Deep Hourglass for Brain Tumor Segmentation

Eze Benson[1(✉)], Michael P. Pound[1], Andrew P. French[1,2], Aaron S. Jackson[1], and Tony P. Pridmore[1]

[1] School of Computer Science, University of Nottingham, Nottingham, UK
{ezenwoko.benson, michael.pound, andrew.p.french, aaron.jackson, tony.pridmore}@nottingham.ac.uk
[2] School of Biosciences, University of Nottingham, Nottingham, UK

Abstract. The segmentation of a brain tumour in an MRI scan is a challenging task, in this paper we present our results for this problem via the BraTS 2018 challenge, consisting of 210 high grade glioma (HGG) and 75 low grade glioma (LGG) volumes for training. We train and evaluate a convolutional neural network (CNN) encoder-decoder network based on a singular hourglass structure. The hourglass network is able to classify the whole tumour (WT), enhancing (ET) tumour and core tumour (TC) in one pass. We apply a small amount of preprocessing to the data before feeding it to the network but no post processing. We apply our method to two different unseen sets of volumes containing 66 and 191 volumes. We achieve an overall Dice coefficient of 92% on the training set. On the first unseen set our network achieves Dice coefficients of 0.66, 0.82 and 0.72 for ET, WT and TC. On the second unseen set our network achieves Dice coefficients of 0.62, 0.79 and 0.65 on ET, WT and TC.

Keywords: Convolutional neural network · Deep learning · Hourglass · Glioma

1 Introduction

Identifying regions of the brain which are tumourous is a task often carried out by medical professionals. Manually classifying segments of the tumour is a subset of a group of problems commonly referred to as semantic segmentation. Semantic segmentation is the task of assigning a class to each pixel within an image, modern automated solutions to this problem often use convolutional neural networks (CNN). The introduction of fully convolutional networks (FCN) [1] established a convolutional neural network architecture that is widely used for the task of semantic segmentation. Architectures such as U-NET [2] achieved success in biomedical imaging by adopting a similar architecture.

We propose the use of an adapted hourglass [3] network to solve the problem of tumour segmentation. The hourglass network improves on U-NET by using bottleneck blocks and adding convolutions to the skip connections. Training a CNN for this problem is a natural choice as they have demonstrated state-of-the-art performance on

A. Crimi et al. (Eds.): BrainLes 2018, LNCS 11384, pp. 419–428, 2019.
https://doi.org/10.1007/978-3-030-11726-9_37

semantic segmentation problems such as the widely used Pascal VOC2012 [9] and cityscapes [10] datasets.

2 Methods

2.1 Data

The dataset of BraTS 2018 [4–8] provides defined training and validation sets. The training set is composed of 210 MRI scans of high grade gliomas (HGG) and 75 MRI scans of low grade gliomas (LGG). Whilst the validation set is a group of 66 mixed HGG and LGG tumours. The MRIs are volumes in the format given by Eq. (1), in that format they have the dimensions $240 \times 240 \times 155$. Each volume has four corresponding modalities FLAIR T1, T2 and T1CE.

$$\mathrm{X} \times \mathrm{Y} \times \mathrm{Z} \tag{1}$$

Where x is the delineation between dimensions and X, Y and Z are the dimensions on a 3D coordinate system.

2.2 PreProcessing

A high variance in intensity in both validation and training set was observed this lead us normalise the training set to be centred around zero with a standard deviation of one. By normalizing the data, we found that the required training time was reduced and the accuracy of the network was increased. The formula for normalization is given in Eq. (2). Each modality was normalized separately due to the variance in intensity profile between modalities.

$$Z = \frac{x - \mu}{\sigma} \tag{2}$$

Where x is the current intensity, μ is the mean of the modality and σ is the standard deviation of the modality.

2.3 Hourglass Architecture

Our approach is to handle 2D slices of each volume separately, a 2D semantic segmentation problem. We performed additional experimentation using a volumetric encoder-decoder but found that the benefit of an end-to-end volumetric approach was outweighed by the significant necessary drop in features at each layer due to memory restrictions.

We design our network using an encoder-decoder structure, adapted from an hourglass network, popularized in the domain of human-pose estimation [3] The structure of the hourglass is similar to other encoder-decoder networks, but contains a denser use of residual blocks throughout.

The encoder starts with an input of 4 channels and contains 7 residual bottleneck blocks [14], after each a max-pooling layer performs spatial downsampling. A further three residual blocks at the lowest spatial resolution derive higher-level features before a series of bilinear upsampling operations return the network to the original spatial resolution. As in the encoder, all upsampling operations of the decoder are interleaved with residual blocks. Skip layers are added between each matching resolution of the encoder and decoder, with each containing an additional residual block to learn an appropriate mapping.

The first residual block has 64 filters, the filter amount doubles after each pooling operation up to a maximum of 512 filters in a convolution.

In order to improve the network's results for the final test set we made architectural changes to improve accuracy whilst keeping memory consumption to a minimum. We found that the choice of upsampling layer (e.g. bilinear, max-unpooling [11]) made little difference to the performance of the network. Unlike the original work [3] we chose not to stack hourglass networks sequentially and perform intermediate supervision, we found this too had a negligible effect on performance. The number of spatial-downsampling layers, 7 in total, were originally chosen based on the input resolution. However, through experimentation we found that using 5 downsampling layers was optimal and save memory. Only one residual block is used at each depth because adding two at all depths immediately doubles memory consumption which surpasses current memory constraints. We also found that replacing elementwise summation with concatenation followed by a 1×1 convolution improved results noticeably. Despite the additional memory consumption of the concatenation and convolutional layer, the increase in performance boost makes the change worthwhile.

2.4 Training

The training was split into two phases pre and post true validation set release. In the first phase the dataset was split into a test set, validation set and training set where each set was 10%, 10% and 80% of the original training set respectively. The data provided is treated as though it is the entire dataset so that our training can be validated and tested in preparation for the true validation set. This allows the network to avoid overfitting and approximate the results expected on the release of the second dataset. Later the network is retrained using a 10% test set and 90% training set split in order to obtain test results on the original data whilst maximizing the training set size. The network is trained for the same number of epochs for all training. The second phase is conducted post true validation set release. In this phase the BraTS dataset is split into 10% validation and 90% training.

The network is trained using an identical training scheme for both the natural and augmented dataset.

The hourglass network implemented in this paper only uses spatial convolutions, to accommodate this we convert MR volumes into a set of 155 images of spatial resolution 240^2. To do this we separate the volume along the depth dimension. For convenience we pad the images to the new resolution 256^2, this allows us to perform pooling operations where the output resolution of a feature map is always 2^x. In turn this allows us to perform concatenations or elementwise summations in the decoder

network without a resolution difference between two feature maps. The 285 volumes therefore become a dataset of 44175 images. All four modalities are used for training and are given to the network as 4 input channels of a single image.

The hourglass network was chosen because it has been successful in other tasks such as human pose estimation [3] and allows the stacking of the network. Stacking the network multiple times sequentially can give performance boosts as shown before [3]. Spatial convolutions were chosen instead of volumetric convolutions because they consume much less memory, volumetric convolutions would exceed available memory if a stacked network was used. In addition, volumetric convolutions are so memory intensive that they do not allow a network to be trained on the entire MR volume at the same time as having a rich set of filters in a deep network. An alternative to this volumetric network is a volumetric network with a subset of a volume included E.g. A $32 \times 32 \times 32$ chunk. However, the problems remain largely unsolved, the performance boost given by depth context is potentially outweighed by the larger number of filters available in a spatial network. This multitude of reasons led to the choice of a spatial network which would be deeper and wider than the equivalent volumetric network given the same memory constraints.

The hourglass is trained on a NVIDIA TITAN X GPU using a cross entropy loss function with a learning rate of 10^{-5} which is decreased by a factor of 10 every 30 epochs. A batch size of 8 is used and the network is trained for a total of 50 epochs therefore the learning rate is only adapted once. The adaptive gradient descent algorithm, RMSProp is used to train the network faster than the typical stochastic gradient descent.

2.5 Data Augmentation

Two methods of data augmentation are used in this paper vertical flipping and random intensity variation. Vertical flipping is used because it matches the natural symmetrical shape of the brain.

Random intensity variation is used because the intensity between MRI scans varies significantly. This is shown by the fact that the standard deviation of the FLAIR modality in the dataset is greater than the mean by almost a factor of 10. E.g. The standard deviation and mean for the FLAIR modality are 529.2 and 61.8 respectively. The T1, T1CE and T2 modalities have similar standard deviations. Intensity variation is performed on the normalised dataset by first rescaling the standard deviation of the dataset and then shifting the mean. This allows the dataset to include image intensities which are not present in the original dataset but could appear on an MRI volume. The range for randomly changing the standard deviation is between zero and two. The mean is shifted between values of 0.4 and –0.4. Values above a standard deviation of two were experimented with but lead to a significant decrease in accuracy. Shifting the mean by over 0.5 and under –0.5 were trialed but also caused an accuracy decrease. The network is trained with and without data augmentation to experimentally ascertain whether augmentation gives any performance increase when using this network on the dataset.

3 Results and Discussion

The results are split into three sections, the results on the training data set, the results on the later released validation set and the results on the final test set. Results are shown for networks trained on the standard data and on augmented data in the validation set.

3.1 Training Dataset

We trained the network on 90% of the data leaving 10% for testing purposes. The network achieved a Dice coefficient of 92% with an IOU of 86%. We find that IOU approximates the network's worst performance on the test set in contrast to Dice which gives an approximate representation of the average case.

3.2 Validation Dataset

The results presented in this section are those achieved when segmenting the validation set using the network trained in Sect. 3.1. Table 1 shows the results of the segmentation without augmentation and Table 2 shows the results with flipping and intensity variation. The metrics provided in both tables are the standard metrics output by the BraTS automatic online evaluation server. Some metrics have been omitted to save space, only the most important evaluation metrics have been included.

Table 1. The results of the hourglass network segmenting the unseen validation set without augmentation in the training data.

	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausedorff TC
Mean	0.59	0.82	0.64	18.12	94.28	130.70
Std	0.28	0.12	0.24	26.62	50.15	42.40
Median	0.71	0.86	0.71	5.732	97.13	132.59
25 quantile	0.48	0.78	0.51	3.162	52.72	103.36
75 quantile	0.80	0.90	0.83	20.03	135.81	163.39

Table 2. The results of the hourglass network segmenting unseen validation set where the network has been trained with augmented data

	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausedorff TC
Mean	0.56	0.82	0.61	14.29	13.57	17.95
Std	0.29	0.13	0.22	23.26	15.32	18.14
Median	0.67	0.87	0.67	5.92	6.59	11.18
25 quantile	0.40	0.78	0.50	2.83	4.18	8.30
75 quantile	0.80	0.90	0.79	12.56	14.97	18.79

After comparing the metrics between a dataset with augmentation and one without we find that in this challenge augmentation appears to give a small increase in accuracy for Dice coefficient and improves the Hausdorff accuracies significantly. It is likely the

case that the frequency at which the network misclassifies pixels remains similar but the network's ability to localize the pixels is increased.

Overall the network segments the whole tumour more accurately than it does the core tumour or enhancing tumour, from the results in previous challenges this result is expected. Naturally the enhancing and core tumour are much more difficult to segment due to the similarity between all classes.

Tables 1 and 2 both show a large disparity between the median and mean accuracy especially with results for the enhancing tumour where the difference is around 10%. The difference is caused by the difficulty of detecting the enhancing tumour and core tumour in some volumes. In most volumes the Dice coefficients are well above the mean however some outliers achieve a score of 0 therefore reducing the mean significantly. When removing these cases the mean Dice coefficient increases by 4% showing that the disparity can be explained by a few very difficult volumes. Some examples of the metrics achieved on these volumes are shown in Table 3.

Table 3. Segmentation results for very difficult volumes using a network trained with augmented data

	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausedorff TC
TCIA09_248_1	0	0.79	0.63	0	14.18	10.82
TCIA10_195_1	0	0.80	0.63	0	15.23	25.98
TCIA11_612_1	0	0.74	0.60	0	52.78	48.52
TCIA12_613_1	0	0.69	0.26	0	49.97	9.00
TCIA13_646_1	0	0.90	0.40	0	35.83	6.48

3.3 Test Dataset

Before the release of the final evaluation dataset we train our network using 95% of the training data. The remaining 5% of the training data is used for on the fly validation of the network to monitor training and prevent overfitting. The network architecture has been adapted to improve the results on the validation set, these architectural changes are discussed in Sect. 2.3. We present the new validation set results along with the test set results. Section 3.2 showed that the network has an increase in Hausdorff95 accuracy when data augmentation was used. The network used for the results in this section was trained using data augmentation.

Table 4. The results of the hourglass network segmenting unseen the validation set

	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausdorff TC
Mean	0.66	0.82	0.72	15.94	26.41	18.87
Std	0.27	0.10	0.23	25.56	23.61	20.56
Median	0.79	0.84	0.80	4.69	17.32	12.47
25 quantile	0.56	0.78	0.62	2.45	7.19	6.61
75 quantile	0.84	0.89	0.89	17.60	38.13	19.60

Table 5. The results of the hourglass network segmenting unseen test set

Label	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausdorff TC
Mean	0.62	0.79	0.65	47.48	13.54	31.58
Std	0.32	0.25	0.34	113.76	23.51	83.24
Median	0.77	0.88	0.82	3.00	5.00	6.40
25 quantile	0.47	0.80	0.48	1.73	3.00	3.32
75 quantile	0.85	0.92	0.90	9.84	9.72	14.80

Table 4 shows the results of the hourglass network on the validation set. The dice scores for the validation set increase by 10% for both ET and TC whilst remaining approximately the same for the whole tumour segmentation. Conversely the Hausdorff scores increase (where a higher score is a decrease in performance) by 1, 13 and 2 for ET, WT and TC respectively. The increase in dice score indicates that the total number of pixels that are being classified correctly has increased but the decrease in Hausdorff score shows that the largest error in the shape of the classified pixels is much higher. The qualitative analysis presented in Sect. 3.4 shows that this may be because misclassification of background pixels far away from the site of the tumour.

The median Hausdorff distance and dice score are significantly better than the mean indicating that the mean results are being distorted by a small subset of difficult to segment brain tumour volumes. This is discussed in Sect. 3.2. The std of both metrics is also very high showing that the networks performance varies largely between volumes.

The network shows a significant improvement in the most problematic volumes highlighted in Table 3. Table 6 shows the modified network's performance on the selected examples. The average Hausdorff distance for the selected examples indicates an overall performance decrease however performance on individual volumes varies significantly when dice scores are compared. The network architecture was modified in order to increase performance on the enhanced tumour, Table 6 shows that on 3 out of 5 selected cases there is an increase of between 4.6% and 38% for the enhancing tumour dice score. The variability in dice score amongst the other two metrics indicates that the training scheme has altered the networks ability to classify the tumour in these volumes.

Table 6. The modified network's segmentation results on a subset of problematic volumes

	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausedorff TC
TCIA09_248_1	0.00	0.80	0.48	0.00	61.26	12.41
TCIA10_195_1	0.00	0.86	0.71	0.00	22.67	30.23
TCIA11_612_1	0.38	0.63	0.40	98.47	59.87	98.25
TCIA12_613_1	0.06	0.94	0.94	58.26	4.12	2.83
TCIA13_646_1	0.05	0.70	0.61	111.19	87.68	15.13

The test set results show a decrease in performance on both dice score and hausdorff distance when compared to validation set results. The median scores for both metrics are noticeably better. This indicates that the validation set contains easier to segment volumes but the ratio between difficult and easy volumes is higher. The test set appears to have much more difficult volumes, this is corroborated by the very high standard deviation values. The results suggest that the percentage of easily segmented volumes in the test set is higher than the validation set.

Despite the differences between the network's performance on the validation and test sets both Tables 4 and 5 indicate the same overall strengths and weaknesses of the network as well as the difficulties within the dataset.

3.4 Qualitative Analysis

In this section we present singular slices taken from the network output. The output has 4 classes which are represented by 4 different colours in the segmentation map. Black, yellow, blue and red represent background, whole tumour, core tumour and enhancing tumour.

The network makes many mistakes when segmenting unseen volumes, most often these errors are misclassifying healthy brain tissues as tumourous. Often the mistakes are of a small area which does not affect the dice score significantly but has a noticeable impact on the hausdorff distance. These errors are important and can be improved upon however for brevity this section will focus on the largest errors associated with the problematic volumes highlighted in Sect. 3.3. Figure 1 shows large errors in classification. The largest errors the network makes occur when the input image has large errors of darkness within the tumour caused by necrosis or an irregular tumour shape. It is unclear why this occurs but could be because the training set contains mostly tumour which have small amounts of necrosis which are masses enveloped by the whole tumour. Therefore when given to the network it is unable to deal with the variance.

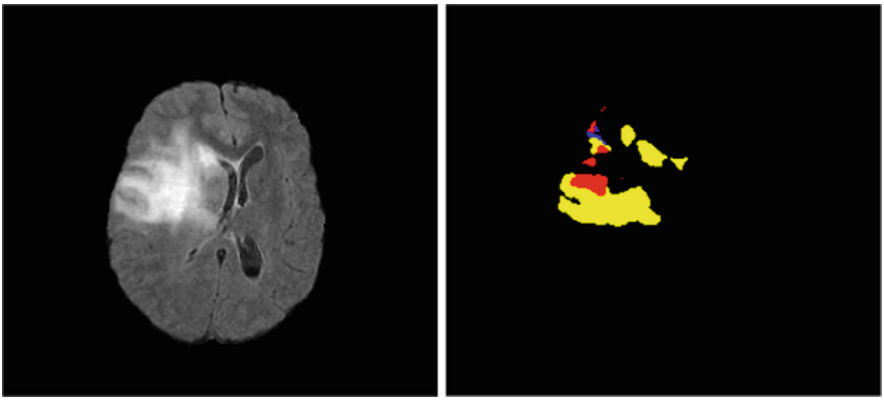

Fig. 1. Left, a FLAIR volume slice containing both brain and tumour tissues. Right, a slice from the network output showing erroneous segmentation results. Two similar looking dark regions on the left side of tumour have been classified different despite having largely the same appearance. These are the most error prone areas for the network.

4 Conclusion

We propose a solution which achieves a 92% Dice coefficient on the training set and 0.66, 0.82 and 0.72 on the validation set. On the test set the network achieves 0.62,0.79 and 0.65 Dice scores. Although the network underperforms on Dice score it can achieve a competitive Hausdorff distance.

Much of the network's underperformance is related to outliers in the set which could be mitigated in future with better preprocessing techniques. Future networks should train more on these difficult volumes using wider public datasets or through synthetic images generated by a CNN. Memory consumption is often a problem when using CNNs, to combat this we plan to add residual blocks in depths which increase the overall accuracy of the network the most. We also plan to add skip connections with an inception block structure [12] as shown in [13] to increase accuracy further.

We show that 2D architectures can segment 3D volumes with success but require fine tuning and a deeper architecture to achieve better results. An approach to bridge the gap may between 2D and 3D may be required. 3D networks outperform 2D networks when depth context is key, how much context is required in most tasks remains unclear. In future works we plan to use a 2.5D approach where each slice has an accompanying adjacent slice either side to provide some depth context.

References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
2. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
3. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
4. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993 (2015)
5. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. data. **5**(4), 170117 (2017)
6. Bakas, et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive, p. 286 (2017)
7. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive, p. 286 (2017)
8. Bakas, S., Reyes, M., Menze, B.: Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. arXiv preprint arXiv:1811.02629 (2018)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)

10. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
11. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561, 2 November 2015
12. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
13. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: The IEEE International Conference on Computer Vision (ICCV), vol. 1, no. 2, p. 4, 1 October 2017
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)