A Study of Acquisition Functions for Medical Imaging Deep Active Learning

Bonaventure F. P. Dossou
McGill University - Mila Quebec AI Institute
Masakhane NLP, Lanfrica
Lelapa AI
bonaventure.dossou@mila.quebec

Abstract

In the real world, especially in the medical imaging context, data scarcity and limited labeled data are recurrent and frequent problems. This is very often a bottleneck to high-performance of recent Deep Learning approaches that are very data-hungry. In this work, we show that active learning could be very effective in data scarcity situations, where obtaining labeled data is expensive. We compare several acquisition functions (BALD, MeanSTD, and MaxEntropy) on the ISIC 2016 Melanoma detection dataset, explore the impact of selecting either the most or least uncertain samples, and leverage the effect of acquired pool sizes on the performance of the model. Our results on the Melanoma detection test set, demonstrate that uncertainty is useful to the Melanoma detection task and that it is more beneficial to select the most uncertain pool samples. These results suggest that active learning could be very useful for medical imaging tasks (in particular) and more generally in low-resource settings.

1 Introduction

Active learning (AL) is generally defined as a semi-supervised machine learning (ML) algorithm whose goal is to use relatively few initial training samples in order to achieve better performance of a given model $\mathcal{M}$. The optimization of $\mathcal{M}$ is done by iteratively training it and making it learn how to choose useful new data samples to label, from a pool of unlabelled data, which will help it find better parameters and improve its overall performance on downstream tasks (e.g., prediction accuracy). The query and acquisition of new samples from the pool of unlabeled data are often done using uncertainty-based measures (5), and selecting the most uncertain samples in the pool of unlabeled data samples. Due to the fact that AL-based methods learn to *smartly* pick useful samples for their learning, this makes AL a prevalent paradigm to cope with data scarcity (which is often a bottleneck to many ML applications (e.g. in the medical where patient data is rare, sensitive, and subject to many privacy issues). The efficiency of active learning (i.e. its ability to produce better performance despite being trained on smaller training data) has been proven in many works of literature (9; 2; 15; 1; 3). In this work, we are exploring epistemic uncertainty (hereafter referred to as *uncertainty*), which refers to the uncertainty of the model in low-resource (lack of training data or availability of a very small amount of data) settings. In order to get some uncertainty score, most existing works make use of kernel-based methods on pair of images in order to capture image similarity (17; 14; 10). Conversely to these methods, in this paper, we make use of Bayesian CNNs (4) which are Convolutional Neural Networks (CNNs) (13) with prior probability distributions placed over a set of model parameters (6). In this paper, exploring the ISIC 2016 Melanoma Diagnosis dataset (7) we attempted to answer the following questions:

5th Deep Learning Indaba Conference (DLI 2023).

- is model uncertainty beneficial to the Melanoma detection task (disguised here as a binary classification)? which acquisition function works better for the given task?
- is it more efficient for medical imaging in general, and in particular for the Melanoma Detection task, to query and acquire the most uncertain samples or least uncertain samples?
- what is the effect of the size of the set of newly acquired data points, on the model's overall performance?

2 Acquisition Functions and Dataset

In this section, we describe the different acquisition functions and the dataset used.

Maximum Entropy (6) This acquisition function aims at selecting the data points which maximize the entropy of the model over each unlabelled data sample and known labels (classes). With the entropy defined as

$$\mathcal{H}[y|x, D_T] = -\sum_c p(y = c|x, D_T) log p(y = c|x, D_T)$$

where D_T is the training set, which is augmented by the set of newly acquired samples at each active learning round.

Mean Standard Deviation (6) The Mean Standard Deviation (for short MeanSTD) is the most commonly used acquisition function. It leverages the variance of the model over classes, given an input x and the parameters w of the model (11; 12). It is mathematically defined as follow:

$$\sigma_c = \sqrt{\mathcal{E}_{q(w)}[p(y = c|x, w)^2] - \mathcal{E}_{q(w)}[p(y = c|x, w)]^2}$$

$$\sigma_x = \frac{1}{C} \sum_c \sigma_c$$

As with the Maximum Entropy, in this scheme, we are also selecting points that maximize the MeanSTD.

BALD (8; 6) is based on mutual information. By definition, the mutual information denoted $\mathcal{I}$ between two random variables X, Y is telling us how much uncertainty we observe in X if we observe Y. BALD focuses on maximizing the mutual information between the predictions of the model and its posterior. BALD is mathematically defined as

$$\mathcal{I}(y, w|x, D_T) = \mathcal{H}[y|x, D_T] - \mathcal{E}_{p(w|D_T)}[\mathcal{H}[y|x, w]]$$

$$= -\sum_c p(y = c|x, D_T) log p(y = c|x, D_T)$$

$$+\mathcal{E}_{p(w|D_T)}[\sum_c p(y = c|x, w) log p(y = c|x, w)]$$

where w are the parameters of the model. In other words, BALD chooses points that are expected to maximize the information gained about the parameters of the model w (6). These points are points on which the model is uncertain on average, but about which some parameters produce disagreeing predictions with high certainty.

Dataset and Task Description The ISIC 2016 dataset (7) has been created for the ISIC 2016 challenge. Its goal was to foster the development of image analysis tools to enable the automated diagnosis of melanoma from dermoscopic images. The ISIC 2016 dataset contains 900 training images, and 350 testing images; a rather small dataset. The task is a binary classification, with the goal to detect whether a given picture is cancerous or not. The initial training dataset has been randomly split into two sets: training (containing 700 images) and evaluation (containing 200 images). The repartition of classes is unbalanced, with a clear and net dominance of negative samples (non-cancerous image samples). The images are colorful (RGB format), and we have downsampled them to the shape of (224, 224).

3 Experiments

We built the experiments using the details and hyperparameters provided in (6). Given the small size of the dataset, we started out with a small set of 100 examples made of 80 positives, and 20 negatives. Each example from all splits has been resized as stated above and normalized. The training images have additionally been augmented with *Center Cropped* and *Random Horizontal Flip* transformations. The CNN architecture is made of two 2-dimensional convolutional layers, each followed by a *relu* activation function. After the second *convolution-activation*, the result is fed to a maximum pooling layer, followed by a dropout. The result is flattened and fed to a fully-connected layer, with later on passed successively through a dropout layer, and classification head (technically another dense layer with output dimension 2). The network has been trained for 100 epochs, with a batch size of 8 and a learning rate of 1e-4. As the authors stated in the paper, we used Adam optimizer with weight decay

$$w = \frac{(1-p) * l^2}{|D_T|}$$

with $p = 0.5$ being the dropout probability, l^2 being the length scale, set to 0.5, and $|D_T|$ the length of the cumulative training dataset. At each active learning round, with a given acquisition function, we perform 20 MC-Dropout forward passes. The $top - k$, $(k = 100)$ most informative samples according to the given acquisition function, are selected, added to the training set, and deleted from the pool of unlabelled data. These hyper-parameters are kept identical across all CNN-based models for each acquisition function explored in this work.

4 Results and Discussion

Our first analysis consisted of checking the importance of uncertainty, in the context of our task description. In order to achieve that, we compared the evaluation losses and accuracies of 4 Bayesian CNNs with and without uncertainty. The results on the test set are reported in Table 1. The *normal* legend corresponds to the model without uncertainty. Table 1, and our training observations, show that uncertainty is important to have a lower and stable training loss, which helps to have better performance. The low performance of mean_std, compared to the normal Bayesian CNN, intuitively makes sense since the method is designed to maximize the variance (gaussian) of the model which could be seen as noise. This noise, coupled with the unbalanced dataset could have negatively impacted the robustness and predictive accuracy. On the other hand, *bald* performs the best: in fact, *bald* maximizes the mutual information between predictions and model posterior, which over time should make the model more accurate in predicting the right label while being robust and coping with the data unbalance. Maximum entropy (referred to as max_entropy) is also efficient and stable. In fact, theoretically, a higher entropy (since here we are maximizing it) means lower information gain: this can be considered a bit as the opposite of the goal of *bald*. Consequently, with the assumption that images from the same class share some specific features, this behavior makes sense: in a way that as the model gets more exposed to non-cancerous images during training, it is more confident about them, thus learning to select samples (images) from the minority (cancerous) class. Therefore, throughout the active learning rounds, the model gets more and more confident about samples from both classes and is more robust in performance.

Method	Testing Loss	Testing Accuracy
normal	0.01538	0.8021
bald	0.0077	**0.8047**
max_entropy	0.0075	0.7784
mean_std	**0.0072**	0.4670

Table 1: Results on the testing set for both with and without uncertainty Bayesian CNN

Method	Testing Loss	Testing Accuracy
bald	**0.009**	0.20
max_entropy	0.0099	0.5876
mean_std	0.0094	**0.8012**

Table 2: Test Results of Bayesian CNNs using the Least Uncertain Samples

Next, we leveraged the impact of the selection of the most and least uncertain samples on the final test set performance. In order to do that, we ran the same experiences (previously done by selecting

the most uncertain samples) but selected at each acquisition round the least uncertain samples. As opposed to Table 1, in Table 2 we can observe overall higher loss values. On the accuracy metric, we can see that in average, *bald* performed worse, which makes sense since technically in this setting we are choosing the points with lower mutual information. The mean_std has better performance because the points selected are the ones minimizing the variance, thus inducing less noise and encouraging better performance. The maximum_entropy kept a relatively normal balance and suggests that it is agnostic of the sampling mode (least uncertain or most uncertain samples i.e. samples respectively minimizing or maximizing the entropy). Thus far, our experiments, insights, and analyses have shown that: (a) uncertainty is beneficial to our Melanoma Detection task, (b) bald is overall the best acquisition function as the authors claimed, and (c) our additional ablation studies have also revealed that, in the context of our Melanoma Detection task, max_entropy has been proven to be agnostic of the acquisition function, offering more robustness and flexibility.

Method	Metric	Query=115	Query=100	Query=90	Query=80	Query=70	Query=60	Query=50
bald	loss	0.0177	0.0174	0.0183	**0.0169**	0.0208	0.0201	0.0185
bald	accuracy	0.8047	**0.8047**	0.7994	0.7942	0.7994	0.8021	0.8047
max_entropy	loss	0.0173	0.0235	0.0192	0.0202	**0.0157**	0.0167	0.0170
max_entropy	accuracy	0.7995	**0.8047**	0.8021	0.7863	0.8021	0.8021	0.7863
mean_std	loss	0.0185	**0.0164**	0.0191	0.0177	0.0164	0.0202	0.0186
mean_std	accuracy	0.7916	0.8021	0.8021	0.8047	0.7889	**0.8074**	0.7968

Table 3: Report of Testing Loss and Testing Accuracy on ISIC 2016 dataset as a function of the different query sizes. For each method and for each metric, the number in bold represents the best value achieved for a given query size.

Finally, we proceeded to leverage the influence of the size of the newly acquired samples (query size). The default pool size value (similarly as in (6)) used is 100. In our ablation study, we tried different additional query sizes: 115, 90, 80, 70, 60, and 50. Our previous experiments revealed that the *learning* happens mainly on the first active learning round. Therefore, we focused on the impact of the query sizes, solely in the first active learning round. The results are presented in Table 3.

In Table 3, we can notice that the scale of the loss and accuracy does not change that much. However, as far as the loss metric is concerned, we can observe that generally, all acquisition functions are impacted by the query size. On the accuracy scale, we can see that max_entropy and mean_std vary a lot compared to *bald*, which consequently offers more stability. In fact, we can speculate that BALD avoided selecting noisy points: nearby images for which there exist multiple noisy labels of different classes (points for which the aleatoric uncertainty is large) (6). Moreover, we can see that the accuracy results are very similar across query sizes and acquisition methods, on the fixed test set. This also demonstrates the difficulties with handling ML performances of ML models in extremely small data regimes. We can see that the loss values are almost similar, while most of the acquisition functions achieved their highest accuracy scores around the original query size of 100 (except for mean_std which performed better in terms of accuracy, with the second-lowest query size).

5 Conclusion and Future Works

In this work, we demonstrated how active learning could be used for a classification downstream task on the Melanoma Dataset. First of all, we showed that using uncertainty (epistemic) is useful for the Melanoma detection task. Next, we demonstrated that it is better for the model to query the most uncertain samples using the designated acquisition functions. Once that was settled, we leveraged several acquisition functions and found out that on average *bald* performs the best. These results demonstrated the viability of active learning in the context of low-resource settings[1]. However, one of our additional extensive analyses of the impact of the query size on the test set performance revealed that despite all the advantages and shortcomings of the different acquisition functions we leveraged, it is still hard to work and generalize in an extremely low data regime. As future work, we could leverage how well these acquisition functions perform on later versions (and bigger) of the ISIC Dataset. Additionally, this work could be extended to the new acquisition function *EPIG* introduced in (16). EPIG measures information gain in the space of predictions rather than parameters and leads to a better performance than BALD.

[1]Due to the requirements and format of the submission, some interesting figures have been removed from this version. Authors would gladly like to include them in the final post Indaba camera ready.

References

[1] Yukun Chen, Thomas A. Lasko, Qiaozhu Mei, Joshua C. Denny, and Hua Xu. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, 58:11–18, 2015. 1

[2] Bonaventure F. P. Dossou, Atnafu Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. 1

[3] Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online, Nov. 2020. Association for Computational Linguistics. 1

[4] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *ArXiv*, abs/1506.02158, 2015. 1

[5] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 1

[6] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017. 1, 2, 3, 4

[7] David A. Gutman, Noel C. F. Codella, M. E. Celebi, Brian Helba, Michael Armando Marchetti, Nabin K. Mishra, and Allan C. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172, 2016. 1, 2

[8] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *ArXiv*, abs/1112.5745, 2011. 2

[9] Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pages 9786–9801. PMLR, 2022. 1

[10] Ajay J. Joshi, Fatih Murat Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009. 1

[11] Michael C. Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 680–688, 2016. 2

[12] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *ArXiv*, abs/1511.02680, 2015. 2

[13] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989. 1

[14] X. Li and Yuhong Guo. Adaptive active learning for image classification. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, 2013. 1

[15] Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017. 1

[16] Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning, 2023. 4

[17] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning*, 2003. 1