CLASSIFICATION OF MRI DATA USING DEEP LEARNING AND GAUSSIAN PROCESS-BASED MODEL SELECTION

Hadrien Bertrand[⋆†] *Matthieu Perrot*[†] *Roberto Ardon*[†] *Isabelle Bloch*[⋆]

[⋆] LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France
[†]Philips Research, France

ABSTRACT

The classification of MRI images according to the anatomical field of view is a necessary task to solve when faced with the increasing quantity of medical images. In parallel, advances in deep learning makes it a suitable tool for computer vision problems. Using a common architecture (such as AlexNet) provides quite good results, but not sufficient for clinical use. Improving the model is not an easy task, due to the large number of hyper-parameters governing both the architecture and the training of the network, and to the limited understanding of their relevance. Since an exhaustive search is not tractable, we propose to optimize the network first by random search, and then by an adaptive search based on Gaussian Processes and Probability of Improvement. Applying this method on a large and varied MRI dataset, we show a substantial improvement between the baseline network and the final one (up to 20% for the most difficult classes).

***Index Terms*—** Deep Learning, Convolutional Neural Networks, MRI, Classification, Model Selection, Gaussian Process

1. INTRODUCTION

In daily clinical practice, automatic analysis of medical images so as to determine the observed anatomy would produce significant benefits in terms of time and cost, given the large number of images acquired each day. This anatomical knowledge can 1) produce reliable search tools on ever increasing datasets (currently based on manually defined tags), supporting follow up of a specific patient's anatomy or finding similar anatomies or pathologies; 2) automate the production of reports on findings (e.g. transcribing tumors location); 3) automatically enrich exams to render an augmented visualization to ease communication between clinicians and patients.

While the imaging modality (CT, MR, US, etc.) is reliably provided by the acquisition systems, the information on the imaged anatomical region (chest, abdomen, spine, etc.) is given by manual annotations. Under the pressure of maximizing equipment exploitation, this information is very prone to errors. Yet, for any other further anatomical analysis, the reliability of this information is crucial. Thus, in the present work, we aim to automatically determine the imaged anatomical region from the image content. Solving this problem within the whole variability of medical imaging (modalities, protocols, patient anatomy, pathologies, etc.) would require a dataset currently inaccessible. Thus, we choose to restrict our study to MRI modality which already encompasses many aspects of the variability of the initial problem (high variability of protocols, anatomies, etc.). Moreover, our target space is limited to four anatomical regions: spine, head, abdomen and pelvis.

The recent surge of popularity and successes of deep learning for computer vision tasks has led to a wealth of applications, including in medical imaging [1]. The commonly adopted approach for classification tasks is to retrain a popular network model (AlexNet [2], VGG [3], etc.) on a specific dataset (sometimes just retraining the last few layers). While this strategy provides often correct results on small and controlled datasets, it reveals insufficient when confronted to the high variability of actual clinical data. This is the case of our dataset, which is only built from such day to day clinical images. Given the high number of degrees of freedom and their correlated impact on network model performances, handcrafting better models often proves very time consuming and inefficient.

In this paper, starting from a baseline network (Section 2.1), we present an automatic strategy to find better architectures given our clinical context. We constrain our search to a space of networks represented by a specific parametrization (Section 2.2) that includes enough diversity and, at the same time, promising models (including AlexNet-like, VGG-like). Inspired by the work in [4], we devise an automatic optimization process (Section 2.3) to produce an ensemble of successful models (Section 2.4). The interest and efficiency of this strategy are demonstrated on our application in Section 3.

2. METHOD

Under the condition that each MRI volume can be classified as one of our targeted anatomies (separating a volume into several ones if it covers several parts of the anatomy), and since MRI data are acquired in a slice by slice approach, we reduce the anatomical region prediction problem to a two dimension classification task (into abdomen, pelvis, head and spine). Assuming that information contained in a slice is rich enough, we effectively augment our dataset size. At the cost of loosing 3D information, we tend to a standard 2D-image classification problem, a good candidate for Deep Learning standard techniques.

2.1. Handcrafted Baseline Architecture

Several successful deep learning experiments in medical imaging are observed using architectures inspired by AlexNet [2] or VGG [3] (refer in particular to this special issue of TMI: [1]). Following these steps, after tedious parameter tuning, we converged to a model presenting a quite satisfactory behavior on our problem, but, at the same time, very difficult to further improve. Using the standard terms used by the deep learning community [5], this baseline model consists of 5 blocks, each comprising a convolution layer of 64 filters of size 3×3, followed by a rectified linear unit (ReLu) activation function and a max-pooling layer. The network ends with 2 fully-connected layers (resp. 4096 and 1024 units) interleaved with ReLU activations and terminated with a softmax decision layer. This network was trained by minimizing the categorical cross-entropy loss weighted by class

978-1-5090-1172-8/17/$31.00 ©2017 IEEE

Name	Range	Baseline
# blocks	$b \in [1; 5]$	5
# conv. layers per block	$c \in [1; 7]$	1
# filters per conv. layer	$2^r, r \in [2, 7]$	64
Size of the filters	$s \in \{3; 5\}$	3
Learning Rate	$10^l, l \in [-7; 0]$	0.001
Batch Size	$2^a, a \in [2; 8]$	8
# epochs	$10e, e \in [1; 10]$	70
Data augmentation	$g \in \{\text{Yes}, \text{No}\}$	Yes

Table 1. Hyper-parameters list. See Section 2.2 for details.

frequency (denoted L in the rest of this paper), using stochastic gradient descent (SGD) with Nesterov momentum ($m = 0.9$) and decay ($d = 10^{-6}$) .

2.2. Parametric Architecture and Hyper-parameters

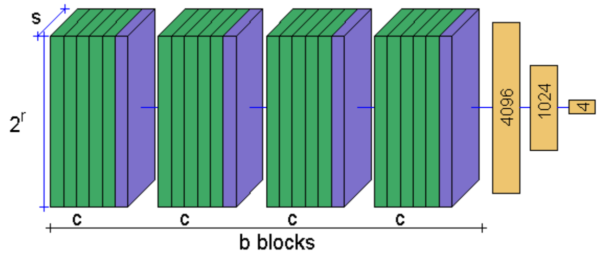


Fig. 1. Explored architectures

By relaxing some structural parts of our baseline architecture we define a wide and rich family of models. This parametric family has the following structure (see Figure 1): (1) a convolution block comprising b sections, each including c convolution layers of 2^r filters of size $s \times s$ interleaved with ReLU activations and terminated by a max-pooling layer, (2) the fully-connected layers as in our baseline architecture, and (3) a final softmax decision layer. Changes within this parametric space of models may drastically transform the optimization landscape, requiring to adjust training setting accordingly (in our case: learning rate, batch size and number of epochs, all other settings remaining identical). Moreover, using or not data augmentation is also considered, since more complex models require an increased amount of information. These architecture parameters and training settings form a collection of model hyper-parameters. Their respective ranges, detailed in Table 1, were defined so as to fulfill memory (less than 12GB) and time constraints (training should last less than one day). To fix ideas, each set of hyper-parameters Θ defines a classification problem that we aim to solve by training a classifier network f with weights W, in order to predict anatomical regions y_i from image slice content x_i: $y_i = f(x_i; W, \Theta)$.

2.3. Hyper-parameters Optimization

At this point, architecture and training parameters (a.k.a. model hyper-parameters) could be optimized with any suitable method. Given that the considered model family is exceedingly huge (more than 400K models) and that training of one network can take up to one day, any exhaustive coverage (like grid search) is intractable. Moreover, *a priori* discarding hyper-parameters with relative small impact remains difficult since it strongly depends on the specificity of the considered dataset [6]. Inspired by the work in [4] and [7], we developed an optimization strategy that consists of a random search warm-up stage followed by a Gaussian Process-guided search.

2.3.1. Random Search

Given the extent of the search space, some hyper-parameters or their combinations are expected to present only a moderate impact on our objective (minimizing slice-wise classification errors). Moreover, interesting hyper-parameter values are often consistent marginally (when integrating over a subset of other hyper-parameters). In this respect, uniform random exploration (Random Search [8]) is by far more efficient than any grid search.

2.3.2. Adaptive Search using a Gaussian Process

At some point, the coverage is sufficient to exploit the regularities of our hyper-parameters space with more advanced optimization techniques so as to accelerate the search. With such assumptions, we can consider to guide the exploration by estimating the performance of any unknown combination of hyper-parameters given performances related to all previously explored hyper-parameters values.

To this purpose, we use a Gaussian Process [9] to regress performances from hyper-parameters values. Smoothness is easily enforced thanks to a Gaussian Kernel whose scale is optimized by maximizing log-marginal-likelihood for each hyper-parameter. Estimation is very efficient; thus, the whole hyper-parameters space can be entirely covered to get, at any point Θ, an estimated loss following a Gaussian model: $\mathcal{N}(\hat{L}\left(\Theta\right), \hat{\sigma}^2\left(\Theta\right))$.

Many optimization strategies can be considered atop of such probabilistic estimations. Practically, the next hyper-parameters proposal is the one that maximizes the probability PI of overcoming a given target L^* [4]:

$$PI(\Theta) = \Phi\left(\frac{L^* - \hat{L}(\Theta)}{\sigma(\Theta)}\right) \tag{1}$$

where Φ denotes the normal cumulative distribution function. To foster a good balance between optimizing locally already identified good proposals and exploring farther regions with potential improvements, the optimization is driven by two targets in parallel: the best loss seen so far and an improvement of 25% from it.

2.4. Ensemble Model

At the end, our optimization process yields a collection of models ranked by their performance. The quality of this assessment is naturally limited since the size of the validation set used in this respect cannot encompass the diversity of clinical reality. Cross-validation could be used to get a better estimator of the performance, but we cannot practically afford its costs. Nevertheless, best models present very similar error rates and, at the same time, a good diversity of architectures (see Figure 2) that can be leveraged. In this paper, we select the top-ten models and build a robust classifier by averaging their predictions (other combinations could be used as well).

3. RESULTS

3.1. Dataset

The dataset consists of MRI images coming from a variety of hospitals and machines across the world (such as the *Centre Hospitalier*

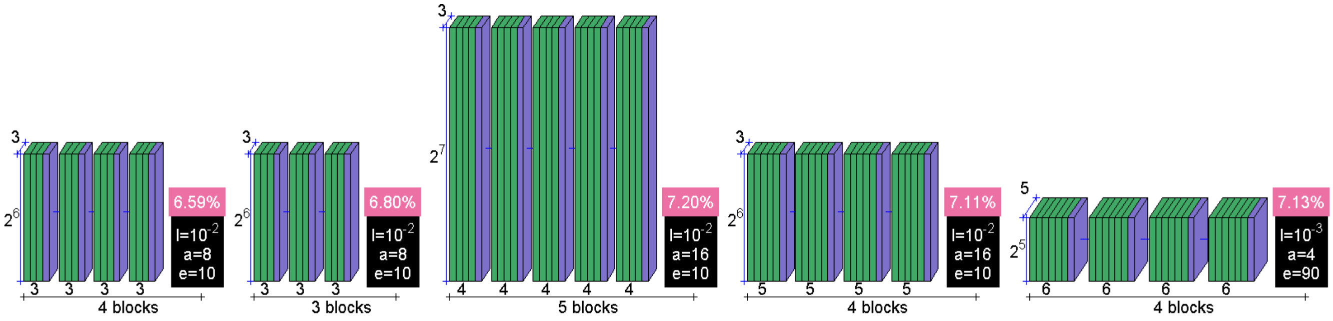


Fig. 2. Architecture of the 5 best models. Height represents the number of filters per layer, depth the size of the filters.

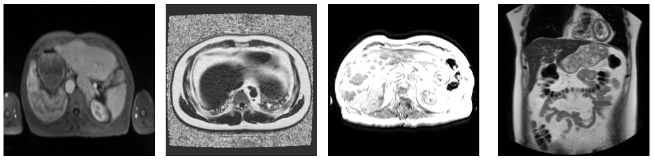


Fig. 3. A selection of axial and coronal abdomen slices showing the diversity and the complexity of our dataset.

Body Part	# Volumes	# Slices
Abdomen	282	11532
Head	301	9032
Pelvis	225	8854
Spine	386	7732

Table 2. Content of the dataset.

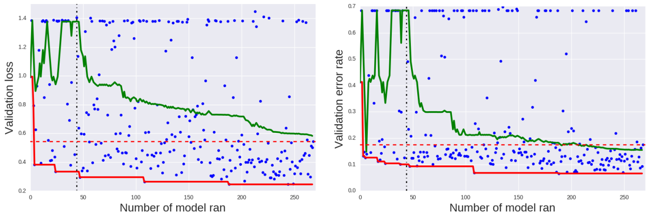


Fig. 4. Loss (left) and error rates (right) on validation dataset in function of the number of models considered so far by the optimization process. In green: running median, in solid red: running min, in dashed red: baseline performance, in dashed black: iteration from which random search is stopped in favor of adaptive search.

Lyon-Sud, France or *Johns Hopkins University, USA*). As a consequence our images display a large variety of protocols (see Figure 3) as well as resolution and number of slices per volume. In this paper, the considered regions are limited to: abdomen, head, pelvis and spine (table 2 sums up the content of our dataset).

Our dataset is splitted in a training set for the optimization of the weights W, a validation set for model selection (optimization w.r.t hyper-parameters) and a test set for model evaluation (resp. 50%, 25%, 25%). The separation is done volume-wise to take into account intra-subject slices correlations. Volumes containing multiple classes are split by anatomical regions and can end up in different sets. This raises the difficulty of the task since, in case of overfitting, predictions will be wrong at validation or testing phases. We also stratified classes across sets, giving us a proportion of slices per class close to the proportion of volumes per class.

Finally, each slice is subject to a unique step of preprocessing: it is resized to 64×64 pixels, a good trade-off between time constraints and quality of information.

Data augmentation consists in generating 80 000 images per epoch, which is 4 times as many images as the training set. The augmentation is done by applying translations, shearing and rotations, zooming, and adding noise.

3.2. Hyper-parameters Optimization

The hyper-parameters were optimized in two steps, 47 iterations of random search followed by an adaptive search (as described in Section 2.3.2). The entire process is depicted in Figure 4. Adaptive search presents quickly an important increase of the proportion of models with good performance (supported by the decreasing running median of the loss). Thus, selected combinations are on average better than random search.

Many proposals present both a high loss and a high error rate. These correspond to models that put all images in the same class, in this case: abdomen, which accounts for around 30% of the dataset (implying 70% of error rate).

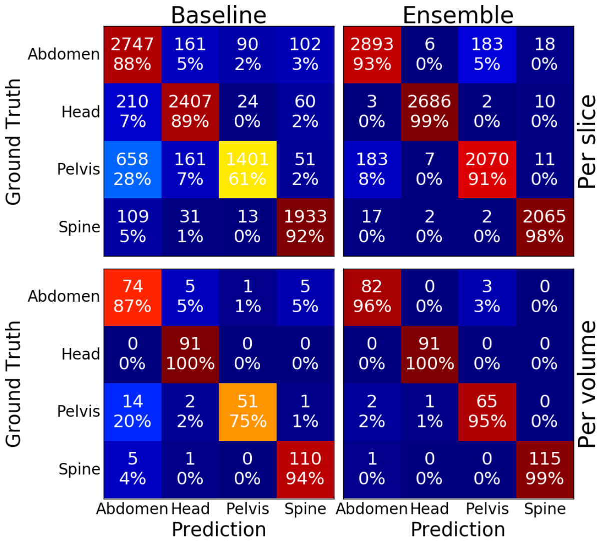
3.3. Test Accuracy

Figure 2 shows the architecture of some of the best models chosen for our ensemble. Despite the first two differing only in the number of blocks, others display variations across all hyper-parameters except data augmentation, which is always turned on. The learning rate is in a small range, either 0.0001 or 0.001, and the batch size is small (less than 16). Those networks tend to be deep (min. 8 convolutional layers) and the other hyper-parameters use a wide range of values.

In terms of accuracy, the ensemble is slightly better than the best model alone, however the ensemble benefits from a reduced bias.

Figure 5 shows the confusion matrices on the test set of the baseline and the ensemble of the 10 best models, demonstrating a substantial improvement on the classification of all anatomical regions. Most of the errors come from pelvis and abdomen, which was expected since the delimitation between those regions is ill-defined. In both cases pelvis is the class with the highest error.

For the volume classification, the choice of class is done by a majority vote on all slices of the volume. This gives us a higher accuracy. The ensemble misclassifies 444 slices from 71 volumes, but only 7 volumes produce errors. The misclassified slices usually



Per slice, Ground Truth	Baseline: Abdomen	Baseline: Head	Baseline: Pelvis	Baseline: Spine	Ensemble: Abdomen	Ensemble: Head	Ensemble: Pelvis	Ensemble: Spine
Abdomen	2747 88%	161 5%	90 2%	102 3%	2893 93%	6 0%	183 5%	18 0%
Head	210 7%	2407 89%	24 0%	60 2%	3 0%	2686 99%	2 0%	10 0%
Pelvis	658 28%	161 7%	1401 61%	51 2%	183 8%	7 0%	2070 91%	11 0%
Spine	109 5%	31 1%	13 0%	1933 92%	17 0%	2 0%	2 0%	2065 98%

Per volume, Ground Truth	Baseline: Abdomen	Baseline: Head	Baseline: Pelvis	Baseline: Spine	Ensemble: Abdomen	Ensemble: Head	Ensemble: Pelvis	Ensemble: Spine
Abdomen	74 87%	5 5%	1 1%	5 5%	82 96%	0 0%	3 3%	0 0%
Head	0 0%	91 100%	0 0%	0 0%	0 0%	91 100%	0 0%	0 0%
Pelvis	14 20%	2 2%	51 75%	1 1%	2 2%	1 1%	65 95%	0 0%
Spine	5 4%	1 0%	0 0%	110 94%	1 0%	0 0%	0 0%	115 99%

Fig. 5. Confusion matrices on the test set, for the baseline network (left) and the ensemble (right), computed on slices (first row) and on volumes (second row).

correspond to the first or last one of a volume, containing little information or being nearly part of another anatomical region.

3.4. Slice by slice analysis of a volume

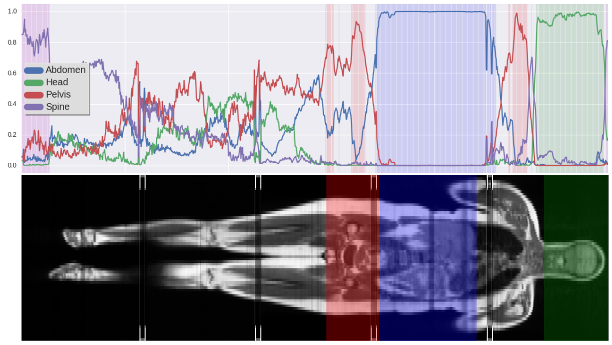


Fig. 6. Slice by slice classification on a full body volume. Top: Class probabilities. Filled areas correspond to decision made when the probability is higher than 0.7. Bottom: Volume and ground truth.

As an interesting example, we analyzed a full body volume by classifying each of its slices through our ensemble model. For each slice, the predicted class is the one with a probability higher than 0.7, and if no class meets this criterion, then we do not choose any. As we can see in Figure 6, the network is doing well at identifying the abdomen and the head. It also identifies correctly the pelvis, with some uncertainty. No class is dominant for most of the legs, however the feet are considered as spine with high probability. It also mistakenly identifies the neck as pelvis.

Those mistakes could be corrected by using a more complex decision criterion than a simple probability cutoff. We also expect that adding more anatomical regions to our dataset will allow for a better localization of the present regions.

4. DISCUSSION AND CONCLUSION

To the problem of finding more accurate networks than handcrafted ones, we have answered with two viable strategies. Random search is as easy to implement as grid search and quickly improves on the baseline, which makes it perfect for time-constrained situations. Without this constraint and at a higher cost in implementation, an adaptive search based on Gaussian Processes explores a range of highly accurate models suited for ensembling.

For the hyper-parameters where there is a “correct” answer, such as the learning rate (0.001) and the presence of data augmentation, the guided search quickly converges and most models inherit their values. We should remove them on further analysis and instead explore a wider range of architectures by adding hyper-parameters controlling the fully-connected layers, such as the number of units, the number of layers, adding new types of layers such as dropout and batch normalization placed across the network or even explore other learning method such as RMSProp or Adadelta.

One limitation of the current system is that some hyper-parameters depend on the value of others. For example we are unable to choose a different number of filters per layer as the number of layers is not fixed. We also limited the range of some hyper-parameters such as the filter size so as to have networks that would fit in memory and be trained in a reasonable amount of time. Only a subset of values combinations would cause a failure. Further work could incorporate constraints on time and memory either by precise estimations when possible or by measures during training to produce estimations with a GP.

Results on volume classification were very satisfactory. Since it is done at slice level, we have obtained a decent localization tool which shows the robustness of our ensemble. Further work will focus on adding more anatomical regions, which might require splitting slices in patches to identify smaller regions such as organs.

5. REFERENCES

[1] “Special Issue on Deep Learning in Medical Imaging,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, 2016.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. 2012.

[3] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv*, 2014.

[4] D. R. Jones, “A Taxonomy of Global Optimization Methods Based on Response Surfaces,” *Journal of Global Optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[5] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning,” The MIT Press, 2016.

[6] J. Bergstra and Y. Bengio, “Random Search for Hyper-parameter Optimization,” *The Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[7] J. Snoek, H. Larochelle, and R. P. Adams, “Practical Bayesian Optimization of Machine Learning Algorithms,” in *Advances in Neural Information Processing Systems*, 2012, pp. 2951–2959.

[8] R. L. Anderson, “Recent advances in finding best operating conditions,” *Journal of the American Statistical Association*, vol. 48, no. 264, pp. 789–798, 1953.

[9] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2005.