Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation

K. Kamnitsas(✉), W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert, and B. Glocker

Biomedical Image Analysis Group, Imperial College London, London, UK
konstantinos.kamnitsas12@imperial.ac.uk

Abstract. Deep learning approaches such as convolutional neural nets have consistently outperformed previous methods on challenging tasks such as dense, semantic segmentation. However, the various proposed networks perform differently, with behaviour largely influenced by architectural choices and training settings. This paper explores Ensembles of Multiple Models and Architectures (EMMA) for robust performance through aggregation of predictions from a wide range of methods. The approach reduces the influence of the meta-parameters of individual models and the risk of overfitting the configuration to a particular database. EMMA can be seen as an unbiased, generic deep learning model which is shown to yield excellent performance, winning the first position in the BRATS 2017 competition among 50+ participating teams.

1 Introduction

Brain tumours are among the most fatal types of cancer [1]. Out of tumours that originally develop in the brain, gliomas are the most frequent [2]. They arise from glioma cells and, depending on their aggressiveness, they are broadly categorized into high and low grade gliomas [3]. High grade gliomas (HGG) develop rapidly and aggressively, forming abnormal vessels and often a necrotic core, accompanied by surrounding oedema and swelling [2]. They are malignant, with high mortality and average survival rate of less than two years even after treatment [3]. Low grade gliomas (LGG) can be benign or malignant, grow slower, but they may recur and evolve to HGG, thus their treatment is warranted. For treatment, patients undergo radiotherapy, chemotherapy and surgery [1].

Firstly for diagnosis and monitoring the tumour's progression, then for treatment planning and afterwards for assessing the effect of treatment, various neuroimaging protocols are employed. Magnetic resonance imaging (MRI) is widely used in both clinical routine and research studies. It facilitates tumour analysis by allowing estimation of extent, location and investigation of its subcomponents [2]. This however requires accurate delineation of the tumour, which proves challenging due

W. Bai, E. Ferrante, S. McDonagh and M. Sinclair—Equal contribution, in alphabetical order.

A. Crimi et al. (Eds.): BrainLes 2017, LNCS 10670, pp. 450–462, 2018.
https://doi.org/10.1007/978-3-319-75238-9_38

to its complex structure and appearance, the 3D nature of the MR images and the multiple MR sequences that need to be consulted in parallel for informed judgement. These factors make manual delineation time-consuming and subject to inter- and intra-rater variability [4].

Automatic segmentation systems aim at providing an objective and scalable solution. Representative early works are the atlas-based outlier detection method [5] and the joint segmentation-registration framework, often guided by a tumour growth model [6–8]. The past few years saw rapid developments of machine learning methods, with Random Forests being among the most successful [9,10]. More recently, convolutional neural networks (CNN) have gained popularity by exhibiting very promising results for segmentation of brain tumours [11–13].

A variety of CNN architectures have been proposed, each presenting different strengths and weaknesses. Additionally, networks have a vast number of meta parameters. The multiple configuration choices for a system influence not only performance but also its behaviour (Fig. 1). For instance, different models may perform better with different types of pre-processing. Consequently, when investigating their behaviour on a given task, findings can be biased. Finally, a configuration highly optimized on a given database may be an over-fit, and not generalise to other data or tasks.

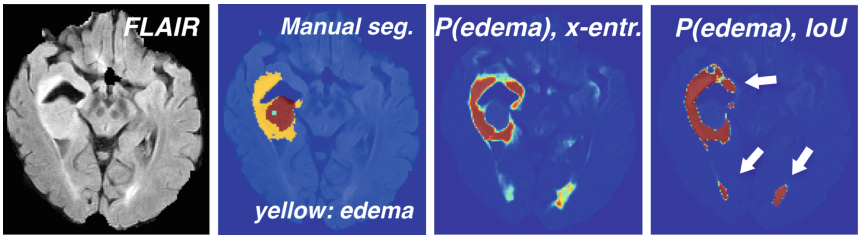

Fig. 1. Left to right: FLAIR; manual annotation of a BRATS'17 subject, where yellow depicts oedema surrounding tumour core; confidence of a CNN predicting oedema, trained with cross-entropy or IoU loss. Although overall performance is similar, training with IoU (or Dice, not shown) loss alters the CNN's behaviour, which tends to output only highly confident predictions, even when false.

In this work we push towards constructing a more *reliable* and *objective* deep learning model. We bring together a variety of CNN architectures, configured and trained in diverse ways in order to introduce high variance between them. By combining them, we construct an *Ensemble of Multiple Models and Architectures* (EMMA), with the aim of *averaging away* the variance and with it model- and configuration-specific behaviours. Our approach leads to: (1) a system robust to unpredictable failures of independent components, (2) enables objective analysis with a generic deep learning model of unbiased behaviour, (3) introduces the new perspective of *ensembling for objectiveness*. This is in contrast to common ensembles, where a single model is trained with small variations such as initial

seeds, which renders the ensemble biased by the main architectural choices. As a first milestone in this endeavour, we evaluated EMMA in the Brain Tumour Segmentation (BRATS) challenge 2017. Our method won the first position in the final testing stage among 50+ competing teams. This indicates the reliability of the approach and paves the way for its use in further analysis.

2 Background: Model Bias, Variance and Ensembling

Feedforward neural networks have been shown capable of approximating any function [14]. They are thus models with zero bias, possible of no systematic error. However they are not a panacea. If left unregularized they can overfit noise in the training data, which leads to mistakes when they are called to generalise. Coupled with the stochasticity of the optimization process and the multiple local minima, this leads to unpredictable inconsistent errors between different instances. This constitutes models with high variance. Regularization reduces the variance but increases the bias, as expressed in the bias/variance dilemma [15]. Regularization can be explicit, such as weight decay that prevents networks from learning rare noisy patterns, or implicit, such as the local connectivity of CNN kernels, which however does not allow the model to learn patterns larger than the its receptive field. Architectural and configuration choices thus introduce bias, altering the behaviour of a network.

One route to address the bias/variance dilemma is ensembling. By combining multiple models, ensembling seeks to create a higher performing model with low variance. The most popular combination rule is averaging, which is not sensitive to inconsistent errors of the singletons [16]. Commonly, instances of a network trained with different initial weights or from multiple final local minima are ensembled, with the majority correcting irregular errors. Intuitively, only inconsistent errors can be averaged out. Lack of consistent failures can be interpreted as statistical independency. Thus methods for de-correlating the instances have been developed. The most popular is *bagging* [17], commonly used for random forests. It uses bootstrap sampling to learn less correlated instances from different subsets of the data.

The above works often discuss ensembling as a means of increasing performance. [18] approached high variance from the scope of *unreliability*. They discussed ensembling as a type of N-version programming, which advocates reliability through redundancy. When producing N-versions of a program, versions may fail independently but through majority voting they behave as a reliable system. They formalize intuitive requirements for reliability: (a) the target function to be covered by the ensemble and (b) the majority to be correct. This in turn advocates diversity, independence and overall quality of the components.

Biomedical applications are reliability-critical and high variance would deter the use of neural networks. For this reason we set off to investigate robustness of diverse ensembles. Diverting from the above works, we introduce another perspective of ensembling: creating an objective, configuration-invariant model to facilitate objective analysis.

3 Ensembles of Multiple Models and Architectures

A variety of CNN architectures has shown promising results in recent literature. Regarding the architectures, they commonly differ in depth, number of filters and how they process multi-scale context among others. Such architectural choices bias the model behaviour. For instance, models with large receptive fields may show improved localisation capabilities but can be less sensitive to fine texture than models emphasizing local information. Strategies to handle class imbalance is another performance relevant parameter. Common strategies are training with class-weighted sampling or class-weighted cross entropy. As analysed in [13], these methods strongly influence the sensitivity of the model to each class. Furthermore, the choice of the loss function impacts results. For example, we observed that networks trained to optimize Intersection over Union (IoU), Dice or similar losses [19] tend to give worse confidence estimations than when trained with cross entropy (Fig. 1). Finally, the setting of hyper-parameters for the optimization can strongly affect performance. It is often observed by practitioners that the choice of the optimizer and its configuration, for instance the learning rate schedule, can make the difference between bad and good segmentation.

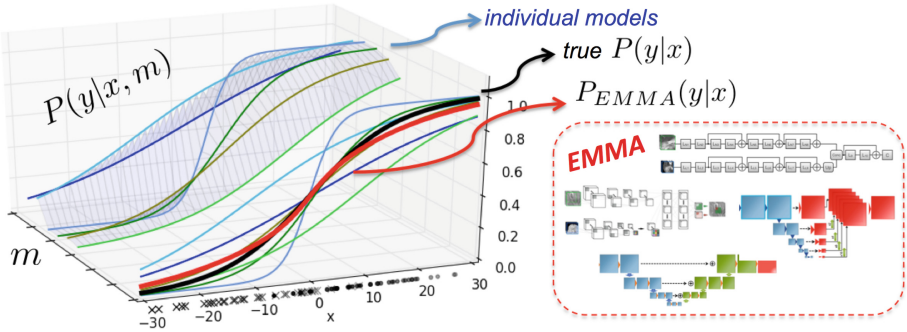

Fig. 2. Our ensemble of diverse networks, EMMA (red), averages out the bias infused by individual model configurations m, to approximate more reliably the true posterior (black), while being robust to suboptimal configurations. Posteriors on the left were obtained from multiple perceptrons, trained to classify clusters centred on 10 and −10 as a toy example, with different losses, regularizations and noise in the training labels. Their ensemble provides reliable estimates. (Color figure online)

The sensitivity to such meta-parameters is a greater problem than merely a time-consuming manual optimization of configurations:

- A configuration setting optimized on one set of training data may be overfitting them and not perform well on unseen data or another task. This can be viewed as another source of high model variance (Sect. 2).
- By biasing the behaviour of the model, it also biases the findings of any analysis performed with it.

We now formalize the problem and our perspective of ensembling as a solution as follows. Given training data X with labels Y, we need to learn the generating process $P(y|x)$. This is commonly approximated by a model $P(y|x;\theta_m, m)$, which has trainable parameters θ_m that are learnt via an optimization process that minimizes:

$$\theta_m = \min_{\theta_m} d(P(Y|X;\theta_m, m), P(Y|X)) \tag{1}$$

where d is a distance (defined by the type of loss) computed at the points given by the training data, while m represents the choice of the meta-parameters. It is commonly neglected although it conditions (biases) the learnt estimator. To take it into account, we instead define m as a stochastic variable over the space of meta-parameter configurations, with a corresponding prior $P(m)$. In order to learn a model of $P(y|x)$ unbiased by m, we marginalize out its effect:

$$\begin{aligned} P(y|x) = \sum_m P(y, m|x) &= \sum_m P(y|x, m)P(m) \\ &\approx \sum_{\forall m \in E} P(y|x;\theta_m, m)\frac{1}{|E|} = P_{EMMA}(y|x) \end{aligned} \tag{2}$$

Here E is the set of models within the ensemble. The prior $P(m)$ is considered uniform over a subspace of m that is covered by the models in E and zero elsewhere. Note we have arrived at the standard ensembling with averaging, by considering that each individual model $P(y|x;\theta_m, m)$ approximates a conditional $P(y|x, m)$ on m, and the true posterior is approximated by the ensemble which marginalizes away effects of m. Note that the case of a single model configured by m can be derived from the above, by setting a dirac prior $P(m) = \delta(m)$. Thus the ensemble relaxes a pre-existing neglected strong prior.

The above formulation presents averaging ensembles from a new perspective: The marginalization over a subspace of the joint $P(y|x, m)$ offers generalisation, regularising the (manual) optimization process of m from falling into minima where $P(Y|X, m)$ overfits $P(Y|X)$ on the given training data (Y, X) (Fig. 2). Moreover, the process leads to a more objective approximation of $P(y|x)$ where the biasing effect of m has been marginalized out. The exposed limitations agree with the requirements for ensembling mentioned in Sect. 2: we need to restrict the subspace of m into an area of relatively high quality models and we need to cover it with a relatively small number of models, thus diversity is key.

In the remainder of this section we describe the main properties of the models used to construct the collection E of EMMA, which cover various contemporary architectures, configured and trained under different settings[1].

3.1 DeepMedic

Model description: We include two DeepMedics in EMMA. The main architecture, originally presented in [13,20], is a fully 3D, multi-scale CNN, designed

[1] Implementation and configuration details considered less important for this work were omitted to avoid cluttering the manuscript.

with a focus on efficient processing of 3D images. For this it employs parallel pathways, with the secondary taking as input down-sampled context, thus avoiding to convolve large volumes at full resolution to remain computationally cheap. Although originally developed for segmenting brain lesions, it was found promising on diverse tasks, such as segmentation of placenta [21], making it a good component for a robust ensemble. The first of the two models we used is the residual version previously employed in BRATS 2016 [22], depicted in Fig. 3. The second is a wider variant, with double the number of filters at each layer.

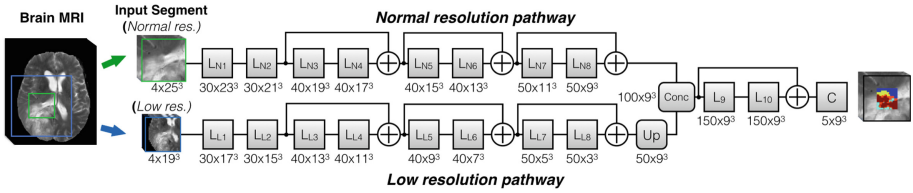

Fig. 3. We used two DeepMedics [13] in our experiments. The smaller of the two is depicted, where the number of feature maps and their dimension at every layer are depicted in the format ($Number \times Size$). The second model used in the ensemble is wider, with double the number of feature maps at every layer. All kernels and feature maps are 3D, even though not depicted for simplicity.

Configuration: The models are trained by extracting multi-scale image segments with a 50% probability centred on healthy tissue and 50% probability on tumour as proposed in [13]. The wider variant is trained on larger inputs, of width 34 and 22 for the two scales respectively. Both are trained with cross-entropy loss. All other meta-parameters were adopted from the original configuration.

3.2 FCN

Model description: We integrate three 3D FCNs [23] in EMMA. A schematic of the first architecture is depicted in Fig. 4. The second FCN is constructed larger, replacing each convolutional layer with a residual block with two convolutions. The third is also residual-based, but with one less down-sampling step. All layers use batch normalisation, ReLUs and zero-padding.

Training details: We draw training patches of width 64 for the first and 80 voxels for the residual-based FCNs, with an equal probability from each label. They were trained using Adam. The first was trained to optimize the IoU loss [19] while the Dice was used similarly for the other two. The trained models are then applied fully convolutionally on whole volumes for inference.

3.3 U-Net

Model description: We employ two 3D versions of the U-Net architecture [24] in our ensemble. The main elements of the first architecture are depicted in Fig. 5.

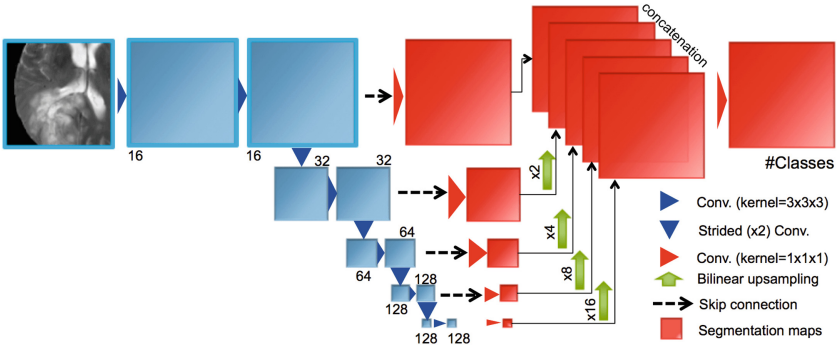

Fig. 4. Schematic of one of the FCN architecture used in EMMA. Shown are number of feature maps per layer. All kernels and feature maps are 3D, even though not depicted for simplicity.

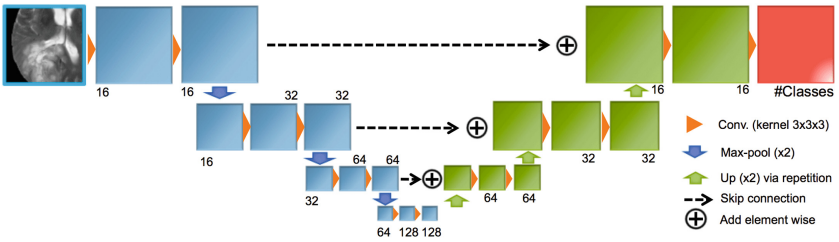

Fig. 5. Schematic of an adapted Unet used in our experiments. Depicted are number of feature maps per layer. All kernels and feature maps are 3D, even though not depicted for simplicity.

In this version we follow the strategy suggested in [25] to reduce model complexity, where skip connections are implemented via summations of the signals in the up-sampling part of the network, instead of the concatenation originally used. The second architecture is similar but concatenates the skip connections and uses strided convolutions instead of max pooling. All layers use batch normalisation, ReLUs and zero-padding.

Training Details: The U-Nets were trained with input patches of size $64 \times 64 \times 64$. The patches were sampled only from within the brain, with equal probability being centred around a voxel from each of the four labels. They were trained minimizing cross entropy via AdaDelta and Adam respectively, with different optimization, regularization and augmentation meta-parameters. The trained models are then applied fully convolutionally on whole volumes for inference.

3.4 Ensembling

The above models are all trained completely separately. At testing time, each model segments individually an unseen image and outputs its class-confidence

maps. The models are then ensembled into EMMA, according to Eq. 2. For this, the ensemble's confidence maps for each class are created by calculating for each voxel the average class confidence of the individual models. The final segmentation is made by assigning to each voxel the class with the highest confidence.

3.5 Implementation Details

The original implementation of DeepMedic was used for the corresponding two models, available on https://biomedia.doc.ic.ac.uk/software/deepmedic/. The FCNs were implemented using DLTK, a deep learning library with a focus on medical imaging applications that allowed quick implementation and experimentation (https://github.com/DLTK/DLTK). Finally, an adaptation of the Unet will be released on https://gitlab.com/eferrante.

4 Evaluation

4.1 Material

Our system was evaluated on the data from the Brain Tumour Segmentation Challenge 2017 (BRATS) [4,26–28]. The training set consists of 210 cases with high grade glioma (HGG) and 75 cases with low grade glioma (LGG), for which manual segmentations are provided. The segmentations include the following tumour tissue labels: (1) necrotic core and non enhancing tumour, (2) oedema, (4) enhancing core. Label 3 is not used. The validation set consists of 46 cases, both HGG and LGG but the grade is not revealed. Reference segmentations for the validation set are hidden and evaluation is carried out via an online system that allows multiple submissions. In the testing phase of the competition, a test set of 146 cases is provided to the teams, and the teams have a 48 hours window for a single submission to the system. For evaluation, the 3 predicted labels are merged into different sets of whole tumour (all labels), the core (labels 1,4) and the enhancing tumour (label 4). For each subject, four MRI sequences are available, FLAIR, T1, T1 contrast enhanced (T1ce) and T2. The datasets are pre-processed by the organisers and provided as skull-stripped, registered to a common space and resampled to isotropic $1\,\mathrm{mm}^3$ resolution. Dimensions of each volume are $240 \times 240 \times 155$.

4.2 Ensembling Multiple Pre-processing Methods

We experimented with three different versions of intensity normalisation as pre-processing: (1) Z-score normalisation of each modality of each case individually, with the mean and stdev of the brain intensities. (2) Bias field correction followed by (1). (3) Bias field correction, followed by piece-wise linear normalisation [29], followed by (1). Preliminary comparisons were inconclusive. We instead chose to average away the normalisation's effect with EMMA. For each of the seven networks in Sect. 3, three instances were trained, each on data processed with different normalisation. They were applied to correspondingly processed images for inference and all results were averaged in EMMA (Fig. 6).

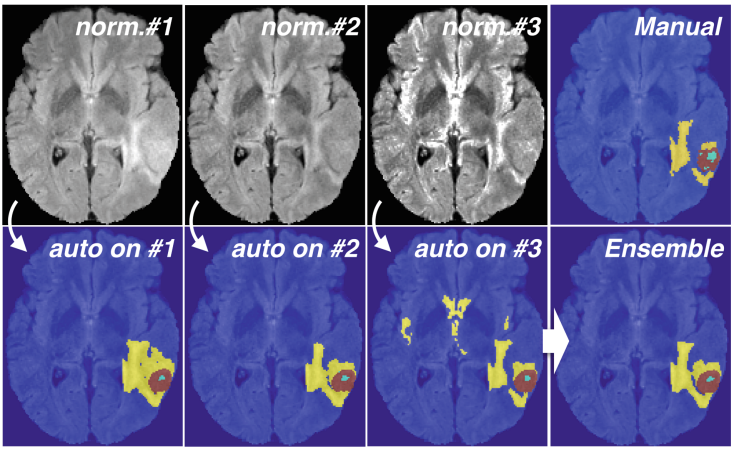

Fig. 6. Results are affected by normalization. To make a system robust to this factor, we introduce in EMMA models trained on differently normalized data.

4.3 Post-processing

The segmentations from EMMA were finally post-processed by removing secondary connected-components smaller than 250 voxels.

4.4 Results

We provide the results that EMMA achieved on the validation and testing set of the BRATS'17 challenge[2] on Table 1. Our system won the competition by achieving the overall best performance in the testing phase, based on Dice score (DSC) and Haussdorf distance. We also show results achieved on the validation set by the teams that ranked in the next two positions at the testing stage.

Table 1. Performance of EMMA on the validation and test sets of BRATS 2017 (submission id biomedia1). Our system achieved the top segmentation performance in the testing stage of the competition. For comparison we show the performance on validation set of the teams that ranked in the next two position. Performance of other teams in the testing stage is not available to us.

	DSC			Sensitivity			Hausdorff_95			#submits
	Enh.	Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	
EMMA (val)	73.8	90.1	79.7	78.3	89.5	76.2	4.50	4.23	6.56	2
UCL-TIG (val)	78.6	90.5	83.8	77.1	91.5	82.2	3.28	3.89	6.48	21
MIC_DKFZ (val)	73.2	89.6	79.7	79.0	89.6	78.1	4.55	6.97	9.48	2
EMMA (test)	72.9	88.6	78.5	–	–	–	36.0	5.01	23.1	1

[2] Leaderboard: https://www.cbica.upenn.edu/BraTS17/lboardValidation.html.

No testing-phase metrics are available to us for these methods. We note that EMMA achieves similar levels of performance on validation and test sets, even though the latter contains data from different sources, indicating the robustness of the method. In comparison, competing methods were very good fits for the validation set, but did not manage to retain the same levels on the testing set. This emphasizes the importance of research towards robust and reliable systems.

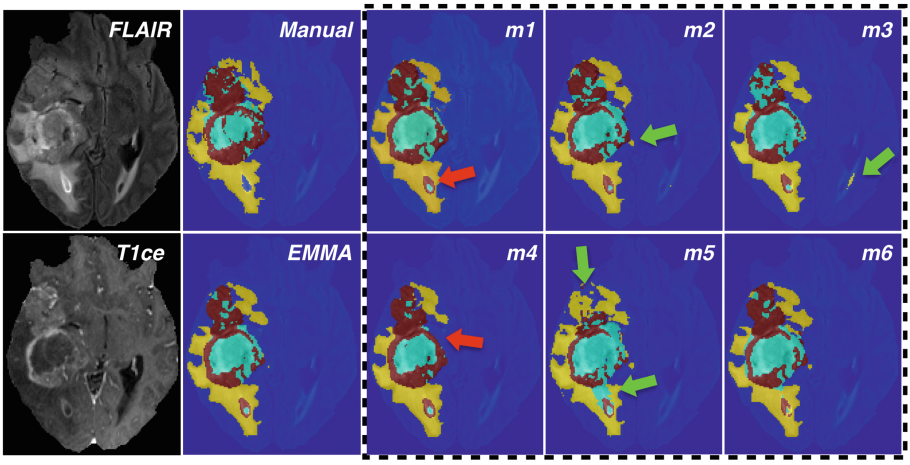

Fig. 7. FLAIR, T1ce and manual annotation of a case in the training set, along with automatic segmentation from preliminary version of EMMA consisting of six models. Green arrows point inconsistent mistakes by the individual models that are corrected by ensembling, while red arrows show consistent mistakes. (Color figure online)

5 Conclusion

Neural networks have been proven very potent, yet imperfect estimators, often making unpredictable errors. Biomedical applications are reliability-critical however. For this reason we first concentrate on improving robustness. Towards this goal we introduced EMMA, an ensemble of widely varying CNNs. By combining a heterogeneous collection of networks we construct a model that is insensitive to independent failures of CNN components and thus generalises well (Fig. 7). We also introduced the new perspective of ensembling for objectiveness. Biased behaviour, introduced by configuration choices, is marginalised out via ensembling, making EMMA a model more fit for objective analysis. Even though the individual networks in this work have straight-forward architectures and were not optimized for the task, EMMA won first place in the final testing stage of the BRATS 2017 challenge among 50+ teams, indicating strong generalisation.

By being robust to suboptimal configurations of its components, EMMA may offer re-usability on different tasks, which we aim to explore in the future. EMMA

may also prove useful for unbiased investigation of factors such as sensitivity of CNNs to different sources of domain shift that affect large-scale studies [30]. Finally, EMMA's uncertainty could serve as a more objective measure of what type of tumours are most challenging to learn.

Computational requirements of ensembles increase with their size. Inference time is commonly of interest. Conveniently, EMMA's models can be parallelised. If multiple GPUs are not available, parallelisation on CPUs may also be practical. As an indication, segmentation of a brain scan with DeepMedic takes five minutes on a single CPU thread. Thus parallelising EMMA's components on different threads allows practical inference times for various applications on modern workstations and CPU cluster. Where computational and storage requirements need to be minimal, knowledge distillation offers an attractive solution [31].

Acknowledgements. This work is supported by the EPSRC (EP/N023668/1, EP/N024494/1 and EP/P001009/1) and partially funded under the 7th Framework Programme by the European Commission (CENTER-TBI: https://www.center-tbi.eu/). KK is supported by the President's PhD Scholarship of Imperial College London. EF is beneficiary of an AXA Research Fund postdoctoral grant. NP is supported by Microsoft Research through its PhD Scholarship Programme and the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1). We gratefully acknowledge the support of NVIDIA with the donation of GPUs for our research.

References

1. DeAngelis, L.M.: Brain tumors. N. Engl. J. Med. **344**(2), 114–123 (2001)
2. Bauer, S., Wiest, R., Nolte, L.P., Reyes, M.: A survey of MRI-based medical image analysis for brain tumor studies. Phys. Med. Biol. **58**(13), R97 (2013)
3. Louis, D., et al.: The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. **131**(6), 803–820 (2016)
4. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE TMI **34**(10), 1993–2024 (2015)
5. Prastawa, M., Bullitt, E., Ho, S., Gerig, G.: A brain tumor segmentation framework based on outlier detection. Med. Image Anal. **8**(3), 275–283 (2004)
6. Gooya, A., Pohl, K.M., Bilello, M., Biros, G., Davatzikos, C.: Joint segmentation and deformable registration of brain scans guided by a tumor growth model. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011. LNCS, vol. 6892, pp. 532–540. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23629-7_65
7. Parisot, S., Duffau, H., Chemouny, S., Paragios, N.: Joint tumor segmentation and dense deformable registration of brain MR images. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7511, pp. 651–658. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33418-4_80
8. Bakas, S., et al.: GLISTRboost: combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Handels, H. (eds.) BrainLes 2015. LNCS, vol. 9556, pp. 144–155. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30858-6_13

9. Zikic, D., et al.: Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7512, pp. 369–376. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33454-2_46
10. Le Folgoc, L., Nori, A.V., Ancha, S., Criminisi, A.: Lifted auto-context forests for brain tumour segmentation. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Winzeck, S., Handels, H. (eds.) BrainLes 2016. LNCS, vol. 10154, pp. 171–183. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55524-9_17
11. Urban, G., Bendszus, M., Hamprecht, F., Kleesiek, J.: Multi-modal brain tumor segmentation using deep convolutional neural networks. In: BRATS-MICCAI (2014)
12. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. IEEE TMI **35**(5), 1240–1251 (2016)
13. 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. Med. Image Anal. **36**, 61–78 (2017)
14. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
15. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural Netw. **4**(1), 1–58 (2008)
16. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
17. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
18. Sharkey, A.J., Sharkey, N.E.: Combining diverse neural nets. Knowl. Eng. Rev. **12**(3), 231–247 (1997)
19. Nowozin, S.: Optimal decisions from probabilistic models: the intersection-over-union case. In: CVPR, pp. 548–555 (2014)
20. Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., Glocker, B.: Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. In: Proceedings of ISLES-MICCAI (2015)
21. Alansary, A., et al.: Fast fully automatic segmentation of the human placenta from motion corrupted MRI. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 589–597. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_68
22. Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B.: DeepMedic for brain tumor segmentation. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Winzeck, S., Handels, H. (eds.) BrainLes 2016. LNCS, vol. 10154, pp. 138–149. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55524-9_14
23. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
24. 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
25. Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdes-Hernandez, M., et al.: White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. arXiv:1706.00935 (2017)

26. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nat. Sci. Data **4**, 170117 (2017)
27. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017)
28. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive (2017)
29. Nyúl, L.G., Udupa, J.K., Zhang, X.: New variants of a method of MRI scale standardization. IEEE TMI **19**(2), 143–150 (2000)
30. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47
31. Bucilu, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Knowledge Discovery and Data Mining, pp. 535–541. ACM (2006)