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.C.H. Lee** **B. Kainz** **D. Rueckert** **B. Glocker**

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

1 Introduction

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 work 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 place in the BRATS 2017 competition.

2 Ensembles of Multiple Models and Architectures

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

[*] Equal contribution, in alphabetical order

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

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). 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: 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. We employ three architectures, DeepMedic, originally presented in [1, 2], 3D FCNs [3], and 3D versions of the U-Net architecture [4].

3 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. 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.

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			
	Enh.	Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	#submits
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

4 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. 1). We also introduced the new perspective of ensembling for objectiveness. By marginalizing out via ensembling the biased behaviour introduced by configuration choices, EMMA is a model more fit for objective analysis. Even though the individual networks have straight-forward architectures and were not optimized for the task, EMMA won the first position in the final testing stage of BRATS 2017 competition 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 could also be useful in unbiased investigation of factors such as sensitivity of CNNs to different sources of domain shift that is strongly affecting large-scale studies [5], or estimating amount of training data required for a task. Finally, EMMA's uncertainty could serve as a more objective measure of what type of patients or tumours are most challenging to learn.

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

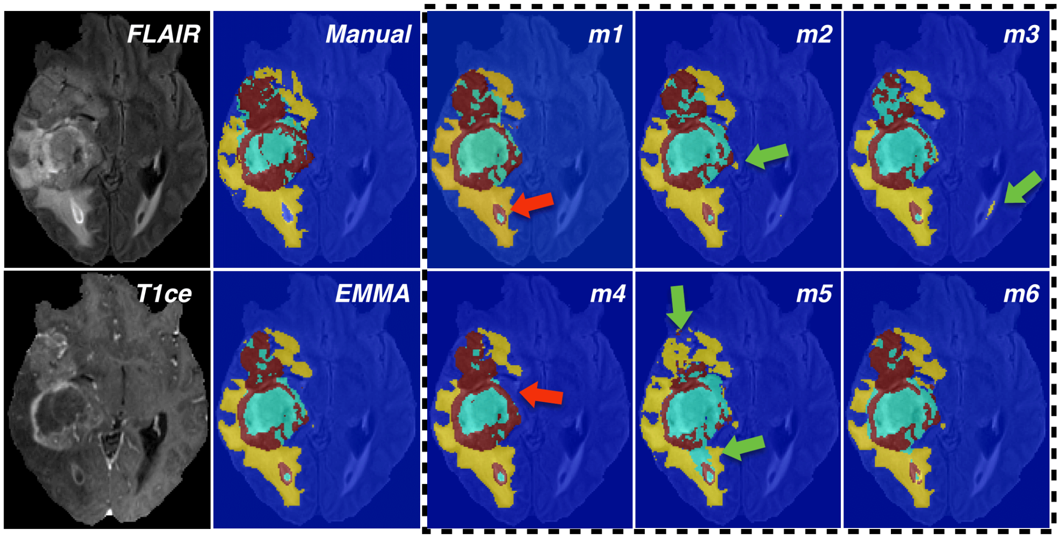

Figure 1: FLAIR, T1ce and manual annotation of a case in the training set, along with automatic segmentation from preliminary version of EMMA consisting of six models. Green arrows point inconsistent mistakes by the individual model that are corrected by the ensembling, while red arrow shows a consistent mistake.

5 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] Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., Glocker, B.: Multi-scane 3d convolutional neural networks for lesion segmentation in brain mri. in proc of ISLES-MICCAI (2015)

[2] 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** (2017) 61–78

[3] Long, J., et al.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431–3440

[4] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, Springer (2015) 234–241

[5] Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Information Processing in Medical Imaging, Springer (2017) 597–609