(四分位数范围)。

		乌内特				TwoPathCNN			
补丁	英商道	WCE	戴斯。莱纳姆:	党卫军	GDL,	WCE	DL2	党卫军	GDL,
	-5	0.71 (0.17)	073 (0.13)	0.37 (0.17)	0.75 (0.14)	0.56 (0.48)	9	0.53 (0.41)	0.49 (0.44)
	-4	0.77 (0.18)	0.76 (0.13)	0.74 (0.16)	0.80 (0.12)	0.80 (0.12)	0.79 (0.11)	0.81 (0.12)	0.80(0.12)
	-3	0.70 (0.17)	0.72 (0.15)	0.39 (0.16)	0.72 (0.15)	0(0)	0 (0)	0.77 (0.11)	0.72 (0.15)
	-5	0.71 (0.23)	0.70 (0.22)	0.65 (0.25)	0740.19)	0(0)	0.73 (0.18)	0.69 (0.21)	0.73 (0.19)
米	4	0.73 (0.18)	0.70 (0.22)	0.61 (0.25)	0.72 (0.19)	0.77 (0.16)	0.76 (0.17)	0.71 (0.18)	0.76 (0.17)
	-3	0.68 (0.23)	0.67 (0.21)	0.70 (0.26)	0.69 (0.22)	0(0)	0.71 (0.22)	0.67 (0.21)	0.72 (0.19)
	-5	0.63 (0.46)	0.62 (0.40)	0.49 (0.42)	0.56 (0.44)	0.62 (0.50)	0.50 (0.41)	0.50 (0.38)	0.56 (0.35)
l	-4	0.68 (0.34)	0.64 (0.44)	0.18 (0.24)	0.66 (0.39)	0.64 (0.42)	(03)9v	0.52 (0.38)	0.64 (0.35)
	-3	0.59 (0.39)	0.57 (0.53)	0.16 (0.22)	0.59 (0.45)	0.77 (0.12)	0.77 (0.14)	0.79 (0.12)	0.79 (0.11)

3D网络被应用于524名受试者呈现与年龄相关的白质高强度的内部数据集。在这两种情况下，t1加权、t2加权和FLAIR数据都通过根据WM强度分布对数据进行z-评分进行强度归一化。训练在1000后被任意停止(resp.;3000)次迭代为2D (resp.;3D)实验，因为它被发现足以允许所有指标的收敛。

3.2 2 d的结果

表2给出了四种损失函数在不同学习率和不同网络下最后200步训练的DSC统计数据，图1给出了在学习率和有效补丁大小空间中对应的等值线，显著说明了GDL对超参数空间的鲁棒性。不同损失函数之间观察到的主要差异是对学习率的鲁棒性，当使用TwoPathCNN时，WCE和DL_2不太能够应对快速学习率(10^{-3})

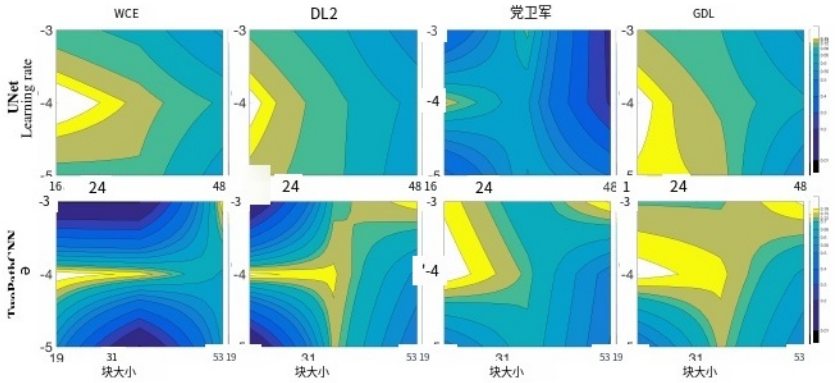

图1所示。在2D环境中，不同有效补丁大小和学习率条件下的DSC(过去200次迭代的中位数)损失函数行为。为了可视化的目的，等值线被线性插值。

表3。DeepMedic和HighResNet在3D环境下200次DSC迭代的比较。结果在中位数格式下(四分位数范围)。

补丁	英商道	WCE	DL2	DeepMedic 党卫军	GDL	WCE	DL2	HighResNet党卫军	GDL
	-5	0.49 (0.17)	0.44 (0.19)	0.42 (0.14)	0.46 (0.17)	0(0)		0(0)	0.06 (0.15)
年代	-4	0.58 (0.20)	0.60 (0.15)	0.61 (0.22)	0.61 (0.18)	0 (0)		0.71 (0.18)	0.34 (0.20)
	-3	0.61 (0.12)	0.59 (0.14)	0.63 (0.15)	0.60 (0.15)	0 (0)		0 (0)	0(0)
	-5	0.05 (0.07)	0.05 (0.07)	0.05 (0.06)	0.04 (0.06)	0 (0)		0.60 (0.27)	0.15 (0.13)
米	-4	0.09 (0.11)	0.07 (0.09)	0.08 (0.09)	0.08 (0.10)	0(0)		0.71 (0.20)	0.20 (0.20)
	-3	0.45 (0.31)	0.42 (0.31)	0.17 (0.24)	0.48 (0.32)	0 (0)		0(0)	0(0)
	-5	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)	0 (0)		0.54 (0.27)	0.03 (0.06)
l	-4	0.01 (0.04)	0.02 (0.04)	0.02 (0.04)	0.01 (0.04)	0(0)		0.57 (0.32)	0.08 (0.19)
	-3	0.21 (0.33)	0.18 (0.30)	0.05 (0.12)	0.20 (0.33)	0 (0)		0.62 (0.18)	0.22 (0.15)

SS的效率更依赖于网络。一个中间的学习率(10−4)似乎在所有情况下都能带来最好的训练。在不同的采样策略中，性能的模式在不同的损失函数中是相似的，在使用较小的patch但较大的批量大小时具有更强的性能。

3.3 3 d结果

与上一节类似，表3展示了3D实验中最近200次迭代的损失函数、样本量和学习率的统计数据，而图2使用等值线绘制了损失函数对参数空间的鲁棒性表示。它对超参数的强依赖性使得DeepMedic对损失函数的选择不可知。在数据不平衡程度较高的3D环境下，与GDL_v相比，WCE无法训练，SS的性能明显下降。DL_2在低学习率下的表现与GLD_v类似，但在更高的训练率下却无法训练。在不同的学习率下观察到与2D情况类似的模式，学习率为

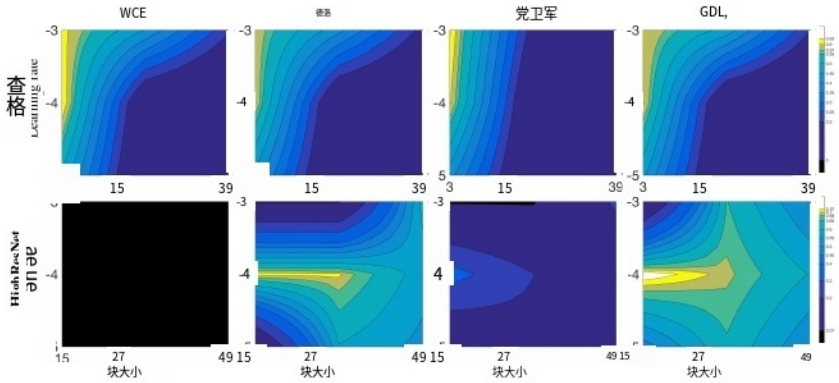

图2所示。基于DSC(过去200次迭代的中位数)的损失函数行为在三维环境中不同条件下的有效补丁大小和学习率。为了可视化的目的，等值线被线性插值。

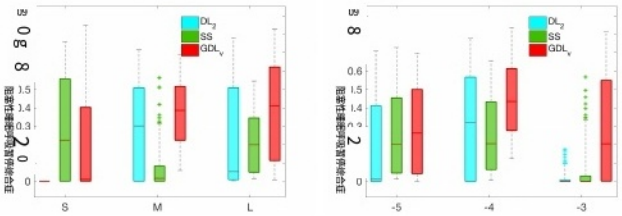

图3所示。测试所有损失函数在补丁大小(左)和学习率(右)上的DSC集。WCE被忽略了，因为它无法应对不平衡。

10^{-5}在3000次迭代后未能在损失函数中提供平台。我们还观察到，对于较小的补丁大小，学习率对网络性能的影响更大，但在适当的条件下(LR=10^{-4})，较小的补丁(和较大的批量大小)导致更高的整体性能。

3D测试集对于3D实验，10%的可用数据被保留用于测试目的。使用最终的HighResNet模型来推断测试数据的分割。图3显示了不同采样策略(右)和不同学习率(左)的损失函数在DSC中的比较。总的来说，GDL_v在不同的实验中被发现比其他损失函数更鲁棒，对于不那么不平衡的样本，相对性能的变化很小。图4是使用HighResNet在10−4的学习率下使用最大的patch尺寸进行三维实验得到的分割示例。

4 讨论

从四个损失函数在两个不同任务/网络中跨学习率和采样策略的训练行为观察来看，似乎大多数为不平衡数据集设计的损失策略都很好地处理了轻微的不平衡。然而，当不平衡程度增加时，基于重叠度量的损失函数显得更鲁棒。可靠性最强的跨

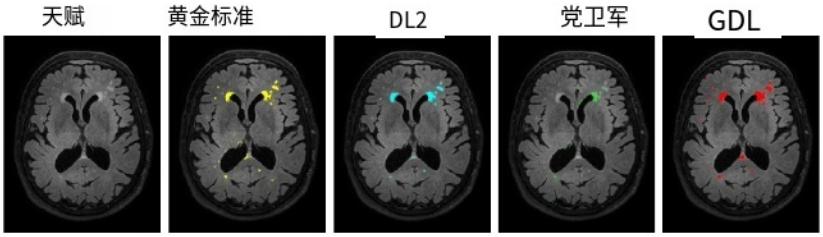

图4所示。使用不同损失函数对随机选择的3D测试集进行分割。注意使用GDL_v时捕获标点病变的能力增加了。损失函数在学习率10^{-4}下每步使用大小为85-³的单个patch进行训练。

使用GDL_v时观察了设置。总的来说，这项工作证明了在深度学习框架中损失函数的选择是多么关键，特别是在处理高度不平衡的问题时。在这项研究中最不平衡的情况下，3D实验的前景背景比率为0.02%(白质病变)。未来的工作将集中在更极端的不平衡情况，例如在检测空洞和血管周间隙的情况下观察到的情况(1/100000)，在这种情况下，深度学习框架必须在学习所有类别的内在解剖变异性和类别不平衡的可容忍水平之间找到平衡。研究的损失函数是作为开源NiftyNet包(http://www.niftynet.io)的一部分实现的。

致谢这项工作使用了Emerald，一种GPU加速的HPC，由科学与工程南方联盟提供，与STFC卢瑟福-阿普尔顿实验室共同拥有所有权。本研究由EPSRC (EP/h04610 /1, EP/J020990/1, EP/K005278, EP/ h04610 /1)，MRC (MR/J01107X/1)，EU-FP7 项目 VPH-DARE@ IT (FP7-ICT-2011-9-601055)，Wellcome Trust (WT101957)，UCL NIHR生物医学研究中心(痴呆症)和NIHR伦敦大学学院医院BRC (NIHR BRC UCLH/UCL高影响倡议- BW.mn.BRC10269)资助。

参考文献

1. Brosch, T., Yoo, Y., Tang, L.Y., Li, D.K., Traboulsee, A., Tam, R.: Deep convo-lutional encoder networks for multiple sclerosis lesion segmentation. In: MICCAI 2015. pp. 3−11. Springer (2015)

2. Crum, W., Camara, O., Hill, D.: Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis. IEEE TMI 25(11), 1451−1461 (nov 2006)

3. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural net-works. MIA 35, 18−31 (2017)

4. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. MIA 36, 61−78 (feb 2017)

5. Lai, M.: Deep learning for medical image segmentation. arXiv:1505.02000 (2015)

6. Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T.: On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task. In: IPMI 2017

7. Menze, B.H.e.a.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE TMI 34(10), 1993−2024 (oct 2015)

8. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully Convolutional Neural Net-works for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565−571. IEEE (oct 2016)

9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomed-ical image segmentation. In: MICCAI. pp. 234−241. Springer (2015)

10. Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D.: 3d deep learning for efficient and robust landmark detection in volumetric data. In: MICCAI 2015. pp. 565−572. Springer (2015)