Skull Stripping: An Active Learning Approach

This work was part of a civilian service deployment.

Jan Riedo
jan.riedo@students.unibe.ch

March 4, 2019

1 Introduction

Skull stripping is widely known as *solved* task. However, state of the art algorithms show poor performance on datasets with lesions or cavities. Even different modalities (different contrast/quality) are still a major challenge for current algorithms[1, 2]. As with other medical image analysis problems, the lack of high quality manually labeled data remains an issue for deep learning (DL) algorithms. Biomedical image annotation is a very labour- and cost-intensive task.

This is where active learning (AL) comes into play. Active learning targets at having a higher accuracy with fewer labeled data, if the algorithm can define which samples to take for training from a large pool of unlabeled data [3]. With this approach, only a fraction of the dataset, namely the samples defined by the AL algorithm needs to be labeled. This reduces the labour work. The key to success in active learning is an optimal sample strategy to select (query) the best sample for training. The best samples are the most informative, which means if the algorithm would know the label of this certain sample, it would perform substantially better. Most query strategies try to find a balance between the most uncertain sample and the sample which is most representative of the underlying distribution.

A variety of different query strategies are found in literature with uncertainty sampling [4] being the most abundant and simplest one; The active learner queries the instances about which it is least confident (closest to 0.5 in a binary segmentation). Query-By-Committee [5] is constituted as a set of competing models. The instance on which they most disagree is queried. The expected model change [6] approach selects the instance that would lead to the greatest change in the current model if the algorithm knew the label. The expected error reduction [7] algorithm queries the instance with minimal expected future error. It was found to work better than uncertainty sampling for certain problem settings [8]. The variance reduction [9], which is used for regression problems tries to reduce the generalization error by

minimizing the output variance. In contrast, density-weighted methods [10] try to reduce error and variance which makes them less prone to querying outliners. Recently, approaches were proposed, where the sample selection is learned, instead of being applied as a fixed algorithm [11, 12, 13].

In this project, skull stripping was performed on three datasets, using an adapted U-Net [14] within an AL framework. Random sampling set the baseline performance, which was surpassed by the majority of applied query strategies. Simple uncertainty sampling was complemented with a representativeness measure. Transfer learning (TL) is applied, employing a model trained conventionally on a full dataset. The model is then refined with the AL strategy for another dataset.

2 Related Work

Since the 2015 ISBI cell tracking challenge, the winning network, U-Net[14], is used for all sorts of biomedical segmentation tasks. The network combines a convolution and deconvolution path to obtain a semantic segmentation[15]. Recent advances were made with the U-Net applied on 2D images in the fields of dermatology [16], furthermore, 3D U-Nets were applied for MR image segmentation[17]. However, Baumgartner et al. [18] found 2D-processing to be more successful in certain MR image segmentation tasks compared to 3D.

Recent progress is made with AL in a deep learning (DL) framework in automated biomedical image annotation [19]. For general AL applications, Bayesian uncertainty sampling is proposed as query strategy [20]. Disagreement of models (uncertainty) combined with representativeness can enhance the performance either in a two-step process [21] (strategy adopted for this work) or in a direct combination via Borda count [22]. Batch-wise sample selection is proposed for query efficiency [23].

Skull stripping with DL is a well researched topic; various approaches in 3D [17, 24], as well as in 2D [25, 26] were made. The 2D models were able to compete with the 3D ones, with an advantage in efficiency. This is one reason for the 2D approach in this work. Another, more important one is, that slicewise annotation by experts is more convenient. State of the art neuroimage software provide routinely used packages for brain extraction [1, 27, 28, 29]. However, they often have a poor performance when certain geometric assumptions do not hold [30].

Transfer learning (TL) aims at transferring the knowledge from one domain to another [31]. The task of skull stripping fits perfectly into the focus of TL, as, as written above, state of the art segmentation software still struggles with domain adaption. Transfer learning from ImageNet [32] was tested for medical images [33, 34], but does not have satisfactory performance. In this work, the model of one dataset was used as base model

for another one.

3 Methods

3.1 Dataset

Two different datasets of healthy human brain magnetic resonance (MR) images and one with post-op images of brain tumor patients were employed (see Tab. 1).

The Internet Brain Segmentation Repository[1] (IBSR) dataset has 18 T1-weighted scans of 256 x 256 x 128 voxels and nonisotropic spacing of 0.94 x 0.94 x 1.5 mm^3. The 2D coronal slices were loaded in full resolution of 256x256 voxels.

The Human Connectome Project[2] HCP dataset has 100 T1- and T2-weighted images of size 260 x 311 x 260 with homogeneous spacing of 0.7 mm^3. In this project, only the 2D axial T1 slices were used.

The Cavity dataset obtained from the Inselspital, Bern University Hospital, has 20 T1-weighted post-contrast images of size 160 x 256 x 256, from which every fourth slice was manually segmented by a non-expert. This results in a voxel spacing of 1 x 1 x 4 mm^3 and 64 axial slices. The images are from subjects immediately (within 24 hours) after a brain tumor removal surgery.

All datasets were split into three subsets (Tab. 1); seed, unlabeled and validation. The seed (assumed to be labeled data) was used to sample for primarily training the network. The unlabeled set served as pool for queries, from which samples were selected, "labeled" and added to the training set. The validation set was used to observe the segmentation performance on unseen data.

Table 1: The three datasets with respective splits used in this work.

set	images / slices	seed	validation	unlabled
IBSR	18 / 2304	2/256	5/640	11/1408
HCP	100 / 26000	10/2600	30/7800	60/15600
Cavity	20 / 1280	3/192	6/384	11/704

3.2 Implementation details

The scenario is a large dataset with few labeled, and a majority of unlabeled (or badly labeled) data. Therefore, pool-based sampling [4], where the best samples are chosen from a pool of unlabeled data, is applied. All trainings started with a seed size of 4 arbitrary samples (slices) and one single sample

[1] http://www.cma.mgh.harvard.edu/ibsr/

[2] http://www.humanconnectomeproject.org

was added each iteration. The query strategy was refined over the course of this project. Each approach was tested against a random sampling.

A learning rate of 10^{-4} was employed with the Adam [35] optimizer and a binary cross entropy (BCE) loss function. A modified version of the U-Net [14] with 31'030'593 trainable parameters was employed. Instead of valid convolutions, same-size convolutions were used. The standard four downsampling steps (max pooling) on the analysis path of the network, as well as the up-sampling (transposed convolutions) steps on the synthesis path were taken as-is from [14]. Dropout regularization was disabled for all runs presented. As the downsampling requires certain dimensions, the HCP slices were padded with zeros (see Tab. 2).

Table 2: Key points of the U-Net applied per dataset.

set	original size	in/out size	lowest size
IBSR	256 x 256	256 x 256	16 x 16
HCP	311 x 260	320 x 288	20 x 18
Cavity	256 x 160	256 x 160	16 x 10

The models were trained and retrained for a maximum of 50 epochs per iteration, for 50 iterations. One iteration of active learning comprises of querying a sample, adding it to the train set and training (retraining) the model on this expanded set of data. The batch size was limited by the GPU memory and set to 10 samples for training and 40 samples for testing. The latest model, as well as the best model per performance measure was saved for later use. Furthermore, the slices selected by the AL algorithm and the predictions were saved for visual examination. Intensity rescaling from 0 to 4096 was carried out during runtime. The implemented version of a kind of early stopping [36] is only based on the training loss. However, the results showed that the approach is feasible for this problem setting. The criterion for stopping was set at three times in a row a change in total loss per epoch of less than 0.001.

The implemented uncertainty sampling is the key of the whole AL pipeline. All slices of the unlabeled subset were fed through the network, the probabilities of the pixels being 1 (brain) were obtained, and the sample with the highest uncertainty was selected for further training. To achieve this, an overall uncertainty measure per slice had to be defined from the voxel-wise uncertainty. The sum of the deviations from 0.5 (perfect uncertainty) showed clear bias towards slices with a lot of brain rim in them. In order to make the uncertainty a bit more independent of the actual number of brain voxels in the slice, the 80 percentile of the probability distribution function of the absolute voxelwise deviations of 0.5 was taken for the slicewise uncertainty estimation.

As the selected slices should not only be the most uncertain ones, but

represent the underlying distribution, a representativeness selection was conducted on top of the uncertainty measure. The representativeness was measured as the correlation coefficient (normalized cross-correlation) Eq. (1) of the sample images and the mean of the candidate set. The candidate set consisted of 64 samples, selected by the uncertainty sampling, from which the most representative one was chosen for training. For some experiments, the most uncertain 256 samples were not taken into account, as this method may show clear bias towards samples with scanning artifacts.

$$CC = \frac{\sum_{i \in \Omega_{A,B}} (I_A(p_i) - \mu_A)(I_B(p_i) - \mu_B)}{\sqrt{\sum_{i \in \Omega_{A,B}} (I_A(p_i) - \mu_A)^2} \sqrt{\sum_{i \in \Omega_{A,B}} (I_B(p_i) - \mu_B)^2}} \tag{1}$$

3.3 Evaluation metrics

For evaluation, Dice and sensitivity were the two metrics reported on the brain mask. The Dice coefficient is found in most publications as the status quo for general segmentation comparison and skull stripping performance measure [37]. Sensitivity is reported to ensure that the cavity coming from tumor removal is still masked as brain. It was more important in this work to have the whole brain masked with the possible drawback of segmenting too much. Specificity is not reported, however was checked during training and showed always satisfactory results.

During the first experiments, it became apparent, that the state of the art performance metrics do not necessarily correlate with the actual image quality, when compared visually. In order to prevent misinterpretations from the metrics, a constrained Dice metric was introduced, which represents the normal Dice, but only applied on a rim of 10 pixels (geodesic distance) of the brain.

4 Experiments and Results

The first experiment was a simple proof of concept on the IBSR dataset. For all experiments, four seed samples was found to be the minimum for a stable training behavior. An upscaling onto the HCP dataset showed, that the metrics do not necessarily correlate with the segmentation quality. A spatial constraint on the evaluation area for the Dice measure showed a slightly better correlation. The AL stragety developed on the IBSR dataset was not able to surpass the random baseline on the HCP set. A more sophisticated query strategy was tested with no success. Furthermore, coronal slicing was found to be superior to axial slicing. The HCP algorithm showed remarkable performance on a set with post-operative samples. Finally, transfer learning was applied with a net trained on the HCP dataset and applied on the Cavity.

IBSR A baseline was set on the IBSR dataset with the query strategy being pure random sample selection. For better comparison, the seed set was hold fixed to a small set of 4 slices. The baseline was challenged by a simple uncertainty selection algorithm described in section 3.2. In Fig. 1, the algorithm showed faster rise and higher overall values for both performance measures compared to random sampling.

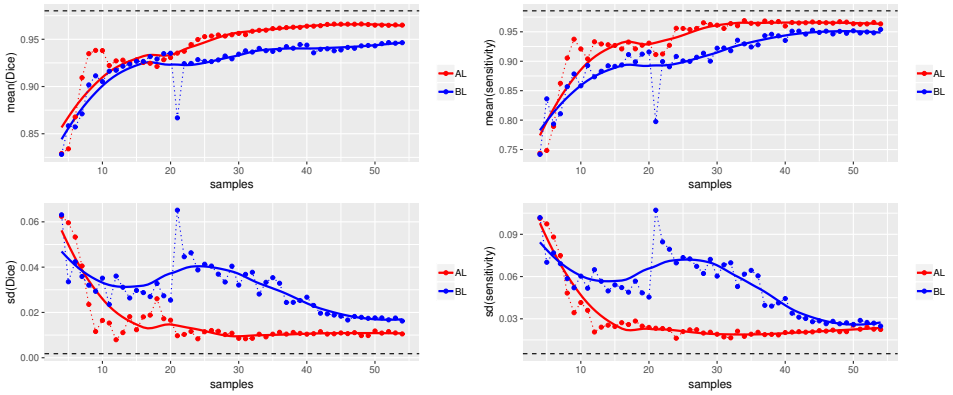


Figure 1: Dice and sensitivity measures for the IBSR dataset. AL: Active Learning algorithm; BL: Baseline random sampling; Dashed: On all 1664 training slices for 477 epochs; Trend: Local Polynomial Regression Fitting (`loess`) function of R with an α of 0.6.

HCP The algorithm developed for the IBSR dataset showed very poor performance on the HCP dataset. The introduction of the two-step procedure with an uncertainty candidate set and a representativeness measure showed substantial improvement in terms of stability. However, the baseline still outperformed the AL algorithm over all iterations. An inspection of the slices queried revealed, that some of the uncertainty samples were very noisy with only little actual information content (see Fig. 2 top row). Ignoring the top uncertain 256 samples prevented such misselections (see Fig. 2 bottom row).

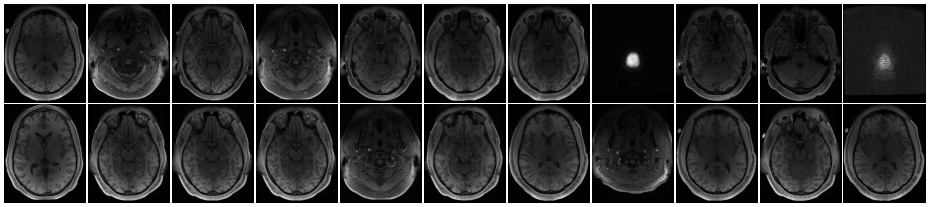


Figure 2: First 11 queried images by the AL algorithm on the HCP set based on most uncertain (top) and secondmost uncertain (bottom) samples.

Some samples in figure 2 come from the same subject and look very

similar; the images come from four (top row), and three (bottom row) images respectively. This is a very small number actually taken into account, compared to the 60 possible choices of the unlabeled set.

Although this adaption showed significant improvement, the AL query algorithm was still unable to outperform the random sampling (see Fig. 3). However, up to a training size of 10, it was better than random and both achieved the same long-term performance. Roughly 12 hours training time were required for 50 iterations. The AL algorithm took approximately double the time as the random sampling. The constrained Dice measure shown in Fig. 3 is qualitatively identical to the normal Dice.

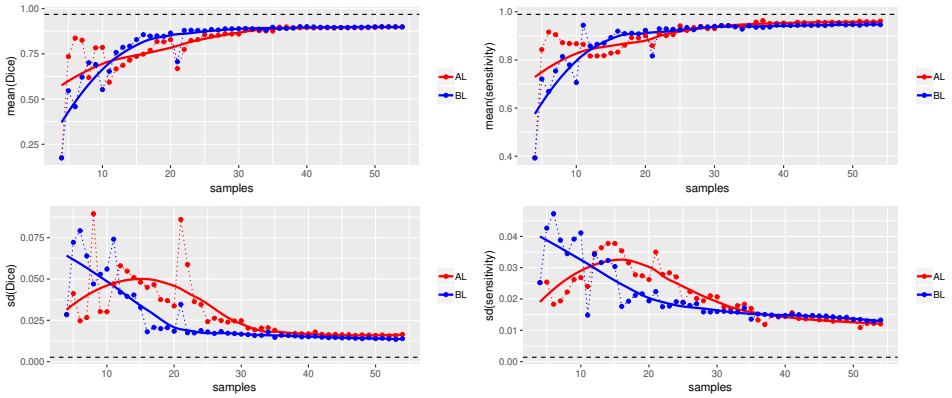


Figure 3: The newly developed constrained Dice and the sensitivity measures for the HCP dataset. AL: Active Learning algorithm; BL: Baseline random sampling; Dashed: On all 18200 training slices for 686 epochs; Trend: Local Polynomial Regression Fitting (`loess`) function of R with an α of 0.6.

HCP coronal To observe the influence of the slice orientation on the performance, one experiment with the HCP dataset was carried out with coronal slices instead of axial ones. It was inspired by the IBSR set, where coronal slices were used, and a better outcome in performance measures was obtained. The query algorithm was the same as before, again without ignoring the top 256 uncertain candidates. The approach outperforms the results in Fig. 3 in terms of overall performance and faster rise. Furthermore, AL was able to outperform random sampling.

Cavity Even though the HCP algorithm only showed mediocre performance, it was applied to the cavity dataset. Thereby, the performance of the algorithm on non-healthy subjects was tested, which is one of the possible uses for this algorithm. Surprisingly, the uncertainty-representativeness sampling (without ignoring the top 256) did a great job on the cavity dataset (see Fig. 4). It was able to outperform the random sampling and even get

close to the passively trained net on all data. A visual examination of the 3D images does confirm the good performance, no issues could be detected with the segmentation of the cavities.

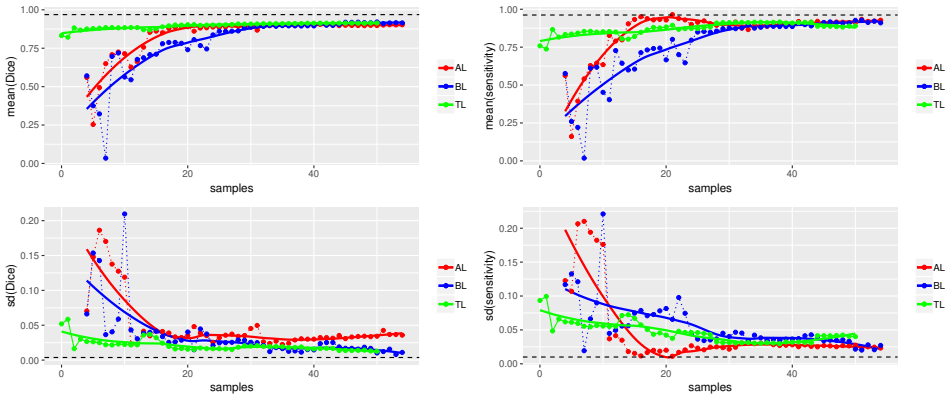
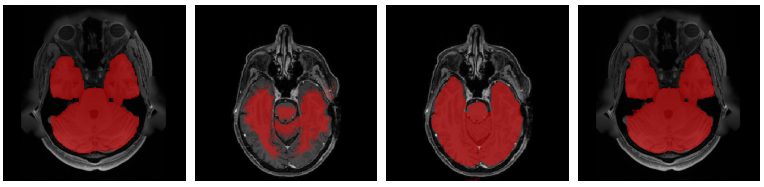


Figure 4: Dice and sensitivity measures for the Cavity dataset. AL: Active Learning algorithm; BL: Baseline random sampling; TL: Transfer Active Learning Dashed: On all 896 training slices for 453 epochs; Trend: Local Polynomial Regression Fitting (`loess`) function of R with an α of 0.6.

Transfer Active Learning In order to get an idea on how good the models generalize, one single experiment of transfer learning and refinement was conducted (Fig. 4, green line); In figure 5(a), the segmentation result of the HCP network, which was trained on all 18200 axial slices for 686 epochs is shown. It was used as initialization for a refinement on the Cavity set. To do so, the Cavity slices were padded to HCP dimensions of 320 x 288. A mean Dice value of over 0.8 was reached without further training (Fig. 4). Visual inspection of the predicted segmentation however revealed a pretty poor performance, visible in figure 5(b).



(a) HCP (b) Cavity (c) Cavity (d) HCP

Figure 5: Visual inspection of the HCP conventional trained network results. (a), (b): Before refinement;(c),(d): After refinement

Despite the fact, that the metrics did not increase greatly afterwards, a substantial improvement in segmentation image quality was observed (Fig. 5(c)). The refinement of the HCP net on the Cavity set did not substantially lower

the segmentation performance (qualitatively) on the HCP set (Fig. 5(d)). Finally, the network being trained from scratch (with AL) on the Cavity dataset only, caught up after 10-20 iterations, with comparable image quality.

5 Discussion and Conclusion

In this work, an active learning algorithm was developed and tested on three different datasets with healthy and impaired subjects. The majority of the experiments demonstrated the superiority of the AL query strategy when performing against random sampling. On the HCP set, the performance was similar to that of random sampling. All algorithms tested showed slightly smaller maximum performance compared to a network trained on the whole datasets.

The two smaller datasets, IBSR with 18 subjects, and Cavity with 20 subjects were no problem for the AL approach to be segmented and to outperform the baseline. However, if applied to the approximately five times larger HCP dataset, the algorithm gets beaten by the baseline (in the case of axial slicing). The rationale behind this could be the characteristics of the individual image modalities and the bigger variance within the dataset. Figure 2 reveals, that several consecutive queries take samples from the same subject and even the same region. This is very unlikely with the random sampling. Furthermore, it could lead to the poorer performance of the AL algorithm, because of the smaller win in information content due to those consecutive slices.

Another reason for the mediocre performance on the HCP set is the characteristic of the sampling strategy. It is well known, that uncertainty sampling can deteriorate the AL performance if noisy images are present in the dataset [38]. We tried to reduce this bias by ignoring the top 256 uncertain images in the candidate set. However, no substantial change was observed. Furthermore, trimming the algorithm to have a fast rise in performance metrics implicitly trains the algorithm to select slices of large brain regions. Those slices lead to a higher probability for brain segmentation and thus (because of the nature of the selected performance measures) to an increase in performance. They also lead to a more stable training behaviour (if no brain is present in the first few samples, nothing will be segmented).

The developed rim Dice is not really useful. There are situations, where the normal Dice overestimates the overall performance, and the constrained Dice gives a better impression of the quality. Generally, the Dice metric has to be interpreted with care when applied to a skull stripping segmentation. A high Dice does not necessarily mean a qualitatively good segmentation, which is apparent when comparing figure 4, green line and figure 5(b).

Tests with coronal slices, as expected, lead to higher overall performance.

The AL algorithm always beat the baseline random sampling. This can be explained with the fact that coronal slices contain less brain-free and head-free regions. Therefore, the algorithm has less possibilities to choose informationless samples. Furthermore, mere noise, one of the pitfalls in uncertainty sampling, are rare in coronal slices, compared to axial ones.

Intensity normalization of the images was implemented at the beginning of this project. It lead to a slower increase of performance with lower maximum. However, intensity rescale improved overall performances. The small amount of data, and the fact that the network is retrained every time one slice is added to the train set, can easily lead to overfitting. The introduction of dropout layers (with p = 0.5 or p = 0.25) in the U-Net did not contribute substantially to an improvement of stability. Moreover, it showed lower overall metrics even after 3.5 times longer training in terms of epochs. Furthermore, the implemented version of an early stopping did prevent overfitting satisfactorily.

The transfer learning (TL) experiment showed nicely, that an algorithm well trained for one set of medical images can still have a relatively poor performance on an anatomically identical dataset with different modalities. Although the metrics showed good performance with fewer samples in figure 4 (green line), the quality of the segmentations, generated with the TL net was not superior to that of the normal AL algorithm. Therefore, for this task, there was no real benefit of the TL approach.

6 Future work

As this project was quite time limited, some interesting topics shall be mentioned, which are left for future investigation.

- As the task of skull stripping is quite well defined, it may be possible to learn or estimate a simple distribution over the range of axial slices, which defines how probable it is to take a certain slice into the training set for AL training.
- The proposed approach should be tested on the MANAGE dataset, which is challenging for current skull stripping algorithms. The existing pipeline enables easy adaption of the main program for other datasets. A framework for manually segmenting slices during runtime would need to be created.
- A simple heuristic for the seed set selection could prevent early breakdown of algorithms. Furthermore, it could rise the overall skull stripping performance.
- The query algorithm could be replaced by the segmentation expert in charge, which could tell the AL framework which region (slice) it

segmented poorly and should be retrained.

Acknowledgements

The author gratefully acknowledges the support of the whole MIA group of the Institute for Surgical Technology and Biomechanics. A special thanks goes to Alain Jungo and Michael Rebsamen, who readily provided one-on-one support.

References

[1] J. E. Iglesias, C. Liu, P. M. Thompson, and Z. Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, Sept 2011.

[2] Kristi Boesen, Kelly Rehm, Kirt Schaper, Sarah Stoltzner, Roger Woods, Eileen Lueders, and David Rottenberg. Quantitative comparison of four brain extraction algorithms. *NeuroImage*, 22(3):1255 – 1261, 2004.

[3] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2010.

[4] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 3–12, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[5] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT 1992*. ACM Press, 1992.

[6] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008.

[7] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[8] Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from*

labeled to unlabeled data in machine learning and data mining, volume 3, 2003.

[9] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[10] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*, 2008.

[11] Meng Fang, Yuan Li, and Trevor Cohn. Learning how to active learn: A deep reinforcement learning approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605. Association for Computational Linguistics, 2017.

[12] Philip Bachman, Alessandro Sordoni, and Adam Trischler. Learning algorithms for active learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 301–310, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[13] M. Woodward and C. Finn. Active one-shot learning. In *NIPS Workshop on Deep Reinforcement Learning*, 2016.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[15] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1520–1528, Washington, DC, USA, 2015. IEEE Computer Society.

[16] Marc Gorriz, Axel Carlier, Emmanuel Faure, and Xavier Giró. Cost-effective active learning for melanoma segmentation. *CoRR*, abs/1711.09168, 2017.

[17] Andrew Beers, Ken Chang, James Brown, Emmett Sartor, CP Mammen, Elizabeth Gerstner, Bruce Rosen, and Jayashree Kalpathy-Cramer. Sequential 3d U-Nets for biologically-informed brain tumor segmentation. *arXiv*, 2017.

[18] Christian F. Baumgartner, Lisa M. Koch, Marc Pollefeys, and Ender Konukoglu. An exploration of 2d and 3d deep learning techniques for cardiac MR image segmentation. In Mihaela Pop, Maxime Sermesant, Pierre-Marc Jodoin, Alain Lalande, Xiahai Zhuang, Guang Yang,

Alistair Young, and Olivier Bernard, editors, *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 111–119, Cham, 2018. Springer International Publishing.

[19] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. *arXiv*, 2018.

[20] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *arXiv*, 2017.

[21] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. *arXiv*, 2017.

[22] Firat Ozdemir, Zixuan Peng, Christine Tanner, Philipp Fuernstahl, and Orcun Goksel. Active learning for segmentation by optimizing content information for maximal entropy. *arXiv*, 2018.

[23] Steven C. H. Hoi, Rong Jin, Jianke Zhu, and Michael R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 417–424, New York, NY, USA, 2006. ACM.

[24] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. Deep MRI brain extraction: A 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460 – 469, 2016.

[25] Raunak Dey and Yi Hong. Compnet: Complementary segmentation network for brain MRI extraction. *arXiv*, 2018.

[26] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Auto-context convolutional neural network for geometry-independent brain extraction in magnetic resonance imaging. *CoRR*, abs/1703.02083, 2017.

[27] Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

[28] Mark Jenkinson, Mickael Pechaud, Stephen Smith, et al. Bet2: MR-based estimation of brain, skull and scalp surfaces. In *Eleventh annual meeting of the organization for human brain mapping*, volume 17, page 167. Toronto., 2005.

[29] Gang Lin, Umesh Adiga, Kathy Olson, John F Guzowski, Carol A Barnes, and Badrinath Roysam. A hybrid 3d watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 56(1):23–36, 2003.

[30] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE transactions on medical imaging*, 36(11):2319–2330, 2017.

[31] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[33] Shin Hoo-Chang, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285, 2016.

[34] Bram Van Ginneken, Arnaud AA Setio, Colin Jacobs, and Francesco Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 286–289. IEEE, 2015.

[35] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[36] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, Aug 2007.

[37] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L. Rubin, and Bradley J. Erickson. Deep learning for brain MRI segmentation: State of the art and future directions. *Journal of Digital Imaging*, 30(4):449–459, Aug 2017.

[38] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 65–72, New York, NY, USA, 2006. ACM.