A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis

Samuel Budd[a], Emma C Robinson[b,1], Bernhard Kainz[a,1]

[a]*Department of Computing, Imperial College London, UK*
[b]*Department of Imaging Sciences, King's College London, UK*

arXiv:1910.02923v1 [cs.LG] 7 Oct 2019

Abstract

Fully automatic deep learning has become the state-of-the-art technique for many tasks including image acquisition, analysis and interpretation, and for the extraction of clinically useful information for computer-aided detection, diagnosis, treatment planning, intervention and therapy. However, the unique challenges posed by medical image analysis suggest that retaining a human end-user in any deep learning enabled system will be beneficial. In this review we investigate the role that humans might play in the development and deployment of deep learning enabled diagnostic applications and focus on techniques that will retain a significant input from a human end user. Human-in-the-Loop computing is an area that we see as increasingly important in future research due to the safety-critical nature of working in the medical domain.

We evaluate four key areas that we consider vital for deep learning in the clinical practice: (1) *Active Learning* to choose the best data to annotate for optimal model performance; (2) *Interpretation and Refinement* - using iterative feedback to steer models to optima for a given prediction and offering meaningful ways to interpret and respond to predictions; (3) *Practical considerations* - developing full scale applications and the key considerations that need to be made before deployment; (4) *Related Areas* - research fields that will benefit human-in-the-loop computing as they evolve.

We offer our opinions on the most promising directions of research and how various aspects of each area might be unified towards common goals.

1. Introduction

Medical imaging is a major pillar of clinical decision making and is an integral part of many patient journeys. Information extracted from medical images is clinically useful in many areas such as computer-aided detection, diagnosis, treatment planning, intervention and therapy. While medical imaging remains a vital component of a myriad of clinical tasks, an increasing shortage of qualified radiologists to interpret complex medical images suggests a clear need for reliable automated methods to alleviate the growing burden on health-care practitioners (of Radiologists, 2017).

In parallel, medical imaging sciences are benefiting from the development of novel computational techniques for the analysis of structured data like images. Development of algorithms for image acquisition, analysis and interpretation are driving innovation, particularly in the areas of registration, reconstruction, tracking, segmentation and modelling.

Medical images are inherently difficult to interpret, requiring prior expertise to understand. Bio-medical images can be noisy and contain many modality-specific artefacts, acquired under a wide variety of acquisition conditions with different protocols. Thus, once trained models do not transfer seamlessly from one clinical task or site to another because of an often yawning domain gap (Kamnitsas et al., 2017; Ben-David et al., 2010). Supervised learning methods require extensive relabelling to regain initial performance in different workflows.

The experience and prior knowledge required to work with such data means that there is often large inter- and intra-observer variability in annotating medical data. This not only raises questions about what constitutes a gold-standard ground truth annotation, but also results in disagreement of what that ground truth truly is. These issues result in a large cost associated with annotating and re-labelling of medical image datasets, as we require numerous expert annotators (oracles) to perform each annotation and to reach a consensus.

In recent years, Deep Learning (DL) has emerged as the state-of-the-art technique for performing many medical image analysis tasks (Tizhoosh and Pantanowitz, 2018; Shen et al., 2017; Litjens et al., 2017; Suzuki, 2017). Developments in the field of computer vision have shown great promise in transferring to medical image analysis, and several techniques have been shown to perform as accurately as human observers (Haenssle et al., 2018; Mar and Soyer, 2018). However, uptake of DL methods within the clinical practice has been limited thus far, largely due to the unique challenges of working with complex medical data, regulatory compliance issues and trust in trained models.

We identify three key challenges when developing DL enabled applications for medical image analysis in a clinical set-

ting:

1. Lack of Training Data: Supervised DL techniques traditionally rely on a large and even distribution of accurately annotated data points, and while more medical image datasets are becoming available, the time, cost and effort required to annotate such datasets remains significant.
2. The Final Percent: DL techniques have achieved state-of-the-art performance for medical image analysis tasks, but in safety-critical domains even the smallest deviation can cause catastrophic results downstream. Achieving clinically credible output may require interactive interpretation of predictions (from an oracle) to be useful in practice.
3. Transparency and Interpretability: At present, most DL applications are considered to be a 'black-box' where the user has limited meaningful ways of interpreting, understanding or correcting how a model has made its prediction. Credence is a detrimental feature for medical applications as information from a wide variety of sources must be evaluated in order to make clinical decisions. Further indication of how a model has reached a predicted conclusion is needed in order to foster trust for DL enabled systems and allow users to weigh automated predictions appropriately.

There is concerted effort in the medical image analysis research community to apply DL methods to various medical image analysis tasks, and these are showing great promise. We refer the reader to a number of reviews of DL in medical imaging (Hesamian et al., 2019; Lundervold and Lundervold, 2019; Yamashita et al., 2018). These works primarily focus on the development of predictive models for a specific task and demonstrate state-of-the-art performance for that task. This review aims to give an overview of where humans will remain involved in this development, deployment and practical use of DL systems for medical image analysis. We focus on medical image segmentation techniques to explore the role of human end users in DL enabled systems. Automating segmentation tasks suffers from all of the drawbacks incurred by medical image data described above. There are many emerging techniques that seek to alleviate the added complexity of working with medical image data to perform automated segmentation of images. Segmentation seeks to divide an image into semantically meaningful regions (sets of pixels) in order to perform a number of downstream tasks, e.g. biometric measurements. Manually assigning a label to each pixel of an image is a laborious task and as such automated segmentation methods are important in practice. Advances in DL techniques such as Active Learning (AL) and Human-in-the-Loop computing applied to segmentation problems have shown progress in overcoming the key challenges outlined above and these are the studies this review focuses on. We categorise each study based on the nature of human interaction proposed and broadly divide them between which of the three key challenges they address.

Section 2 introduces Active Learning, a branch of Machine Learning (ML) and Human-in-the-Loop Computing that seeks to find the most *informative* samples from an unlabelled distribution to be annotated next. By training on the most informative subset of samples, related work can achieve state-of-the-art performance while reducing the costly annotation burden associated with annotating medical image data.

Section 3 evaluates techniques used to refine model predictions in response to user feedback, guiding models towards more accurate per-image predictions. We evaluate techniques that seek to improve interpretability of automated predictions and how models provide feedback on their own outputs to guide users towards better decision making.

Section 4 evaluates the key practical considerations of developing and deploying Human-in-the-Loop DL enabled systems in practice and outlines the work being done in these areas that addresses the three key challenges identified above. These areas are human focused and assess how human end users might interact with these systems.

Section 5 introduces related areas of ML and DL research that are having an impact on AL and Human-in-the-Loop Computing and are beginning to influence the three key challenges outlined.

In Section 6 we offer our opinions on the future directions of Human-in-the-Loop DL research and how many of the techniques evaluated might be combined to work towards common goals.

2. Active Learning

In this section we assume a scenario in which a large pool of un-annotated data U is available to us, and that we have an oracle (or group of oracles) from which we can request annotations for every un-annotated data point x_U to add to an annotated set L. We wish to train some model $f(x|L^*)$ where $L^* \subseteq L$ and consider methods that rely on annotated data to do so. A brute-force solution to this problem would be to ask the oracle(s) to annotate every x_U such that $L^* = L$, but this is rarely a practical or cost-effective solution due to the unique challenges associated with annotating biomedical image data. It is theorised that there is some L^* that achieves equivalent performance to L, i.e. $f(x|L^*) \approx f(x|L)$. A model trained on some optimal subset L^* of a dataset might achieve equivalent performance to a model trained on the entire, annotated dataset. Active Learning (AL) is the branch of machine learning that seeks to find this optimal subset L^* given a current model $f'(x|L')$, where L' is an intermediate annotated dataset, and an un-annotated dataset U. AL methods aim to iteratively seek the most informative data-points x_i^* for training a model, under the assumption that both the model and the un-annotated dataset will evolve over time, rather than selecting a fixed subset once to be used for training. In a wider context and before the advent of DL, Settles (2009) reviewed this field as a state-of-the-art ML methodology.

A typical AL framework, as outlined in Figure 1, consists of a method to evaluate the *informativeness* of each un-annotated data point x_U given $f'(x_U|L')$, tied heavily to the choice of *query type*, after which all chosen data-points are required to be annotated. Once new annotations have been acquired, the AL framework must use the new data to improve the model. This is normally done by either *retraining* the entire model using all available annotated data L', or by *fine-tuning* the network using the most recently annotated data-points x_i^*. Using this

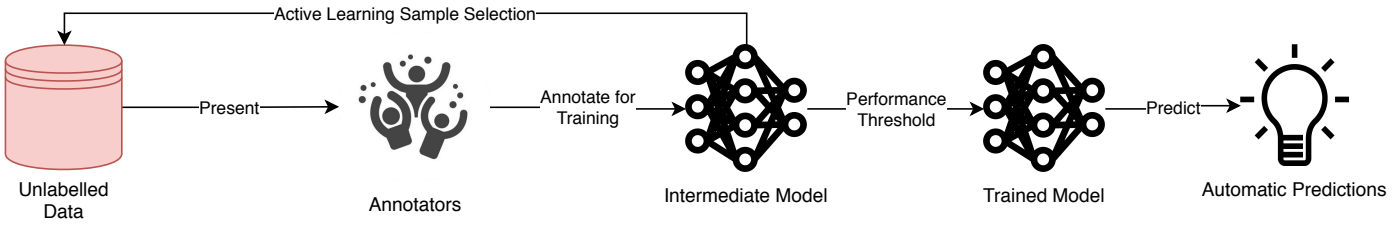

Fig. 1. Overview of Active Learning frameworks.

approach, state-of-the-art performance can be achieved using fewer annotations for several bio-medical image analysis tasks, as shown in the methods discussed in this section, thus widening the annotation bottleneck and reducing the costs associated with developing DL enabled systems from un-annotated data.

2.1. Query Types

In every AL framework the first choice to be made is what type of *query* we wish to make using a model and un-annotated dataset. There are currently three main choices available and each lends itself to a particular scenario dependant on what type of un-annotated data we have access to, and what question we wish to ask the oracle(s).

Stream-based Selective Sampling assumes a continuous stream of incoming un-annotated data-points x_U. The current model and an *informativeness* measure $I(x_U)$ are used to decide, for each incoming data-point, whether or not to ask the oracle(s) for an annotation. This query type is usually computationally inexpensive but offers limited performance benefits due to the isolated nature of each decision: the wider context of the underlying distribution is not considered, thus balancing exploration and exploitation of the distribution is less well captured than in other query types.

Membership Query Synthesis assumes that rather than drawing from a real-world distribution of data-points, we instead generate a data-point x^*_G that needs to be annotated. The generated data-point is what the current model 'believes' will be most informative to itself. This data-point is then annotated by the oracle(s). This approach may suffer from the same drawbacks as *Stream-based* methods as a model may have no knowledge of unseen areas of the distribution, and thus be unable to request annotations of those areas. Nevertheless, recent advances of *Generative Adversarial Networks* (GANs) show potential for generating data-points that mimic real-world distributions for many different types of data, including biomedical images, which we discuss in Section 5.2.

Pool-based Sampling assumes a large un-annotated real-world dataset U to draw samples from and seeks to select a batch of N samples $x^*_0, ..., x^*_N$ from the distribution to request labels for. *Pool-based* methods usually use the current model to make a prediction on each un-annotated data point to obtain a ranked measure of *informativeness* $I(x_U|f'(x_U|L'))$ for every data-point in the un-annotated set, and select the top N samples using this metric to be annotated by the oracle(s). These methods can be computationally expensive as every iteration requires a metric evaluation for every data-point in the distribution. However, these methods have shown to be the most promising when combined with DL methods, which inherently rely on a batch-based training scheme.

2.2. Evaluating Informativeness

In developing an AL framework, once a query type has been selected, the next question to ask is how to measure the informativeness $I(x_U)$ of each of the data-points? Many varying approaches have been taken to quantifying the informativeness of a sample given a model and an underlying distribution. Here we sort these metrics by the level of human interpretability they offer.

Traditionally, AL methods employ hand-designed heuristics to quantify what we as humans believe makes something informative. A variety of model specific metrics seek to quantify what the effect of using a sample for training would have on the model, e.g., the biggest change in model parameters. However, these methods are less prevalent than human designed heuristics due to the computational challenge of applying these to DL models due to the usually very high number of parameters. Finally some methods are emerging that are completely agnostic to human interpretability of informativeness and instead seek to learn the best selection policy from available data and previous iterations, as discussed in detail in Section 2.2.3.

2.2.1. Uncertainty

The main family of informativeness measure falls into calculating uncertainty. It is argued that the more *uncertain* a prediction is, the more information we can gain by including the ground truth for that sample in the training set.

There are several ways of calculating uncertainty from different ML/DL models. When considering DL for segmentation the most simple measure is the sum of lowest class probability for each pixel in a given image segmentation. It is argued that more certain predictions will have high pixel-wise class probabilities, so the lower the sum of the minimum class probability over each pixel in an image, the more certain a prediction is considered to be - this is a fairly intuitive way of thinking about uncertainty and offers a means to rank uncertainty of samples within a distribution. We refer to the method above as *least confident* sampling where the samples with the highest uncertainty are selected for labelling (Settles, 2009). A drawback of *least confident* sampling is that it only considers information about the most probable label, and discards the information about the remaining label distribution. Two alternative methods have been proposed that alleviate this concern. The first, called *margin sampling* (Settles, 2009), can be used in a multi-class setting and considers the first and second most probable labels

under the model and calculates the difference between them. The intuition here is that the larger the margin is between the two most probable labels, the more confident the model is in assigning that label. The second, more popular approach is to use *entropy* as an uncertainty measure. For binary classification, *entropy* sampling is equivalent to *least confident* and *margin* sampling, but for multi-class problems *entropy* generalises well as an uncertainty measure. Using one of the above measures, un-annotated samples are ranked and the most 'uncertain' cases are chosen for the next round of annotation.

Wang et al. (2017b) propose the Cost-Effective Active Learning (CEAL) method for deep image classification that involves *complementary sampling* in which the framework selects from an unlabelled data-set a) a set of uncertain samples to be labelled by an oracle, and b) a set of highly certain samples that are 'pseudo-labelled' by the framework and included in the labelled data-set.

Wen et al. (2018) propose an active learning method that uses uncertainty sampling to support quality control of nucleus segmentation in pathology images. Their work compares the performance improvements achieved through active learning for three different families of algorithms: Support Vector Machines (SVM), Random Forest (RF) and Convolutional Neural Networks (CNN). They show that CNNs achieve the greatest accuracy, requiring significantly fewer iterations to achieve equivalent accuracy to the SVMs and RFs.

Another common method of estimating informativeness is to measure the agreement between multiple models performing the same task. It is argued that more disagreement found between predictions on the same data point implies a higher level of uncertainty. These methods are referred to as *Query by consensus* and are generally applied when *Ensembling* is used to improve performance - i.e, training multiple models to perform the same task under slightly different parameters/settings Settles (2009). Ensembling methods have shown to measure informativeness well, but at the cost of computational resources - multiple models need to be trained and maintained, and each of these needs to be updated in the presence of newly selected training samples.

Nevertheless, Beluch Bcai et al. (2018) demonstrate the power of ensembles for active learning and compare to alternatives to ensembling. They specifically compare the performance of acquisition functions and uncertainty estimation methods for active learning with CNNs for image classification tasks and show that ensemble based uncertainties outperform other methods of uncertainty estimation such as 'MC Dropout'. They find that the difference in active learning performance can be explained by a combination of decreased model capacity and lower diversity of MC dropout ensembles. A good performance is demonstrated on a diabetic retinopathy diagnosis task.

Gal et al. (2017) introduce the use of Bayesian CNNs for Active Learning, and show that the use of Bayesian CNNs outperform deterministic CNNs in the context of Active Learning. Bayesian CNNs model uncertainty of predictions directly, and its argued this property allows them to outperform deterministic CNNs. In this work several different query strategies (or acquisition functions as they are referred to in the text) are used for Active Learning to demonstrate improved performance from fewer training samples than random sampling. They demonstrate their approach for skin cancer diagnosis from skin lesion images to show significant performance improvements over uniform sampling using the BALD method for sample selection, where the BALD method seeks to maximise the mutual information between predictions and model posterior.

Konyushkova et al. (2019) propose an active learning approach that exploits geometric smoothness priors in the image space to aid the segmentation process. They use traditional uncertainty measures to estimate which pixels should be annotated next, and introduce novel criteria for uncertainty in multi-class settings. They exploit geometric uncertainty by estimating the entropy of the probability of supervoxels belonging to a class given the predictions of its neighbours and combine these to encourage selection of uncertain regions in areas of non-smooth transition between classes. They demonstrate state-of-the-art performance on mitochondria segmentation from EM images and on an MRI tumour segmentation task for both binary and multi-class segmentations. They suggest that exploiting geometric properties of images is useful to answer the questions of where to annotate next and by reducing 3D annotations to 2D annotations provide a possible answer to how to annotate the data, and that addressing both jointly can bring additional benefits to the annotation method, however they acknowledge that it would impossible to design bespoke selection strategies this way for every new task at hand.

2.2.2. Representativeness

Many AL frameworks extend selection strategies to include some measure of *representativeness* in addition to an uncertainty measure. The intuition behind including a representativeness measure is that methods only concerned with *uncertainty* have the potential to focus only on small regions of the distribution, and that training on samples from the same area of the distribution will introduce redundancy to the selection strategy, or may skew the model towards a particular area of the distribution. The addition of a representativeness measure seeks to encourage selection strategies to sample from different areas of the distribution, thus improving AL performance. A sample with a high representativeness covers the information for many images in the same area of the distribution, so there is less need to include many samples covered by a representative image.

To this end, Yang et al. (2017) present Suggestive Annotation, a deep active learning framework for medical image segmentation, which uses an alternative formulation of uncertainty sampling combined with a form of representativeness density weighting. Their method consists of training multiple models that each exclude a portion of the training data, which are used to calculate an ensemble based uncertainty measure. They formulate choosing the most representative example as a generalised version of the maximum set-cover problem (NP Hard) and offer a greedy approach to selecting the most representative images using feature vectors from their models. They demonstrate state-of-the-art performance using 50% of the available data on the MICCAI Gland segmentation challenge and a lymph node segmentation task.

Smailagic et al. (2018) propose *MedAL*, an active learning framework for medical image segmentation. They propose a sampling method that combines uncertainty, and distance between feature descriptors, to extract the most informative samples from an unlabelled data-set. Another contribution of this work is an approach which generates an initial training set by leveraging existing computer vision image descriptors to find the images that are most dissimilar to each other and thus cover a larger area of the image distribution. They show good results on three different medical image analysis tasks, achieving the baseline accuracy with less training data than random or pure uncertainty based methods.

Ozdemir et al. (2018) propose a Borda-count based combination of an uncertainty and a representativeness measure to select the next batch of samples. Uncertainty is measured as the voxel-wise variance of N predictions using MC dropout in their model. They introduce new representativeness measures such as 'Content Distance', defined as the mean squared error between layer activation responses of a pre-trained classification network. They extend this contribution by encoding representativeness by maximum entropy to optimise network weights using an novel entropy loss function.

Sourati et al. (2018) propose a novel method for ensuring diversity among queried samples by calculating the Fisher Information (FI), for the first time in CNNs. Here, efficient computation is enabled by the gradient computations of propagation to allow FI to be calculated on the large parameter space of CNNs. They demonstrate the performance of their approach on two different flavours of task: a) semi-automatic segmentation of a particular subject (from a different group/different pathology not present in the original training data) where iteratively labelling small numbers of voxels queried by AL achieves accurate segmentation for that subject; and b) using AL to build a model generalisable to all images in a given data-set. They show that in both these scenarios the FI-based AL improves performance after labelling a small percentage of voxels, outperformed random sampling and achieved higher accuracy than than entropy based querying.

2.2.3. Learning Active Learning

The methods discussed so far are all hand designed heuristics of informativeness, but some works have emerged that attempt to learn what the most informative samples are through experience of previous sample selection outcomes. This offers a potential way to select samples more efficiently but at the cost of interpretability of the heuristics employed. Many factors influence the performance and optimality of using hand-crafted heuristics for data selection. Konyushkova et al. (2017) propose 'Learning Active Learning', where a regression model learns data selection strategies based on experience from previous AL outcomes. Arguing there is no way to foresee the influence of all factors such as class imbalance, label noise, outliers and distribution shape. Instead, their regression model 'adapts' its selection to the problem without explicitly stating specific rules. Bachman et al. (2017) take this idea a step further and propose a model that leverages labelled instances from different but related tasks to learn a selection strategy, while simultaneously adapting its representation of the data and its prediction function.

2.3. Fine-tuning vs Retraining

The final step of each AL framework is to use newly acquired annotations to improve a model. Two main approaches are used to train a model on new annotations. These are retraining the model using all available data including the newly acquired annotations or to fine-tune the model using only new annotations or the new annotations plus a subset from the existing annotations.

Tajbakhsh et al. (2016) investigate using transfer learning and fine-tuning in several medical image analysis tasks and demonstrate that the use of a pre-trained CNN with fine-tuning outperformed a CNN trained from scratch and that these fine-tuned CNNs were more robust to the size of the training sets. They also showed that neither shallow nor deep tuning was the optimal choice for a particular application and present a layer-wise training scheme that could offer a practical way to reach optimal performance for the chosen task based on the amount of data available. The methods employed in this work perform one-time fine-tuning where a pre-trained model is fine-tuned just once with available training samples, however this does not accommodate an active selection process or continuous fine-tuning.

Zhou et al. (2017) propose a continuous fine-tuning method that fine-tunes a pre-trained CNN with successively larger datasets and demonstrate that this approach converges faster than repeatedly fine-tuning the pre-trained CNN. They also find that continuously fine-tuning with only newly acquired annotations requires careful meta-parameter adjustments making it less practical across many different tasks.

Retraining is computationally more expensive than fine-tuning but it provides a consistent means to evaluate AL framework performance. Fine-tuning is used across a number of different ML areas such as one or few shot learning, and transfer learning and the best approach to this is still an open question and as such is less prevalent in AL frameworks, as fine tuning improves we may see a shift towards its use in AL frameworks. It is important to establish baseline fine-tuning and retraining schemes to effectively compare the DL/AL methods in which they are applied in order to isolate the effects of these schemes from the improvements made in other areas.

3. Interpretability and Refinement

So far we have considered the role of humans in annotating data to be used to train a model, but once a model is trained, we still require a human-in-the-loop to interpret model predictions and potentially to refine them to acquire the most accurate results for unseen data as outlined in Figure 2.

3.1. Interpretability

While DL methods have become a standard state-of-the-art approach for many medical image analysis tasks, they largely remain black-box methods where the end user has limited

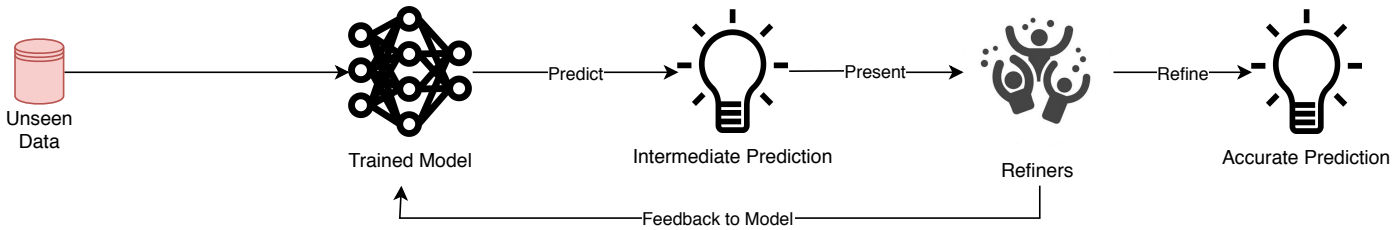

Fig. 2. Overview of Refinement frameworks.

meaningful ways of interpreting model predictions. This feature of DL methods is a significant hurdle in the deployment of DL enabled applications to safety-critical domains such as medical image analysis. We want models to be highly accurate and robust, but also explainable and interpretable.

Recent EU law[1] has led to the 'right for explanation', whereby any subject has the right to have automated decisions that have been made about them explained. This even further highlights the need for transparent algorithms which we can reason about [Goodman and Flaxman (2016), Edwards and Veale (2017a), Edwards and Veale (2017b)].

It is important for users to understand how a certain decision has been made by the model, as even the most accurate and robust models aren't infallible, and false or uncertain predictions must be identified so that trust in the model can be fostered and predictions are appropriately weighted in the clinical decision making process. It is vital the end user, regulators and auditors all have the ability to contextualise automated decisions produced by DL models. Here we outline some different methods for providing interpretable ways of reasoning about DL models and their predictions.

Typically DL methods can provide statistical metrics on the uncertainty of a model output, many of the uncertainty measures discussed in Section 2 are also used to aid in intepretability. While uncertainty measures are important, these are not sufficient to foster complete trust in DL model, the model should provide human-understandable justifications for its output that allow insights to be drawn elucidating the inner workings of a model. Chakraborty et al. (2017) discuss many of the core concerns surrounding model intepretability and highlight various works that have demonstrated more sophisticated methods of making a DL model interpretable across the DL field. Here we evaluate some of the works that have been applied to medical image segmentation and refer the reader to [Stoyanov et al. (2018), Holzinger et al. (2017)] for further reading on interpretability in the rest of the medical imaging domain.

Oktay et al. (2018) introduce 'Attention Gating' to guide networks towards giving more 'attention' to certain image areas, in a visually interpretable way - potentially aiding in the subsequent refinement of annotations.

Ng et al. (2018) explore different uncertainty estimates for a U-Net based cardiac MRI segmentation in order to detect inaccurate segmentations, as having the ability to know when a segmentation is less accurate can be useful to reduce down stream errors, and demonstrate that by setting a threshold on the quality of segmentations we can remove poor segmentations for manual correction.

In Budd et al. (2019) we propose a visual method for interpreting automated head circumference measurements from ultrasound images, using MC Dropout at test-time to acquire N head segmentations to calculate an upper and lower bound on the head circumference measurement in real-time. These bounds were displayed over the image to guide the sonographer towards views in which the model predicts with the most confidence. This upper lower bound is presented as a measure of model compliance of the unseen image rather than uncertainty. Finally, variance heuristics are proposed to quantify the confidence of a prediction in order to either accept or reject head circumference measurements, and it is shown these can improve overall performance measures once 'rejected' images are removed.

Wang et al. (2019b) propose using test-time augmentation to acquire a measure of aleatoric (image-based) uncertainty and compare their method with epistemic (model) uncertainty measures and show that their method provides a better uncertainty estimation than a test-time dropout based model uncertainty alone and reduces overconfident incorrect predictions.

Paschali et al. (2019) propose a novel interpretation method for histological Whole Slide image processing by combing a deep neural network with a Multiple instance Learning branch to enhance the models expressive power without guiding its attention. A logit heat-map of model activations is presented, in order to interpret its decision making process. Two expert pathologists provided feedback that the interpretability of the method has potential for integration into several clinical applications.

Jungo and Reyes (2019) evaluate several different voxel-wise uncertainty estimation methods applied to medical image segmentation with respect to their reliability and limitations and show that current uncertainty estimation methods perform similarly. Their results show that while uncertainty estimates may be well calibrated at the dataset level (capturing epistemic uncertainty), they tend to be mis-calibrated at a subject-level (aleatoric uncertainty). This compromises the reliability of these uncertainty estimates and highlights the need to develop subject-wise uncertainty estimates. They show auxiliary networks to be a valid alternative to common uncertainty methods

[1] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1

as they can be applied to any previously trained segmentation model.

Developing transparent systems will enable faster uptake in clinical practice and including humans within the deep learning clinical pipelines will ease the period of transition between current best practices and the breadth of possible enhancements that deep learning has to offer.

We suggest that ongoing work in improving interpretability of DL models will also have a positive impact on AL, as the majority of methods to improve intepretability are centred on providing uncertainty measures for a models prediction, these same uncertainty measures can be used for AL selection strategies in place of existing uncertainty measures that are currently employed. As intepretability and uncertainty measures improve we expect to see a similar improvement of AL frameworks as they incorporate the most promising uncertainty measures.

3.2. Refinement

If we can develop accurate, robust and interpretable models for medical image applications we still cannot guarantee clinical grade accuracy for every unseen data-point presented to a model. The ability to generalise to unseen input is a cornerstone of deep learning applications, but in real world distributions, generalisation is rarely perfect. As such, methods to rectify these discrepancies must be built into applications used for medical image analysis. This iterative refinement must save the end user time and mental effort over performing manual annotation. Many interactive image segmentation systems have been proposed, and more recently these have built on the advances in deep learning to allow users to refine model outputs and feedback the more accurate results to the model for improvement.

Amrehn et al. (2017) introduced UI-Net, that builds on the popular U-Net architecture for medical image segmentation Ronneberger et al. (2015). The UI-Net is trained with an *active user model*, and allows for users to interact with proposed segmentations by providing *scribbles* over the image to indicate areas that should be included or not, the network is trained using simulated user interactions and as such responds to iterative user scribbles to refine a segmentation towards a more accurate result.

Conditional Random fields have been used in various tasks to encourage segmentation homogeneity. Zheng et al. (2015) propose CRF-CNN, a recurrent neural network which has the desirable properties of both CNNs and CRFs. Wang et al. (2017a) propose DeepIGeoS, an interactive geodesic framework for medical image segmentation. This framework uses two CNNs, the first performs an initial automatic segmentation, and the second takes the initial segmentation as well as user interactions with the initial segmentation to provide a refined result. They combine user interactions with CNNs through geodesic distance transforms Criminisi et al. (2008), and these user interactions are integrated as hard constraints into a Conditional Random Field, inspired by Zheng et al. (2015). They call their two networks P-Net (initial segmentation) and R-Net (for refinement). They demonstrate superior results for segmentation of the placenta from 2D fetal MRI and brain tumors from 3D FLAIR images when compared to fully automatic CNNs. These segmentation results were also obtained in roughly a third of the time taken to perform the same segmentation with traditional interactive methods such as GeoS or ITK-SNAP.

Graph Cuts have also been used in segmentation to incorporate user interaction - a user provides *seed points* to the algorithm (e.g. mark some pixel as foreground, and another as background) and from this the segmentation is calculated. Wang et al. (2018) propose BIFSeg, an interactive segmentation framework inspired by graph cuts. Their work introduces a deep learning framework for interactive segmentation by combining CNNs with a bounding box and scribble based segmentation pipeline. The user provides a bounding box around the area which they are interested in segmenting, this is then fed into their CNN to produce an initial segmentation prediction, the user can then provide scribbles to mark areas of the image as mis-classified - these user inputs are then weighted heavily in the calculation of the refined segmentation using their graph cut based algorithm.

Bredell et al. (2018) propose an alternative to BIFSeg in which two networks are trained, one to perform an initial segmentation (they use a CNN but this initial segmentation could be performed with any existing algorithm) and a second network they call interCNN that takes as input the image, some user scribbles and the initial segmentation prediction and outputs a refined segmentation, they show that with several iterations over multiple user inputs the quality of the segmentations improve over the initial segmentation and achieve state-of-the-art performance in comparison to other interactive methods.

The methods discussed above have so far been concerned with producing segmentations for individual images or slices, however many segmentation tasks seek to extract the 3D shape/surface of a particular region of interest (ROI). Kurzendorfer et al. (2017) propose a dual method for producing segmentations in 3D based on a Smart-brush 2D segmentation that the user guides towards a good 2D segmentation, and after a few slices are segmented this is transformed to a 3D surface shape using Hermite radial basis functions, achieving high accuracy. While this method does not use deep learning it is a strong example of the ways in which interactive segmentation can be used to generate high quality training data for use in deep learning applications - their approach is general and can produce segmentations for a large number of tasks. There is potential to incorporate deep learning into their pipeline to improve results and accelerate the interactive annotation process.

Jang and Kim (2019) propose an interactive segmentation scheme that generalises to any previously trained segmentation model, which accepts user annotations about a target object and the background. User annotations are converted into interaction maps by measuring the distance of each pixel to the annotated landmarks, after which the forward pass outputs an initial segmentation. The user annotated points can be mis-segmented in the initial segmentation so they propose BRS (back-propogating refinement scheme) that corrects the mis-labelled pixels. They demonstrate that their algorithm outperforms conventional approaches on several datasets and that BRS can generalise to medical image segmentation tasks by transforming existing CNNs into user-interactive versions.

In this section we focus on applications concerned with iteratively refining a segmentation towards a desired quality of output. In the scenarios above this is performed on an un-seen image provided by the end user, but there is no reason the same approach could be taken to generate iteratively more accurate annotations to be used in training, e.g., using active learning to select which samples to annotate next, and iteratively refining the prediction made by the current model until a sufficiently accurate annotation is curated. This has the potential to accelerate annotation for training without any additional implementation overhead. Much work done in AL ignores the role of the oracle and merely assumes we can acquire an accurate label when we need it, but in practice this presents a more significant challenge. We foresee AL and HITL computing become more tightly coupled as AL research improves it's consideration for the oracle providing the annotations.

4. Practical Considerations

We have so far discussed the core body of work behind AL, model interpretation and prediction refinement, and while the works discussed above go a long way in covering the majority of research being done, there are several practical considerations for developing and deploying DL enabled applications that we must consider. In this section we outline the main practical areas that still require research to fully understand their impact on the DL enabled application development pipeline and suggest where we might look next.

4.1. Noisy Oracles

Gold-standard annotations for medical image data are acquired by aggregating annotations from multiple expert oracles, but as previously discussed, this is rarely feasible to obtain for large complex datasets due to the expertise required to perform such annotations. Here we ask what effect on performance we might incur if we acquire labels from oracles without domain expertise, and what techniques can we use to mitigate the suspected degradation of annotation quality when using non-expert oracles, to avoid any potential loss in accuracy.

Li et al. (2015) and Zhang and Chaudhuri (2015) propose active learning methods that assume data will be annotated by a crowd of non-expert or 'weak' annotators, and offer approaches to mitigate the introduction of bad labels into the data set. They simultaneously learn about the quality of individual annotators so that the most informative examples can be labelled by the strongest annotators.

Cheplygina et al. (2017) explore using Amazon's MTurk to gather annotations of airways in CT images. Results showed that the novice oracles were able to interpret the images, but that instructions provided were too complex, leading to many unusable annotations. Once the bad annotations were removed, the annotations did show medium to high correlation with expert annotations, especially if annotations were aggregated.

Rodrigues and Pereira (2017) describe an approach to assess the reliability of annotators in a crowd, and a crowd layer used to train deep models from noisy labels from multiple annotators, internally capturing the reliability and biases of different annotators to achieve state-of-the-art results for several crowd-sourced data-set tasks.

We can see that by using a learned model of oracle annotation quality we can mitigate the effects of low quality annotations and present the most challenging cases to most capable oracles. By providing clear instructions we can lower the barriers for non-expert oracles to perform accurate annotation, but this is not generalisable and would be required for every new annotation task we wish to perform.

4.2. Alternative Query Types

Most segmentation tasks require pixel-wise annotations, but these are not the only type of annotation we can give an image. Segmentation can be performed with 'weak' annotations, which include image level labels e.g. modality, organs present etc. and annotations such as bounding boxes, ellipses or scribbles. It is argued that using 'weaker' annotation formulations can make the task easier for the human oracle, leading to more accurate annotations. 'Weak' annotations have been shown to perform well in several segmentation tasks, Rajchl et al. (2016a) demonstrate obtaining pixel-wise segmentations given a data-set of images with 'weak' bounding box annotations. They propose DeepCut, an architecture that combines a CNN with an iterative dense CRF formulation to achieve good accuracy while greatly reducing annotation effort required. In a later study, Rajchl et al. (2017) examine the impact of expertise required for different 'weak' annotation types on the accuracy of liver segmentations. The results showed a decrease in accuracy with less expertise, as expected, across all annotation types. Despite this, segmentation accuracy was comparable to state-of-the-art performance when using a weakly labelled atlas for outlier correction. The robust performance of their approach suggests 'weak' annotations from non-expert crowds could be used to obtain accurate segmentations on many different tasks, however their use of an atlas makes this approach less generalisable than is desired.

In Rajchl et al. (2016b) they examine using super pixels to accelerate the annotation process. This approach uses a pre-processing step to acquire a super-pixel segmentation of each image, non-experts are then used to perform the annotation by selecting which super-pixels are part of the target region. Results showed that the approach largely reduces the annotation load on users. Non-expert annotation of 5000 slices was completed in under an hour by 12 annotators, compared to an expert taking three working days to establish the same with an advanced interface. The non-expert interface is web-based demonstrating the potential of distributed annotation collection/crowd-sourcing. An encouraging aspect of this paper is that the results showed high performance on the segmentation task in question compared with expert annotation performance, but may not be suitable for all medical image analysis tasks.

It has been shown that we can develop high performing models using weakly annotated data, and as weak annotations requires less expertise to perform, they can be acquired faster and from a non-expert crowd with a smaller loss in accuracy than gold-standard annotations. This is very promising for future research as datasets of weakly annotated data might be much easier and more cost-effective to curate.

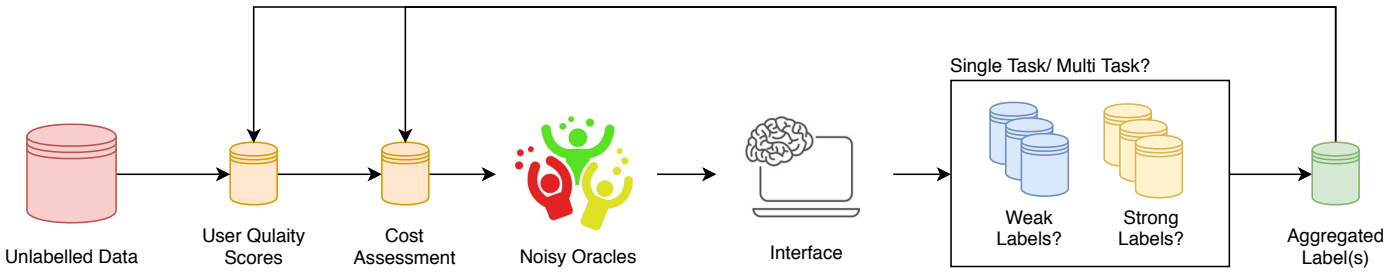

Fig. 3. Overview of Practical Considerations

4.3. Multi-task learning

Many works aim to train models or acquire training data for several tasks at once, it is argued that this can save on cost as complementary information may result in higher performance over multiple different tasks (Moeskops et al., 2016). Wang et al. (2019a) propose a dual network for joint segmentation and detection task for lung nodule segmentation and cochlea segmentation from CT images, where only a part of the data is densely annotated and the rest is weakly labelled by bounding boxes, using this they show that their architecture out-performs several baselines. At present this work only handles the case for two different label types but they propose extending the framework for a true multi-task scenario.

This is a promising area but, as of yet, it has not been incorporated into an active learning setting. As such, it may be elucidating to analyse the differences in samples chosen by different AL methods when the model is being training for multiple tasks simultaneously. However, Lowell et al. (2018) raise concerns over the transferability of actively acquired datasets to future models due to the inherent coupling between active learning selection strategies and the model being trained, and show that training a successor model on the actively acquired dataset can often result in worse performance than from random sampling. They suggest that, as datasets begin to outlive the models trained on them, there is a concern for the efficacy of active learning, since the acquired dataset may be disadvantageous for training subsequent models. An exploration of how actively acquired datasets perform on multiple models may be required to explain the effects of an actively acquired dataset coupled with one model on the performance of related models.

4.4. Annotation Interface

So far the majority of Human-in-the-loop methods assume a significant level of interaction from an oracle to annotate data and model predictions, but few consider the nature of the interface with which an oracle might interact with these images. The nature of medical images require special attention when proposing distributed online platforms to perform such annotations. While the majority of techniques discussed so far have used pre-existing data labels in place of newly acquired to demonstrate their performance, it is important to consider the effects of accuracy of annotation that the actual interface might incur.

Nalisnik et al. (2015) propose a framework for the online classification of Whole-slide images (WSIs) of tissues. Their interface enables users to rapidly build classifiers using an active learning process that minimises labelling efforts and demonstrates the effectiveness of their solution for the quantification of glioma brain tumours.

Khosravan et al. (2017) propose a novel interface for the segmentation of images that tracks the users gaze to initiate seed points for the segmentation of the object of interest as the only means of interaction with the image, achieving high segmentation performance. Stember et al. (2019) extend this idea and compare using eye tracking generated training samples to traditional hand annotated training samples for training a DL model. They show that almost equivalent performance was achieved using annotation generated through eye tracking, and suggest that this approach might be applicable to rapidly generate training data. They acknowledge that there is still improvements to be made integrate eye tracking into typical clinical radiology work flow in a faster, more natural and less distracting way.

Tinati et al. (2017) evaluate the player motivations behind EyeWire, an online game that asks a crowd of players to help segment neurons in a mouse brain. The gamification of this task has seen over 500,000 players sign up and the segmentations acquired have gone onto be used in several research works [Kim et al. (2014)]. One of the most exciting things about gamification is that when surveyed, users were motivated most by making a scientific contribution rather than any potential monetary reward. However this is very specialised towards this particular task and would be difficult to apply across other types of medical image analysis task.

There are many different approaches to developing annotation interfaces and the ones we consider above are just a few that have been applied to medical image analysis. As development increases we expect to see more online tools being used for medical image analysis and the chosen format of the interface will play a large part in the usability and overall success of these applications.

4.5. Variable Learning Costs

When acquiring training data from various types of oracle it is worth considering the relative cost associated with querying a particular oracle type for that annotation. We may wish to acquire more accurate labels from an expert oracle, but this is likely more expensive to obtain than from a non-expert oracle. The trade off, of course, being accuracy of the obtained label - less expertise of the oracle will likely result in a lower quality of annotation. Several methods have been proposed to model

this and allow developers to trade off between cost and overall accuracy of acquired annotations.

Kuo et al. (2018) propose a cost-sensitive active learning approach for intracranial haemorrhage detection. Since annotation time may vary significantly across examples, they model the annotation time and optimize the return on investment. They show their approach selects a diverse and meaningful set of samples to be annotated, relative to a uniform cost model, which mostly selects samples with massive bleeds which are time consuming to annotate.

Shah et al. (2018) propose a budget based cost minimisation framework in a mixed-supervision setting (strong and weak annotations) via dense segmentation, bounding boxes, and landmarks. Their framework uses an uncertainty and a representativeness ranking strategy to select samples to be annotated next. They demonstrate state-of-the-art performance at a significantly reduced training budget, highlighting the important role of choice of annotation type on the costs of acquiring training data.

The above works each show an improved consideration for the economic burden that is incurred when curating training data. A valuable research direction would be to assess the effects of oracle expertise level, annotation type and image annotation cost in a unified framework as these three factors are very much linked and may have a profound influence over each other.

5. Related Areas

Several related areas of ML/DL research show potential benefits for developing accurate and robust models from limited training data, and have also been applied in AL scenarios for medical image analysis. Here we outline where research done in these related areas has also aimed at solving one or more our the three main challenges we outline in this review.

5.0.1. Semi-supervised Learning

In the presence of large data-sets, but the absence of labels, unsupervised and semi-supervised approaches offer a means by which information can be extracted without requiring labels for all the data-points. This could potentially have a massive impact on the medical image analysis field where this is often the case.

In a semi-supervised learning (SSL) scenario we may have some labelled data, but this is often very limited. We do however have a large set of un-annotated instances (much like in active learning) to draw information from, the goal being to improve a model (trained only on the labelled instances) using the un-labelled instances. From this we derive two distinct goals: a) predicting labels for future data (inductive SSL) and b) predicting labels for the available un-annotated data (transductive SSL) Cheplygina et al. (2018).

An popular family of SSL approaches employ a technique called *self-training* where a classifier is trained using only the labelled data, following training, inference is performed on the unlabelled instances. A decision is made about each of the new annotations as to whether they should be included in the training set in the next iteration. One proposed approach to making this decision is the use of an oracle to decide if the label is accurate enough for use during training, guiding this towards an active learning approach. Self-training is popular in many segmentation tasks, but less so for detection and diagnosis applications Cheplygina et al. (2018).

It has been shown that increasing the number of samples improves performance, but that the advantages of SSL methods decrease as we acquire more labelled data. SSL methods provide a powerful way of extracting useful information from unannotated image data and we believe that progress in this area will be beneficial to AL systems that desire a more accurate model for initialisation to guide data selection strategies.

5.1. Reinforcement Learning

Reinforcement learning (RL) is a branch of ML that enables an 'agent' to learn in an interactive environment, by trial and error, using feedback from its own actions and experiences, working towards achieving the defined goal of the system.

Woodward et al. (2017) propose a one-shot learning method that combines with RL to allow the model to decide, during inference, which examples are worth labelling. A stream of images is presented and a decision is made either to predict the label, or pay to receive the the correct label. Through the choice of RL reward function they are able to achieve higher prediction accuracy than a purely supervised task, or trade prediction accuracy for fewer label requests.

Fang et al. re-frame the data selection process as a RL problem, and explicitly learn a data selection policy. This is agnostic to the data selection heuristics common in AL frameworks, providing a more general approach, demonstrating improvements in entity recognition, however this is yet to be applied to medical image data.

Milletari et al. (2019) propose the application of RL to ultrasound care, guiding a potentially inexperienced user to the correct sonic window and enabling them to obtain clinically relevant images of the anatomy of interest. This human-in-the-loop application is an example of the novel applications possible when combining DL/RL with real-time systems enabling users to respond to model feedback to acquire the most accurate information available.

RL methods offer a different approach to AL and Human-in-the-Loop problems that is well aligned with aiding real-time feedback between a DL enabled application and its end users, however it requires task specific goals that may not be generalisable across different medical image analysis tasks.

5.2. Generative Adversarial Networks

Generative Adversarial Network (GAN) based methods have been applied to several areas of medical imaging such as denoising, modality transfer and abnormality detection, but more relevant to AL has been the use of GANs for image synthesis, this offers an alternative (or addition) to the many data augmentation techniques used to expand limited data-sets Yi et al. (2015).

Last et al. (2018) propose a conditional GAN (cGAN) based method for active learning where they use the discriminator D

output as a measure of uncertainty of the proposed segmentations, and use this metric to rank samples from the unlabelled data-set. From this ranking the most uncertain samples are presented to an oracle for segmentation and the least uncertain images are included in the labelled data-set as *pseudo ground truth* labels. They show their method approaches increasing accuracy as the percentage of interactively annotated samples increases - reaching the performance of fully supervised benchmark methods using only 80% of the labels. This work also motivates the use of GAN discriminator scores as a measure of prediction uncertainty.

Mahapatra et al. (2018) also use a cGAN to generate chest X-Ray images conditioned on a real image, and using a Bayesian neural network to assess the informativeness of each generated sample, decide whether each generated sample should be used as training data. If so, is used to fine-tune the network. They demonstrate that the approach can achieve comparable performance to training on the fully annotated data, using a dataset where only 33% of the pixels in the training set are annotated, offering a huge saving of time, effort and costs for annotators.

Zhao et al. (2019) present an alternative method of data synthesis to GANs through the use of learned transformations. From a single manually segmented image, they leverage other un-annotated images in a SSL like approach to learn a transformation model from the images, and use the model along with the labelled data to synthesise additional annotated samples. Transformations consist of spatial deformations and intensity changes to enable to synthesis of complex effects such as anatomical and image acquisition variations. They train a model in a supervised way for the segmentation of MRI brain images and show state-of-the-art improvements over other one-shot bio-medical image segmentation methods.

The above works demonstrate the power of using synthetic data conditioned on a very small amount of annotated data to generate new training samples that can be used to train a model to a high accuracy, this is of great value to AL methods where we usually require a initial training set to train a model on before we can employ a data selection policy. These methods also demonstrate the efficient use of labelled data and allow us to generate multiple training samples from a individually annotated image, this may allow the annotated data obtained in AL/Human-in-the-Loop methods to be used more effectively through generating multiple training samples for a single requested annotation, further reducing the annotation effort required to train state-of-the-art models.

5.3. Transfer Learning

Transfer Learning (TL) and domain adaptation are branches of DL that aim to use pre-trained networks as a starting point for new applications. Given a pre-trained network trained for a particular task, it has been shown that this network can be 'fine-tuned' towards a target task from limited training data. Tajbakhsh et al. (2016) demonstrated the applicability of TL for a variety of medical image analysis tasks, and show, despite the large differences between natural images and medical images, CNNs pre-trained on natural images and fine-tuned on medical images can perform better than medical CNNs trained from scratch. This performance boost was greater where fewer target task training examples were available. Many of the methods discussed so far start with a network pre-trained on natural image data.

Zhou et al. (2018b) propose AFT*, a platform that combines AL and TL to reduce annotation efforts, which aims at solving several problems within AL. AFT* starts with a completely empty labelled data-set, requiring no seed samples. A pre-trained CNN is used to seek 'worthy' samples for annotation and to gradually enhance the CNN via continuous fine-tuning. A number of steps are taken to minimise the risk of catastrophic forgetting. Their previous work Zhou et al. (2017) applies a similar but less featureful approach to several medical image analysis tasks to demonstrate equivalent performance can be reached with a heavily reduced training data-set. They then use these tasks to evaluate several patterns of prediction that the network exhibits and how these relate to the choice of AL selection criteria.

Zhou et al. (2018a) have gone onto to use their AFT framework for annotation of CIMT videos, a clinical technique for characterisation of Cardiovascular disease. Their extension into the video domain presents its own unique challenges and thus they propose a new concept of an Annotation Unit - reducing annotating a CIMT video to just 6 user mouse clicks, and by combining this with their AFT framework reduce annotation cost by 80% relative to training from scratch and by 50% relative to random selection of new samples to be annotated (and used for fine-tuning).

Kushibar et al. (2019) use TL for supervised domain adaptation for sub-cortical brain structure segmentation with minimal user interaction. They significantly reduce the number of training images from different MRI imaging domains by leveraging a pre-trained network and improve training speed by reducing the number of trainable parameters in the CNN. They show their method achieves similar results to their baseline while using a remarkably small amount of images from the target domain and show that using even one image from the target domain was enough to outperform their baseline.

The above methods and more discussed in this review demonstrate the applicability of TL to reducing the number of annotated sample required to train a model on a new task from limited training data. By using pre-trained networks trained on annotated natural image data (there is an abundance) we can boost model performance and further reduce the annotation effort required to achieve state-of-the-art performance.

5.4. Continual Lifelong Learning and Catastrophic Forgetting

In many of scenarios described in this review, models continuously receive new annotations to be used for training, and in theory we could continue to retrain or fine-tune a model indefinitely, but is this practical and cost effective? It is important to quantify the long term effects of training a model with new data to assess how the model changes over time and whether or not performance has improved, or worse, declined. Learning from continuous streams of data has proven more difficult than anticipated, often resulting in 'catastrophic forgetting' or 'interference' Parisi et al. (2019). We face the *stability-plasticity-dilemma*. Avoiding catastrophic forgetting in neural networks

when learning from continuous streams of data can be broadly divided among three conceptual strategies: a) Retraining the the whole network while regularising (to prevent forgetting of previously learned tasks). b) selectively train the network and expand it if needed to represent new tasks, and c) retaining previous experience to use memory replay to learn in the absence of new input. We refer the reader to Parisi et al. (2019) for a more detailed overview of these approaches.

Baweja et al. (2018) investigate continual learning of two MRI segmentation tasks with neural networks for countering catastrophic forgetting of the first task when a new one is learned. They investigate elastic weight consolidation, a method based on Fisher information to sequentially learn segmentation of normal brain structures and then segmentation of white matter lesions and demonstrate this method reduces catastrophic forgetting, but acknowledge there is a large room for improvement for the challenging setting of continual learning.

It is important to quantify the performance and robustness of a model at every stage of its lifespan. One way to consider stopping could evaluate when the cost of continued training outweighs the cost of errors made by the current model. An existing measure that attempts to quantify the economical value of medical intervention is the Quality-adjusted Life year (QALY), where one QALY equates to one year of healthy life NICE (2013). Could this metric be incorporated into models? At present we cannot quantify the cost of errors made by DL medical imaging applications but doing so could lead to a deeper understanding of how accurate a DL model really ought to be.

As models are trained on more of the end users own data, will this cause the network to perform better on data from that users system despite performing worse on data the model was initially trained on? Catastrophic forgetting suggests this will be the case, but is this a bad thing? It may be beneficial for models to gradually bias themselves towards high performance for the end users own data, even if this results in the model becoming less transferable to other data.

6. Future Prospective and Unanswered Questions

In Sections 2 & 3 we discuss methods through which a user might gather training data to build a model, use their model to predict on new data and receive feedback to iteratively refine the model output towards a more accurate result. Each of these techniques assume some human end user will be present to interact with the system at the point of initial annotation, interpretation and refinement. Each of these areas seeks to achieve a shared goal of achieving the highest performing model from as little annotated data as possible - with a means to weigh conclusions of models predictions appropriately.

Active learning assumes the presence of a user interface to perform annotations but is only concerned with which data to annotate. Refinement assumes we can generate an annotation through iterative interaction with the current model prediction. Hence, it would be desirable to combine these two in future work. If we can train a model with a tiny amount of training data, and then ask annotators to refine model predictions towards a more accurate label, we can expedite the annotation process by reducing the initial annotation workload and reduce additional interface work for use with unseen data. This would be the same interface used to create the training annotations. By combining the efforts of active learning and iterative refinement into a unified framework we can rapidly produce annotations to train our model, as well as acquiring high quality results from our models from the beginning. This should also have the added side effect of training the model on data from the same distribution that it will be predicting on, reducing domain shift effects in unseen distributions.

By incorporating our end user at each stage of the model life cycle we could also use human feedback on model performance to add a more 'human interpretable' metric of model confidence as each user could rank the performance of the model for each input as it sees it, potentially giving a metric of confidence based on human interpretation of the model output. This of course requires experts to be using the system.One might argue that the models initial predictions may impart some influence over the human user but by crowd-sourcing the initial annotations to a less expert multi-label crowd we could reduce this bias.

Developments in uncertainty quantification will benefit both AL selection heuristics and interpretation of model outputs, but there is no guarantee that the best performing uncertainty metrics for selecting new samples to be annotated will be the same metrics that are the most interpretable to a human user.

Figure 4 outlines the core methods being used in human-in-the-loop computing for each of the papers discussed in this review. This figure shows that there is significant overlap of research goals for many areas of human-in-the-loop computing but there are large gaps that need to be filled in order to understand the relationships between different methods and how these might affect their performance.

As the many areas of DL research converge towards shared goals of working with limited training data to achieve state-of-the-art results, we expect to see more systems emerge that exploit the advances made in the range of sub-fields of ML described here. We have already seen the combination of several methods into individual frameworks but as of yet no works combine all of the approaches discussed into a single framework. As different combinations of approaches begin to appear it is important to consider the measure by which we assess their performance, as isolating individual developments becomes more difficult. Developing baseline human-in-the-loop methods to compare to will be vital to assess the contributions of individual works in each area and to better understand the influences of competing improvements in these areas.

7. Conclusions

In this review we have explored the large body of emerging medical image analysis work in which a human end user is at the centre. Deep Learning has all the ingredients to induce a paradigm shift in our approach to a myriad of clinical tasks. The direct involvement of humans is set to play a core role in this shift. The works presented in this review each offer their own approaches to including humans in the loop and we suggest that there is sufficient overlap in many methods for them to

be considered under the same title of Human-in-the-Loop computing. We hope to see new methodologies emerge that combine the strengths of AL and HITL computing into end-to-end systems for the development of deep learning applications that can be used in clinical practice. While there are some practical limitations as discussed, there are many proposed solutions to such issues and as research in these directions continues it is only a matter of time before deep learning applications blossom into fully-fledged, accurate and robust systems to be used for daily routine tasks. We are in an exciting era for medical image analysis, with endless opportunity to innovate and improve the current state-of-the-art and to leverage the powers of deep learning to make a real impact in health care across the board. With diligent research and development we should see more and more applications boosted by deep learning capabilities finding their way onto the market, allowing users to achieve better results, faster, and with less expertise than before, freeing up expert time to be used on the most challenging cases. The field of Human-in-the-loop computing will play a crucial role to achieve this.

Acknowledgments

SB is supported by the EPSRC Centre for Doctoral Training in Smart Medical Imaging EP/S022104/1. This work was in part supported by EP/S013687/1, Intel and Nvidia. We thank The Wellcome Trust IEH Award iFind project [102431] and Innovate UK: London Medical Imaging & Artificial Intelligence Centre for Value-Based Healthcare [104691] for funding this research.

References

Amrehn, M., Gaube, S., Unberath, M., Schebesch, F., Horz, T., Strumia, M., Steidl, S., Kowarschik, M., Maier, A., 2017. UI-Net: Interactive Artificial Neural Networks for Iterative Image Segmentation Based on a User Model. Technical Report. URL: https://arxiv.org/pdf/1709.03450.pdf.

Bachman, P., Sordoni, A., Trischler, A., 2017. Learning Algorithms for Active Learning. Technical Report. URL: http://proceedings.mlr.press/v70/bachman17a/bachman17a.pdf.

Baweja, C., Glocker, B., Kamnitsas, K., 2018. Towards continual learning in medical imaging. Technical Report. URL: https://www.doc.ic.ac.uk/~bglocker/public/mednips2018/med-nips_2018_paper_82.pdf.

Beluch Bcai, W.H., Nürnberger, A., Bcai, J.M.K., 2018. The power of ensembles for active learning in image classification. Technical Report. URL: http://openaccess.thecvf.com/content_cvpr_2018/CameraReady/1487.pdf.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Machine learning 79, 151–175.

Bredell, G., Tanner, C., Konukoglu, E., 2018. Iterative Interaction Training for Segmentation Editing Networks. Technical Report. URL: https://arxiv.org/pdf/1807.08555.pdf.

Budd, S., Sinclair, M., Khanal, B., Matthew, J., Lloyd, D., Gomez, A., Toussaint, N., Robinson, E., Kainz, B., 2019. Confident Head Circumference Measurement from Ultrasound with Real-time Feedback for Sonographers URL: http://arxiv.org/abs/1908.02582.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R.M., Kelley, T.D., Braines, D., Sensoy, M., Willis, C.J., Gurram, P., 2017. Interpretability of deep learning models: A survey of results, in: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE. pp. 1–6. URL: https://ieeexplore.ieee.org/document/8397411/, doi:10.1109/UIC-ATC.2017.8397411.

Cheplygina, V., De Bruijne, M., Pluim, J.P.W., 2018. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Technical Report. URL: https://arxiv.org/pdf/1804.06353.pdf.

Cheplygina, V., Perez-Rovira, A., Kuo, W., Tiddens, H.A.W.M., De Bruijne, M., 2017. Early Experiences with Crowdsourcing Airway Annotations in Chest CT URL: https://arxiv.org/pdf/1706.02055v1.pdf.

Criminisi, A., Sharp, T., Blake, A., 2008. GeoS: Geodesic Image Segmentation, Springer, Berlin, Heidelberg, pp. 99–112. URL: http://link.springer.com/10.1007/978-3-540-88682-2_9, doi:10.1007/978-3-540-88682-2{_}9.

Edwards, L., Veale, M., 2017a. Enslaving the Algorithm: From a Right to an Explanationn to a Right to Better Decisionss? SSRN Electronic Journal URL: https://www.ssrn.com/abstract=3052831, doi:10.2139/ssrn.3052831.

Edwards, L., Veale, M., 2017b. Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for. SSRN Electronic Journal URL: https://www.ssrn.com/abstract=2972855, doi:10.2139/ssrn.2972855.

Fang, M., Li, Y., Cohn, T., . Learning how to Active Learn: A Deep Reinforcement Learning Approach URL: https://arxiv.org/pdf/1708.02383v1.pdf.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep Bayesian Active Learning with Image Data URL: http://arxiv.org/abs/1703.02910.

Goodman, B., Flaxman, S., 2016. European Union regulations on algorithmic decision-making and a "right to explanation" URL: http://arxiv.org/abs/1606.08813http://dx.doi.org/10.1609/aimag.v38i3.2741, doi:10.1609/aimag.v38i3.2741.

Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., Uhlmann, L., Alt, C., Arenbergerova, M., Bakos, R., Baltzer, A., Bertlich, I., Blum, A., Bokor-Billmann, T., Bowling, J., Braghiroli, N., Braun, R., Buder-Bakhaya, K., Buhl, T., Cabo, H., Cabrijan, L., Cevic, N., Classen, A., Deltgen, D., Fink, C., Georgieva, I., Hakim-Meibodi, L.E., Hanner, S., Hartmann, F., Hartmann, J., Haus, G., Hoxha, E., Karls, R., Koga, H., Kreusch, J., Lallas, A., Majenka, P., Marghoob, A., Massone, C., Mekokishvili, L., Mestel, D., Meyer, V., Neuberger, A., Nielsen, K., Oliviero, M., Pampena, R., Paoli, J., Pawlik, E., Rao, B., Rendon, A., Russo, T., Sadek, A., Samhaber, K., Schneiderbauer, R., Schweizer, A., Toberer, F., Trennheuser, L., Vlahova, L., Wald, A., Winkler, J., Wölbing, P., Zalaudek, I., 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology 29, 1836–1842. URL: https://academic.oup.com/annonc/article/29/8/1836/5004443, doi:10.1093/annonc/mdy166.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. Journal of Digital Imaging 32, 582–596. URL: http://link.springer.com/10.1007/s10278-019-00227-x, doi:10.1007/s10278-019-00227-x.

Holzinger, A., Malle, B., Kieseberg, P., Roth, P.M., Müller, H., Reihs, R., Zatloukal, K., 2017. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology URL: http://arxiv.org/abs/1712.06657.

Jang, W.D., Kim, C.S., 2019. Interactive image segmentation via backpropagating refinement scheme, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition.

Jungo, A., Reyes, M., 2019. Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation. Technical Report. URL: https://github.com/alainjungo/reliability-challenges-uncertainty.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: International conference on information processing in medical imaging, Springer. pp. 597–609.

Khosravan, N., Celik, H., Turkbey, B., Cheng, R., McCreedy, E., McAuliffe, M., Bednarova, S., Jones, E., Chen, X., Choyke, P., Wood, B., Bagci, U., 2017. Gaze2Segment: A Pilot Study for Integrating Eye-Tracking Tech-

nology into Medical Image Segmentation, Springer, Cham, pp. 94–104. URL: http://link.springer.com/10.1007/978-3-319-61188-4_9, doi:10.1007/978-3-319-61188-4{_}9.

Kim, J.S., Greene, M.J., Zlateski, A., Lee, K., Richardson, M., Turaga, S.C., Purcaro, M., Balkam, M., Robinson, A., Behabadi, B.F., Campos, M., Denk, W., Seung, H.S., EyeWirers, t., 2014. Spacetime wiring specificity supports direction selectivity in the retina. Nature 509, 331–336. URL: http://www.nature.com/articles/nature13240, doi:10.1038/nature13240.

Konyushkova, K., Sznitman, R., Fua, P., 2017. Learning Active Learning from Data. URL: https://papers.nips.cc/paper/7010-learning-active-learning-from-data.

Konyushkova, K., Sznitman, R., Fua, P., 2019. Geometry in active learning for binary and multi-class image segmentation. Computer Vision and Image Understanding 182, 1–16. URL: https://www.sciencedirect.com/science/article/pii/S107731421930013X, doi:10.1016/J.CVIU.2019.01.007.

Kuo, W., Häne, C., Yuh, E., Mukherjee, P., Malik, J., 2018. Cost-Sensitive Active Learning for Intracranial Hemorrhage Detection, Springer, Cham, pp. 715–723. URL: http://link.springer.com/10.1007/978-3-030-00931-1_82, doi:10.1007/978-3-030-00931-1{_}82.

Kurzendorfer, T., Fischer, P., Mirshahzadeh, N., Pohl, T., Brost, A., Steidl, S., Maier, A., 2017. Rapid Interactive and Intuitive Segmentation of 3D Medical Images Using Radial Basis Function Interpolation . Annual Conference on Medical Image Understanding and Analysis , 11–13URL: www.mdpi.com/journal/jimaging, doi:10.3390/jimaging3040056.

Kushibar, K., Valverde, S., González-Villà, S., Bernal, J., Cabezas, M., Oliver, A., Lladó, X., 2019. Supervised Domain Adaptation for Automatic Sub-cortical Brain Structure Segmentation with Minimal User Interaction. Scientific Reports 9, 6742. URL: http://www.nature.com/articles/s41598-019-43299-z, doi:10.1038/s41598-019-43299-z.

Last, F., Klein, T., Ravanbakhsh, M., Nabi, M., Batmanghelich, K., Tresp, V., 2018. Human-Machine Collaboration for Medical Image Segmentation. Technical Report. URL: https://pdfs.semanticscholar.org/4e0c/535386e3a3d307cee45e97b9417eff4da92e.pdf.

Li, S.Y., Jiang, Y., Zhou, Z.H., 2015. Multi-Label Active Learning from Crowds URL: http://arxiv.org/abs/1508.00722.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135, doi:10.1016/J.MEDIA.2017.07.005.

Lowell, D., Lipton, Z.C., Wallace, B.C., 2018. Practical Obstacles to Deploying Active Learning. Technical Report. URL: https://arxiv.org/pdf/1807.04801.pdf.

Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik 29, 102–127. URL: https://www.sciencedirect.com/science/article/pii/S0939388918301181, doi:10.1016/J.ZEMEDI.2018.11.002.

Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Reyes, M., 2018. Efficient Active Learning for Image Classification and Segmentation Using a Sample Selection and Conditional Generative Adversarial Network, Springer, Cham, pp. 580–588. URL: http://link.springer.com/10.1007/978-3-030-00934-2_65, doi:10.1007/978-3-030-00934-2{_}65.

Mar, V.J., Soyer, H.P., 2018. Artificial intelligence for melanoma diagnosis: how can we deliver on the promise? Annals of Oncology 29, 1625–1628. URL: https://academic.oup.com/annonc/article/29/8/1625/5004449, doi:10.1093/annonc/mdy193.

Milletari, F., Birodkar, V., Sofka, M., 2019. Straight to the point: reinforcement learning for user guidance in ultrasound URL: http://arxiv.org/abs/1903.00586.

Moeskops, P., Wolterink, J.M., van der Velden, B.H.M., Gilhuijs, K.G.A., Leiner, T., Viergever, M.A., Išgum, I., 2016. Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities, Springer, Cham, pp. 478–486. URL: http://link.springer.com/10.1007/978-3-319-46723-8_55, doi:10.1007/978-3-319-46723-8{_}55.

Nalisnik, M., Gutman, D.A., Kong, J., Cooper, L.A., 2015. An Interactive Learning Framework for Scalable Classification of Pathology Images. Proceedings : ... IEEE International Conference on Big Data. IEEE International Conference on Big Data 2015, 928–935. URL: http://www.ncbi.nlm.nih.gov/pubmed/27796014http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5082843, doi:10.1109/BigData.2015.7363841.

Ng, M., Guo, F., Biswas, L., Wright, G.A., 2018. Estimating Uncertainty in Neural Networks for Segmentation Quality Control. Technical Report. URL: https://www.doc.ic.ac.uk/~bglocker/public/mednips2018/med-nips_2018_paper_105.pdf.

NICE, 2013. Judging whether public health interventions offer value for money — Guidance and guidelines — NICE URL: https://www.nice.org.uk/advice/lgb10.

Oktay, O., Schlemper, J., Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention u-net: Learning where to look for the pancreas URL: https://arxiv.org/pdf/1804.03999.pdf.

Ozdemir, F., Peng, Z., Tanner, C., Fuernstahl, P., Goksel, O., 2018. Active Learning for Segmentation by Optimizing Content Information for Maximal Entropy, Springer, Cham, pp. 183–191. URL: http://link.springer.com/10.1007/978-3-030-00889-5_21, doi:10.1007/978-3-030-00889-5{_}21.

Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S., 2019. Continual Lifelong Learning with Neural Networks: A Review. Technical Report. URL: https://arxiv.org/pdf/1802.07569v2.pdf.

Paschali, M., Naeem, M.F., Simson, W., Steiger, K., Mollenhauer, M., Navab, N., 2019. Deep Learning Under the Microscope: Improving the Interpretability of Medical Imaging Neural Networks URL: http://arxiv.org/abs/1904.03127.

of Radiologists, T.R.C., 2017. Clinical radiology UK workforce census 2017 report. Technical Report. URL: https://www.rcr.ac.uk/system/files/publication/field_publication_files/bfcr185_cr_census_2017.pdf.

Rajchl, M., Koch, L.M., Ledig, C., Passerat-Palmbach, J., Misawa, K., Mori, K., Rueckert, D., 2017. Employing Weak Annotations for Medical Image Analysis Problems URL: https://arxiv.org/pdf/1708.06297v1.pdf.

Rajchl, M., Lee, M.C.H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., Rueckert, D., 2016a. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Transactions on Medical Imaging 36, 674–683.

Rajchl, M., Lee, M.C.H., Schrans, F., Davidson, A., Passerat-Palmbach, J., Tarroni, G., Alansary, A., Oktay, O., Kainz, B., Rueckert, D., 2016b. Learning under Distributed Weak Supervision URL: https://arxiv.org/pdf/1606.01100v1.pdf.

Rodrigues, F., Pereira, F.C., 2017. Deep Learning from Crowds, in: AAAI. URL: https://arxiv.org/pdf/1709.01779v2.pdf.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, Springer, Cham, pp. 234–241. URL: http://link.springer.com/10.1007/978-3-319-24574-4_28, doi:10.1007/978-3-319-24574-4{_}28.

Settles, B., 2009. Active learning literature survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

Shah, M.P., Bhalgat, Y.S., Awate, S.P., 2018. Annotation-cost Minimization for Medical Image Segmentation using Suggestive Mixed Supervision Fully Convolutional Networks. Technical Report. URL: https://www.doc.ic.ac.uk/~bglocker/public/mednips2018/med-nips_2018_paper_30.pdf.

Shen, D., Wu, G., Suk, H.I., 2017. Deep Learning in Medical Image Analysis. Annual review of biomedical engineering 19, 221–248. URL: http://www.ncbi.nlm.nih.gov/pubmed/28301734http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5479722, doi:10.1146/annurev-bioeng-071516-044442.

Smailagic, A., Noh, H.Y., Costa, P., Walawalkar, D., Khandelwal, K., Mirshekari, M., Fagert, J., Galdran, A., Xu, S., 2018. MedAL: Deep Active Learning Sampling Method for Medical Image Analysis. undefined URL: https://www.semanticscholar.org/paper/MedAL%3A-Deep-Active-Learning-Sampling-Method-for-Smailagic-Noh/fa23dc7a8b3927953d83f5ce46e0b622b7cac456.

Sourati, J., Gholipour, A., Dy, J.G., Kurugol, S., Warfield, S.K., 2018. Active Deep Learning with Fisher Information for Patch-Wise Semantic Segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S.... volume 11045, pp. 83–91. URL: http://www.ncbi.nlm.nih.gov/pubmed/

30450490http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6235453http://link.springer.com/10.1007/978-3-030-00889-5_10, doi:10.1007/978-3-030-00889-5{_}10.

Stember, J.N., Celik, H., Krupinski, E., Chang, P.D., Mutasa, S., Wood, B.J., Lignelli, A., Moonis, G., Schwartz, L.H., Jambawalikar, S., Bagci, U., 2019. Eye Tracking for Deep Learning Segmentation Using Convolutional Neural Networks. Journal of Digital Imaging 32, 597–604. URL: http://link.springer.com/10.1007/s10278-019-00220-4, doi:10.1007/s10278-019-00220-4.

Stoyanov, D., Taylor, Z., Kia, S.M., Oguz, I., Reyes, M., Martel, A., Maier-Hein, L., Marquand, A.F., Duchesnay, E., Löfstedt, T., Landman, B., Cardoso, M.J., Silva, C.A., Pereira, S., Meier, R. (Eds.), 2018. Understanding and Interpreting Machine Learning in Medical Image Computing Applications. volume 11038 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham. URL: http://link.springer.com/10.1007/978-3-030-02628-8, doi:10.1007/978-3-030-02628-8.

Suzuki, K., 2017. Overview of deep learning in medical imaging. Radiological Physics and Technology 10, 257–273. URL: http://www.ncbi.nlm.nih.gov/pubmed/28689314http://link.springer.com/10.1007/s12194-017-0406-5, doi:10.1007/s12194-017-0406-5.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging 35, 1299–1312. URL: http://ieeexplore.ieee.org/document/7426826/, doi:10.1109/TMI.2016.2535302.

Tinati, R., Luczak-Roesch, M., Simperl, E., Hall, W., 2017. An investigation of player motivations in Eyewire, a gamified citizen science project. Computers in Human Behavior 73, 527–540. URL: https://ac.els-cdn.com/S0747563216309037/1-s2.0-S0747563216309037-main.pdf?_tid=abaecc23-8276-4fab-ad47-300db95a3b56&acdnat=1523979569_9e91cf49a50416e7c8772bfb73614a6b, doi:10.1016/j.chb.2016.12.074.

Tizhoosh, H.R., Pantanowitz, L., 2018. Artificial Intelligence and Digital Pathology: Challenges and Opportunities. Journal of pathology informatics 9, 38. URL: http://www.ncbi.nlm.nih.gov/pubmed/30607305, doi:10.4103/jpi.jpi{_}53{_}18.

Wang, D., Li, M., Ben-Shlomo, N., Corrales, C.E., Cheng, Y., Zhang, T., Jayender, J., 2019a. Mixed-Supervised Dual-Network for Medical Image Segmentation URL: http://arxiv.org/abs/1907.10209.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019b. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45. doi:10.1016/J.NEUCOM.2019.01.103.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T., 2018. Interactive Medical Image Segmentation using Deep Learning with Image-specific Fine-tuning. Technical Report. URL: https://arxiv.org/pdf/1710.04043.pdf.

Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ebastien Ourselin, S., Vercauteren, T., 2017a. DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation. Technical Report. URL: https://arxiv.org/pdf/1707.00652.pdf.

Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L., 2017b. Cost-Effective Active Learning for Deep Image Classification. IEEE Transactions on Circuits and Systems for Video Technology 27, 2591–2600. doi:10.1109/TCSVT.2016.2589879.

Wen, S., Kurc, T.M., Hou, L., Saltz, J.H., Gupta, R.R., Batiste, R., Zhao, T., Nguyen, V., Samaras, D., Zhu, W., 2018. Comparison of Different Classifiers with Active Learning to Support Quality Control in Nucleus Segmentation in Pathology Images. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2017, 227–236. URL: http://www.ncbi.nlm.nih.gov/pubmed/29888078.

Woodward, M., Finn, C., Research, B.A., 2017. Active One-shot Learning. Technical Report. URL: https://arxiv.org/pdf/1702.06559.pdf.

Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K., 2018. Convolutional neural networks: an overview and application in radiology. Insights into Imaging 9, 611–629. URL: https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9, doi:10.1007/s13244-018-0639-9.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation, Springer, Cham, pp. 399–407. URL: http://link.springer.com/10.1007/978-3-319-66179-7_46, doi:10.1007/978-3-319-66179-7{_}46.

Yi, X., Walia, E., Babyn, P., 2015. Generative Adversarial Network in Medical Imaging: A Review. Technical Report 8. URL: https://arxiv.org/pdf/1809.07294.pdf.

Zhang, C., Chaudhuri, K., 2015. Active Learning from Weak and Strong Labelers URL: http://arxiv.org/abs/1510.02847.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V., 2019. Data augmentation using learned transformations for one-shot medical image segmentation URL: http://arxiv.org/abs/1902.09383.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S., 2015. Conditional Random Fields as Recurrent Neural Networks. Technical Report.

Zhou, Z., Shin, J., Feng, R., Hurst, R.T., Kendall, C.B., Liang, J., 2018a. Integrating Active Learning and Transfer Learning for Carotid Intima-Media Thickness Video Interpretation. Journal of Digital Imaging URL: http://www.ncbi.nlm.nih.gov/pubmed/30402668http://link.springer.com/10.1007/s10278-018-0143-2, doi:10.1007/s10278-018-0143-2.

Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., Liang, J., 2017. Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally. Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2017, 4761. URL: http://www.ncbi.nlm.nih.gov/pubmed/30337799http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6191179, doi:10.1109/CVPR.2017.506.

Zhou, Z., Shin, J.Y., Gurudu, S.R., Gotway, M.B., Liang, J., 2018b. AFT * : Integrating Active Learning and Transfer Learning to Reduce Annotation Efforts URL: https://arxiv.org/pdf/1802.00912v1.pdf.

Supplementary Material

		Active Learning				Interpretation				Refinement			Related					Practical				
Author	**Year**	Uncertainty	Representativeness	Learning AL	Bespoke	Uncertainty	Bayesian	Visual	Bespoke	Interactive	Task Prior	Bespoke	Transfer	SSL	Reinforcement	Generative	Continual	User Model	Weak Labels	Multi-task	Cost sensitive	Interface
Wang et al	2017	✓												✓							✓	
Wen et al	2018	✓				✓																
Beluch et al	2018	✓																				
Gal et al	2017	✓			✓	✓	✓															
Konyushkova et al	2019	✓			✓															✓		✓
Yang et al	2017	✓	✓																			
Smailagic et al	2018	✓	✓		✓																	
Ozdemir et al	2018	✓	✓		✓																	
Sourati et al	2018		✓		✓					✓									✓			
Konyushkova et al	2017			✓	✓																	
Bachman et al	2017			✓	✓															✓		
Tajbakhsh et al	2016												✓				✓					
Oktay et al	2018								✓													
Ng et al	2018					✓																
Budd et al	2019					✓	✓	✓	✓	✓												✓
Wang et al	2019					✓		✓	✓													
Paschali et al	2019							✓	✓													✓
Jungo et al	2019					✓																
Amrehn et al	2017									✓								✓	✓			✓
Zheng et al	2015									✓	✓	✓										✓
Wang et al	2017									✓	✓	✓							✓			✓
Wang et al	2017									✓	✓	✓					✓		✓			✓
Bredell et al	2018									✓		✓							✓			✓
Jang et al	2019									✓	✓	✓										✓
Li et al	2015	✓			✓													✓				✓
Zhang et al	2015	✓			✓													✓			✓	✓
Cheplygina et al	2017																					✓
Rodrigues et al	2017																	✓				
Rajchl et al	2017										✓							✓	✓			✓
Moeskops et al	2016																			✓		
Wang et al	2019																		✓	✓		
Nalisnik et al	2015	✓	✓																✓			✓
Khosravan et al	2017																		✓			✓
Stember et al	2019																		✓			✓
Tinati et al	2017																		✓			✓
Kuo et al	2018	✓	✓																		✓	
Shah et al	2018	✓	✓																✓		✓	
Woodward et al	2017			✓											✓						✓	
Fang et al	2017			✓											✓							
Milletari et al	2019									✓		✓			✓							✓
Last et al	2018	✓			✓									✓		✓						
Mahapatra et al	2018	✓	✓			✓	✓									✓						
Zhao et al	2019													✓		✓						
Tajbakhsh et al	2016												✓									
Zhou et al	2017	✓	✓		✓								✓									
Zhou et al	2018	✓	✓		✓								✓				✓		✓			✓
Baweja et al	2018																✓			✓		

Fig. 4. Table of features demonstrated by work discussed in this review