Inducing natural invariance in deep segmentation pipelines for generalizable detection of colorectal polyps

Any short subtitle

Birk Sebastian Frostelid Torpmann-Hagen

Thesis submitted for the degree of
Master in Robotics and Intelligent Systems
60 credits

Department of Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2022

Inducing natural invariance in deep segmentation pipelines for generalizable detection of colorectal polyps

Any short subtitle

Birk Sebastian Frostelid Torpmann-Hagen

Inducing natural invariance in deep segmentation pipelines for generalizable detection of colorectal polyps

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

Contents

List of Figures

List of Tables

Preface

Chapter 1

Introduction

Colorectal cancer is one of the leading causes of cancer related deaths, causing approximately 900 thousand deaths worldwide per year (cite). Early detection thereof is as a consequence of significant importance. Polyps are often an early warning-sign of developing tumor, and early detection thereof can as a result significantly reduce fatality rates. Polyps are, however, often missed during colonoscopies, owing to the significant variability in the shapes an sizes of polyps, as well a the high degrees of similarity to surrounding tissue. Hence, automatic segment polyps via deep learning has the potential to significantly increase the likelihood of early detection and effective treatment.

However, clinical applications of deep learning are known to fail in deployment, despite exhibiting excellent performance during development. This is known as generalization failure, and is ubiquitous in the domain. While there has been a growing body of research dedicated to identifying and analyzing its root causes, there is still limited research into approaches to mitigating generalizability failure.

This thesis presents a novel approach to increasing generalizability, whereby the model is trained to not only minimize segmentation-loss, but also minimize the effects of the data being perturbed by an ensemble of transformations, including color-transformations, geometric transformations, additive noise, and adding extra polyps to the image using a GAN-inpainter. This endows the pipeline with the ability to more readily infer causally viable inductive biases by explicitly forcing the model to be robust to any combination of the aforementioned transformations.

Generalizability is then measured by evaluating several vanilla-pipelines consisting of several models on a number of separate datasets, which itsthen compared toroot causes, the results of the modified piresults show that (...)

Chapter 2

Background

2.1 Colorectal Polyps, Medical Imaging, and Deep Learning

Polyps are small growths found in and around the inner lining of the large intestine. These polyps, also referred to as adenomas, can in time develop into cancerous tumours, or carcinomas, in a process known as the adenoma-carcinoma sequence [13]. Though the majority of polyps do not undergo this process, identifying polyps nonetheless constitutes an important step towards preventing colorectal cancer. Indeed, resection of these polyps has been shown to reduce the incidence of colorectal cancer by a significant margin [17].

Though colorectal cancer remains as one of the leading causes of cancer-related death worldwide (source), mortality rates have nonetheless declined in large part to increased use of screening colonoscopy, which in turn has facilitated the use of more preemptive treatment. Polyps are, however, by nature somewhat difficult to detect and are routinely missed by clinicians, with miss rates ranging upwards of 27% for diminutive (<2.5mm) polyps [10, 14].

Reducing this miss rate has the potential to further reduce incidence rates. As a result, there has been a significant body of work dedicated to developing systems and techniques to aid in optimizing and effectivizing the screening procedure. One such example, referred to as chromoendoscopy, has been shown to reduce miss rates by (...) merely by employing the use of specific dyes prior to the colonoscopy. Similarly, the use of narrow-band imaging techniques, wherein light of specific wavelengths specifically designed to highlight the textural differences between the polyps and the surrounding tissue, has been shown to reduce miss rates by (...)

These systems do, however, require more equipment, training and expertise to effectively employ. Thus, automatic polyp segmentation using deep learning and convolutional neural networks (CNNs) has been identified as a possible diminutive detection method. This requires minimal training time on the part of the clinician, no additional equipment, and has been show to significantly increase detection rates when deployed in a clinical setting [3].

This has spurred on a large body of research dedicated to improving on the performance and expanding the capabilities of deep-learning based systems for polyp detection and segmentation. Several challenges been also held, namely the Endotect challenge [11], EndoCV2020 [2], EndoCV2021 [1], and more.

There are, however, still several hurdles to overcome; recent research has shown that even state of the art deep-learning pipelines are prone to generalisation failure when deployed in practical settings, particularly when exposed to distributional shifts such as changes in demographics, imaging equipment, noise, and more despite exhibiting high performance on hold-out sets [4, 6, 9, 18]. As a result, the EndoCV2021 challange emloyed training data from several centers, with the data from one of the centers being hidden and used as generalisation test data. The results from this challange demomstrated the pervasiveness of generalisation failure, with every submitted model exhibiting significant performance reductions when evaluated on their hidden dataset (cite summary here).

Naturally, automatic segmentation systems are rendered practically useless should they fail to perform sufficiently outside of the very carefully controlled conditions upon which they are trained. Consequently, for any such system to have any practical merit, it has to have the capacity to infer causally reasonable patterns in the data that generalize well to other hospitals, demographics, imaging equipment, resolutions, and so on. Though a human would not find this type of generalisation very

2.2 Generalisation failure in broader contexts

2.2.1 Generalisation failure in Medical Imaging

Generalisation failure is not, of course, unique to the gastrointestinal domain. Indeed, though medical imaging has in recent years proven to be one of the most promising applications of artificial intelligence and deep learning, having the capacity to significantly improve both the accuracy and efficiency of detection, diagnosis, and treatment of a wide variety of diseases [16], they are nonetheless highly prone to generalisation failure. In addition to the already limited capabilities of deep neural networks to generalise, medical domains are subject to a number of other exacerbating factors that make generalisation all the more diffcult. Training data is often scarce, the pathologies that constitute the classification targets are unevenly distributed and often exhibit high degrees of inter-class and within-class variability. Moreover, due to the sheer scope of the data involved, there are inevitably a significant number of confounding variables both during training and in deployment.

For instance, a deep-learning based classifier which successfully detected pneumonia in X-ray scans across a number of hospitals with striking accuracy was determined to be basing its predictions not on any lesions or otherwise pathologically relevant features in the images, but rather on a hospital-specific metal token that was on every image, which it used in conjunction with learning the prevalence rate of pneumonia for the hospitals from which the data was collected. As a result, when deployed on data from hospitals that it had not seen during training, the system failed to generalize [18].

In another study, it was shown that a classifier intended to detect diabetic retinopathy exhibited significant variability in performance depending on the type of camera used. The same study also showed that the same type of performance variability could be found when detecting skin-conditions across demographics with differing skin tones. [6].

2.2.2 Generalisation failure in other domains

Naturally, non-medical domains are in no way immune to generalisation failure. In fact, one could easily argue that the vast majority of deep-learning pipelines fail to generalize altogether, and instead merely infer some set of inductive biases that, although perhaps causally incorrect, perform sufficiently well for general use. It has for instance been shown that CNNs trained on imagenet, one of the largest and most diverse datasets in the domain of computer vision, are heavily biased towards textural features[8]. Naturally, this is not necessarily causally accurate; a cat is not a cat because it has cat-like fur; nor is an elephant an elephant only because it has skin of an elephant. By manually increasing shape bias, it has been shown that the performance of such CNNs improves both in robustness to perturbations and iid accuracy.

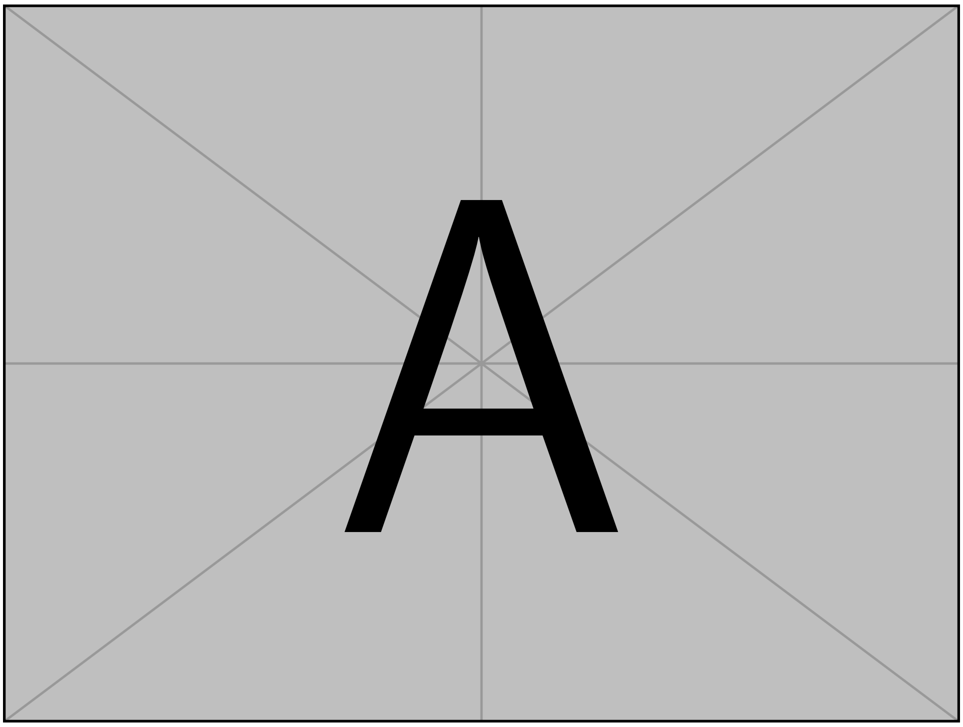

Figure 2.1:

Another characteristic of deep learning that supports this argument is the effectiveness of adversarial attacks [12], which specifically target weaknesses in the inductive biases within DNNs through any number of means in an attempt to induce high rates of incorrect, yet highly confident predictions. Gradient-based adversarial attacks, for instance, use the gradients of the model to break even the most sophisticated and well-trained pipelines merely by adding some carefully crafted, yet visually imperceptible noise to the inputs [5]. Even without access to the gradients, there exists a multitude of so-called black-box attacks that only use output samples to generate similarly effective attacks (cite). Finally, it has been shown that adding minor visual distractions to objects, for example adding bits of tape or graffiti to stop signs, dramatically increases misclassification rates [7].

Even benign, but nonetheless confounding perturbations also have the potential to induce failure. It has for instance been shown that sophisticated natural language processing models can and readily do fail if one adds peripheral information to the input. (Example, citation)

2.3 Generalisability Theory

Exactly why and how DNNs seem to so persistently fail to generalize is a topic of ongoing research, and the available literature seems to suggest that the problem is multifaceted. This section is an attempt to summarize and distill the findings and analysis performed in the field. It will cover the theoretical basis of generalisation and why one might expect DNNs to generalize, discuss the key characteristics of generalisation failure and their origins, and finally introduce a probabilistic perspective of generalisation.

2.3.1 Generalisation through Empirical Risk Minimization

Naturally, deep learning would not have experienced as much of a revolution in the last decade or so if there was not some semblance of an expectation that their striking performance was generalisable and performant also outside the idealized settings typically involved in research. The theoretical basis that informs this belief in (most) modern deep learning pipelines is the idea of so-called empirical risk minimization, wherein it is assumed that the dataset upon which the model is trained is a representative sample of the distribution of all possible samples in the relevant domain. In other words, it assumes that the dataset is independently and identically distributed (iid) to the domain distribution. To better understand this assumption, it is beneficial to consider the it from first principles:

At the most fundamental level, the goal of machine learning is to learn a mapping between two spaces of objects X and Y. This mapping, namely the function $f : X \rightarrow Y$, maps some input object $x \in X$, an image for example, to a corresponding and application-relevant output object $y \in Y$, for instance a segmentation mask or a class probabilities. It is worth noting, however, that f is not as much a function in the mathematical sense as much as it is an abstraction of whatever ground-truth relationship that the deep learning system is intended to capture, and consequently cannot typically be modelled explicitly. Instead, machine learning systems aim to find a representation of this mapping automatically by leveraging a training set $\{x_i, y_i\}_{0...n}$ to find a sufficiently performant approximation of f. This is referred to as supervised learning, and the resulting approximation found using the training set is denoted by $h : X \rightarrow \hat{Y}$, and typically referred to as a hypothesis.

To find such an approximation, we assume that there exists a joint probability distribution over X and Y, namely $P(x, y)$, and that the training data $\{x_i, y_i\}_{0...n}$ is drawn from this probability distribution such that the resulting sample distribution is independent and identically distributed to $P(x, y)$. This is the so-called iid assumption. By modelling the mapping as a joint probability distribution, one can model uncertainty in the predictions by expressing the output as a conditional probability $P(y|x)$. In conjunction with a loss-function $L(h(x), y)$ which measures the discrepancy between the hypothesis and the ground truth, these assumptions allows us to quantify the expected performance of a given hypothesis:

$$R(h) = \boldsymbol{E}[L(h(x), y)] = \int L(h(x), y) dP(x, y) \tag{2.1}$$

Using this framework, one can then find an iid-optimal hypothesis, often called a predictor, by finding the predictor h^* among a fixed class of functions (defined by network architecture) $\mathcal{H}$ that minimizes risk:

$$h^* = \underset{h \in \mathcal{H}}{\arg\min}\, R(h) \tag{2.2}$$

Since $P(x, y)$ is not known, however, one cannot compute $R(h)$ explicitly. Instead, the expected risk has to be estimated empirically, i.e by finding the arithmetic average of the risk associated with each prediction by the hypothesis over the training set:

$$R_{emp}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i) \tag{2.3}$$

This risk can in turn be minimized with respect to the hypothesis class. This is called empirical risk minimization (ERM):

$$\hat{h} = \underset{h \in \mathcal{H}}{\arg\min}\, R_{emp}(h) \tag{2.4}$$

To reiterate, the central idea with this approach to machine learning is that the training data can be considered a finite iid sampling of the underlying distribution. As such, by the central limit theorem, the hold-out performance of the computed hypothesis will approach iid-optimal performance given a sufficient amount of training data and some sufficiently capable and regularized training procedure. This should in theory allow deep learning systems to be able to generalize, since the empirical risk in theory can approximate the true risk arbitrarily well given sufficient training data support.

2.3.2 A taxonomy of Generalisation Failure Modes

As the analysis in 2.2 shows, ERM nonetheless readily fails to generate generalisable predictors with respect to out-of-distribution data). Understanding exactly why this is the case is a subject of ongoing study, and the literature around the matter is highly fragmented. Consequently, in an attempt to synthesize the findings from the often conflicting literature, it is beneficial to construct a taxonomy over the various forms of generalisation failure, the properties that induce them, and the how these properties emerge with respect to empirical risk minimization. First, consider the assumptions upon which ERM is based, namely that:

1. The empirical risk is a good approximation of the true risk $R(h)$
2. f exists in $\mathcal{H}$
3. $\{x_i, y_i\}$ is an IID sampling of $P(x, y)$. This is the aforementioned iid assumption.
4. $P(x, y)$ (and, given that 3 is true, $\{x_i, y_i\}$) is independently and identically distributed to full space of input-output pairs one might expect in deployment, henceforth denoted by $P_\infty(x, y)$.
5. $\hat{h}$ is unique in in $\mathcal{H}$
6. The optimizer consistently finds $\hat{h}$, given that it exists and is unique

The behavior that violations of assumptions 2 and 1 is well understood and fairly easy to detect, corresponding to underfitting and overfitting respectively, but violations of the remaining assumptions result in more subtle forms of generalisation failure. The general consensus is that generalisation failure can in broad strokes be attributed to either underspecification or structural misalignment. The following sections will attempt to summarize and synthesize the analyses within the literature, and connect each of the generalisation failure modes they identify to the above violations.

2.3.3 Structural Misalignment

Generalisation failure is often attributed to a structural misalignment between the training domain and the deployment domain, and that deep learning pipelines are particularly predisposed to inferring causally inviable but nonetheless empirical-risk-minimizing inductive biases that lack robustness to structural differences. For example: assume hospital A employs normal white-light endoscopy, and hospital B employs narrow-band endoscopy. Though one might expect the pipeline to learn features that are invariant to lighting conditions, and thus that a predictor trained on data from hospital A may generalize to hospital B, this is in no way guaranteed since the difference between the respective modalities cannot be directly accounted for through conventional ERM.

Though this not altogether preclude the pipeline from generating a generalisable predictor, it does make it highly unlikely. Since ERM really only minimises empirical risk, the model will instead learn highly predictive, but nonetheless causally inconsistent features. Using the same example as above, the model may instead learn to associate the characteristic texture of polyps under narrow-band imaging with a positive prediction, and rely entirely on this property in order to make predictions, which naturally does not translate well to white-light imaging. Similarly, the model trained on the white-light modality may learn to focus more on particular color- and shape-based features, which likewise may not transfer well to narrow-band imaging. This corresponds to a violation of assumption 4. If training on a dataset from hospital A, even assuming an infinite or otherwise theoretically optimal dataset that fully encapsulates the space of polyp images using white-light imaging - i.e $P(x,y)$, the pipeline will rarely if ever learn causally viable patterns that transfer well to the space of all possible polyps images across all possible non-destructive lighting conditions, or $P_\infty(x,y)$.

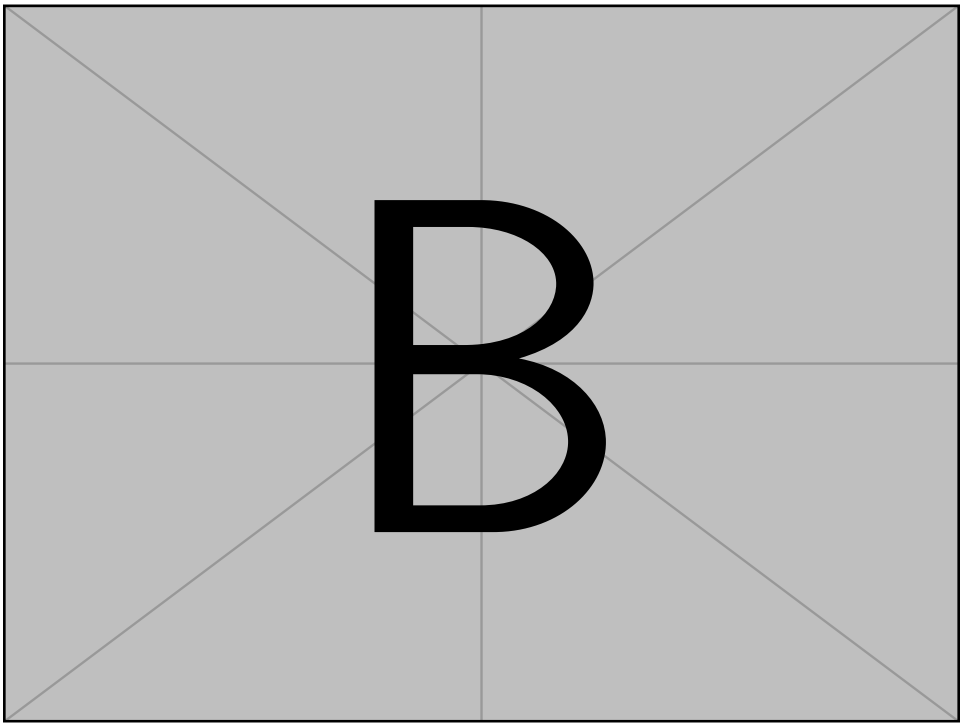

Figure 2.2:

In the above example, the patterns the respective pipelines infer, though not particularly representative of the broader context of what makes a polyp a polyp, make sense from a perspective of causality when considered from the perspective of either of the two hypothetical datasets. When considering only narrow-band imaging, it makes sense to heavily weigh the texture of the polyps. When considering only white-light imaging, it makes sense to heavily weight the shape and colour of the polyps. Generalisation failure is then in this case and by this line of reasoning entirely dependent on the structural misalignment and distributional shift corresponding to the change in imaging techniques as opposed to any erroneous logic in the pipeline itself. Ideally, the pipeline should of course detect patterns that generalize well regardless of lighting conditions, but it is not reasonable to expect the pipeline to draw this conclusion autonomously. Instead, the pipeline has to be "told" to keep this invariance in mind during training a priori. If some hypothetical change to the pipeline were to manage to induce this invariance, the structural misalignment between the dataset would no longer be an issue, ceteris paribus.

Spurious correlations and shortcut learning

In the above example, it is assumed that the inductive biases inferred in either domain are causally viable within for the respective imaging modalities. This is, as it turns out, not necessarily the case. As examplified in section 2.2, deep learning pipelines readily exploit predictive but nonetheless causally incorrect patterns.

2.3.4 Underspecification

[6] argues that this purely structural view of generalisation failure is flawed,

2.3.5 A Bayesian perspective of generalisation

2.4 Inducing robust features

2.4.1 Adversarial Training

2.4.2 Adversarial attacks and defences

2.4.3 Improved Risk estimation

Data augmentation

Distributional modelling

2.4.4 Bayesian Learning

2.4.5

2.5 Generalizability of polyp segmentation

2.6 Putting it all together

Chapter 3

Methodology

As described in earlier sections, good generalizability can only be achieved if the pipeline can reliably produce predictors that infer the right inductive biases and consequently generate robust features. Naturally, the set of correct inductive biases are not known, so any such pipeline instead has to learn to not infer the wrong inductive biases. To achieve this, a model of natural variance is constructed, which aims to encapsulate all the variability one might expect to see in the domain. This model can then be leveraged to force the pipeline to be robust to natural variance through contrastive learning. The central idea, then, is that it is more likely that the model learns to infer generalizable inductive biases as opposed to learning to learning to simply be robust to all possible configurations of a large amount of transformations.

To evaluate this, several predictors are trained from several pipelines with and without the influence of (algorithm name). Their performance is then evaluated on both a stress-test, and two separate polyp datasets, namely Etis-larib and EndoCV2021).

3.1 PLACEHOLDER ALGORITHM NAME

Summarizing the key points made in chapter 2, generalization is in large part a function of the set of inductive biases that one imbues in the pipeline and how well one can describe these inductive biases in a way that permits the model to learn them. The problem then boils down to the following questions:

1. How can behaviour consistent with the inductive biases be rewarded for the specific task?

2. How can the pertinent inductive biases be expressed in the context of deep learning?

3. How can we robustly optimize for consistent inductive biases?

3.1.1 Consistency Loss

Consistency can be expressed as the degree to which a certain model resists change when subjected to some form of perturbation. For a segmentation task, this corresponds to the model outputting corresponding segmentation masks for both the perturbed and unperturbed inputs. Of course, since the perturbation is not excluded from modifying the segmentation polyps themselves, it is necessary to also take into account the expected change due to the perturbation, and discount this from the overall change from the unperturbed to the perturbed domain. In simpler terms, the loss needs to describe the discrepancy between the expected change in the segmentation predictions due to the perturbation(s) and the actual change in the segmentation predictions. This loss will be referred to as the naked consistency loss (for reasons that will soon be made clear). Formally, this can be expressed as follows:

Let $Y := \{y, \hat{y} := f(x)\}$ be the set consisting of the segmentation predictions and masks for the unperturbed samples, and $\epsilon(\cdot)$ be some perturbation function. Then, let $A := \{a := \epsilon(y), \hat{a} := f(\epsilon(x))\}$ be the set consisting of segmentation predictions and masks for the perturbed samples. The naked consistency loss can then be expressed as follows:

$$L_c = \frac{1}{\{y \cup a\}} \sum \{y \ominus \hat{y} \ominus a \ominus \hat{a}\} \tag{3.1}$$

Where $\ominus$ denotes the symmetric difference.

The attentive reader may have noticed that this loss is zero not only if the predictions are both correct and consistent with one another, but also if the predictions are both incorrect, as long as they are consistent with one another. The reasoning behind this is that consistent behavior should be rewarded even if the model has not quite learned how to perform to an adequate standard. To illustrate, consider once more the example from chapter 2 with the narrow-band and white-light polyp-segmentation datasets. Assume that the perturbation function simply maps between the respective lighting modalities. In this case, the loss will reward the model if it predicts identical segmentations regardless of lighting conditions, even if they are both mostly incorrect. It will nevertheless be trying to infer patterns which are invariant to lighting, however, and consequently be causally stronger than a pipeline wherein the model is permitted to be inconsistent depending on lightning conditions.

Of course, using this loss in isolation is not really practical. For one, the model will have no way of knowing what the actual intent behind it is, and moreover the model will most likely learn the simplest possible interpretation of consistency and simply predict the same segmentation every time, regardless of the lighting conditions and for that matter regardless of the input. These are some pretty broad local minima, and there would naturally be a significant number of risk-equivalent predictors, which of course in accordance with the analysis in 2 constitutes generalisation failure on its own. Thus, it has to be combined with a task-specific loss, which for the polyp-segmentation task could be Dice-loss, Jaccard-Loss, binary cross entropy, etc.

Naturally, jointly optimizing for these two often conflicting objectives - overall task performance vs consistency - is not as straight forward as it may seem. The naive approach would be to simply add the task-loss and the naked consistency:

$$L = L_{task} + L_c \tag{3.2}$$

3.1.2 Model of natural variation

In order to account for any natural variation one may expect to find in deployment, it is necessary to construct a model which can parameterize the variability that is encountered, in other words a model of natural variability (MNV) Naturally, there is no way of knowing the full extent of all the types of variability one may find in the wild, but it may nonetheless be sufficient to model some subset thereof. This, naturally, requires some knowledge of the domain from which the dataset is collected. Similarly to how adding rotational augmentations is a bad idea for classification of hand-written numbers, certain transformations may or may not be suitable for use within a MNV.

In the case of polyp-segmentation, it is clear that it is necessary to account for variability in for instance lighting, polyp-size, polyp-shape, polyp-location, camera-quality, color-shifts, blurs, optical distortions, and affine transformations. Thus, a model is required that can (more or less) parametrize this variability. Broadly speaking, these transformations can be categorized as follows:

- Pixel-wise variability, which affect only the image, i.e color-shifts, brightness shifts, contrast-shifts, lighting, blurs etc

- Geometric variability, which affect both the image and the segmentation mask by some parametrizable quantity, i.e affine transforms and distortions

- Manifold variablity, which affects both the image and the segmentation mask depending on a learned model of the distribution, i.e the size, shape and location of polyps

Pixel-wise variability and geometric variability can be modeled fairly trivially through the use of the same transformations typically used for data-augmentation. Manifold-variability, however, is somewhat more difficult. Similar to how [15] employs cross-dataset style-transfer, it is necessary to find some way to model the distributional properties of the data, and then apply perturbations using the resulting model. Since both the size, shape, and position of polyps can be expected to vary, a model that can change all these factors is necessary. To this end, an in-painting model can be constructed. In particular, a GAN-inpainter.

Gan-based polyp inpainting

Geometric and pixel-wise transformations

3.1.3 Adversarial sampling

3.1.4 Difficulty annealing

3.2 Baselines

Several models were tested (...)

3.3 Datasets

3.4 Metrics and evaluation

Chapter 4

Results

Chapter 5

Analysis

5.1 Augmentation Robustness and Consistency Loss

As the results show, the performance of the pipeline that merely used augmentations is more or less equivalent to the performance exhibited by the modified pipeline. There is a very good reason for this: Consistency loss is mathematically equivalent to data augmentation, up to the choice of hyperparemeters - i.e augmentation probability, learning rates, etc. This section presents a proof of this fact, along with a theoretical analysis of how data augmentation affects the pipeline.

5.1.1 Data augmentation

Let $Y := \{y, \hat{y} := f(x)\}$ be the set consisting of the segmentation predictions and masks for the unaugmented samples, and $A := \{a := MNV(y), \hat{a} := f(MNV(x))\}$ be the set consisting of segmentation predictions and masks for the augmented samples. Finally, let $Z := \{z, \hat{z}\} \in_R \{Y, A\}$. The loss function subject to data augmentation can then be expressed as $L(Z \in_R Y, A)$, where L is any loss function. For the sake of simplicity in remaining calculations, this will be treated as the Jaccard loss, i.e $L(y, \hat{y}) := 1 - \sum y \cap \hat{y} / \sum y \cup \hat{y}$

$$L(Z \in_R Y, A)$$

5.1.2 Consistency loss

$$L_s = \frac{1}{\sum \hat{y} \cup \hat{a}} \sum \hat{y} \ominus y \tag{5.1}$$

$$L_c = \frac{1}{\sum \hat{y} \cup \hat{a}} \sum [\hat{y} \ominus y \ominus \hat{a} \ominus a] \tag{5.2}$$

$$L_{c+s} = L_c(Y, A) + L_s(Y) \tag{5.3}$$

$$= \frac{1}{\sum \hat{y} \cup \hat{a}} \left[\sum \{\hat{y} \ominus y \ominus \hat{a} \ominus a\} + \sum \{\hat{y} \ominus y\} \right] \tag{5.4}$$

$$\begin{aligned}
&= \frac{1}{\sum \hat{y} \cup \hat{a}} \left[\sum \{\hat{y} \ominus y\} + \sum \{\hat{a} \ominus a\} \right. \\
&- \sum \{\backslash y \cap \hat{y} \cap a \cap \hat{a}\} \cup \{y \cap \backslash \hat{y} \cap a \cap \hat{a}\} \cup \\
&\{y \cap \hat{y} \cap \backslash a \cap \hat{a}\} \cup \{y \cap \hat{y} \cap a \cap \backslash \hat{a}\} \\
&- \sum \{\backslash y \cap \hat{y} \cap \backslash a \cap \hat{a}\} \cup \{y \cap \backslash \hat{y} \cap \backslash a \cap \hat{a}\} - \\
&\cup \{\backslash y \cap \hat{y} \cap a \cap \backslash \hat{a}\} \cup \{y \cap \backslash \hat{y} \cap a \cap \backslash \hat{a}\} \\
&\left. + \sum \{\hat{y} \ominus y\} \right]
\end{aligned} \tag{5.5}$$

$$\begin{aligned}
&= 2L_s(y, \hat{y}) + L_s(a, \hat{a}) + \frac{1}{\sum \hat{y} \cup \hat{a}} \left[\vphantom{\sum} \right. \\
&- \sum \{\backslash y \cap \hat{y} \cap a \cap \hat{a}\} \cup \{y \cap \backslash \hat{y} \cap a \cap \hat{a}\} \cup \\
&\{y \cap \hat{y} \cap \backslash a \cap \hat{a}\} \cup \{y \cap \hat{y} \cap a \cap \backslash \hat{a}\} \\
&- \sum \{\backslash y \cap \hat{y} \cap \backslash a \cap \hat{a}\} \cup \{y \cap \backslash \hat{y} \cap \backslash a \cap \hat{a}\} - \\
&\left. \cup \{\backslash y \cap \hat{y} \cap a \cap \backslash \hat{a}\} \cup \{y \cap \backslash \hat{y} \cap a \cap \backslash \hat{a}\} \right]
\end{aligned} \tag{5.6}$$

The non-loss terms in equation 5.6 are proper subsets of the symmetric difference of the mask and segmentation across either dataset. The component of the loss that corresponds to these terms consequently grows in proportion to both $L_s(y, \hat{y})$ and $L_s(a, \hat{a})$. L_{c+s} and L_{sy+sa} are therefore monotonically correlated - i.e, when one grows, the other grows with it, and when one falls, the other one falls with it.

5.1.3 Adversarial Dice

$$L = \frac{1}{2} L(a, \hat{a}) + \frac{1}{2} L(y, \hat{y})$$

This should be asymptotically equivalent to data augmentation wit p=0.5

Chapter 6

Discussion

asdasdf

Bibliography

[1] Sharib Ali et al. 'EndoCV 2021 3rd International Workshop and Challenge on Computer Vision in En-doscopy'. In: ().

[2] Sharib Ali et al. 'Preface to: EndoCV2020Computer Vision in Endoscopy'. In: *CEUR Workshop Proceedings*. Vol. 2595. CEUR Workshop Proceedings. 2020.

[3] Ishita Barua et al. 'Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis'. en. In: *Endoscopy* 53.3 (Mar. 2021), pp. 277–284.

[4] Emma Beede et al. 'A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy'. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376718. URL: https://doi.org/10.1145/3313831.3376718.

[5] Battista Biggio et al. 'Evasion Attacks against Machine Learning at Test Time'. In: *Lecture Notes in Computer Science* (2013), 387–402. ISSN: 1611-3349. DOI: 10.1007/978-3-642-40994-3_25. URL: http://dx.doi.org/10.1007/978-3-642-40994-3_25.

[6] Alexander D'Amour et al. *Underspecification Presents Challenges for Credibility in Modern Machine Learning*. 2020. arXiv: 2011.03395 [cs.LG].

[7] Ivan Evtimov et al. 'Robust Physical-World Attacks on Machine Learning Models'. In: *CoRR* abs/1707.08945 (2017). arXiv: 1707.08945. URL: http://arxiv.org/abs/1707.08945.

[8] Robert Geirhos et al. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. 2019. arXiv: 1811.12231 [cs.CV].

[9] Robert Geirhos et al. 'Shortcut learning in deep neural networks'. In: *Nature Machine Intelligence* 2.11 (2020), 665–673. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00257-z. URL: http://dx.doi.org/10.1038/s42256-020-00257-z.

[10] D Heresbach et al. 'Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies'. en. In: *Endoscopy* 40.4 (Apr. 2008), pp. 284–290.

[11] Steven Hicks et al. 'The EndoTect 2020 Challenge: Evaluation and Comparison of Classification, Segmentation and Inference Time for Endoscopy'. In: Feb. 2021, pp. 263–274. ISBN: 978-3-030-68792-2. DOI: 10.1007/978-3-030-68793-9_18.

[12] Andrew Ilyas et al. *Adversarial Examples Are Not Bugs, They Are Features*. 2019. arXiv: 1905.02175 [stat.ML].

[13] A Leslie et al. 'The colorectal adenoma-carcinoma sequence'. en. In: *Br. J. Surg.* 89.7 (July 2002), pp. 845–860.

[14] D K Rex et al. 'Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies'. en. In: *Gastroenterology* 112.1 (Jan. 1997), pp. 24–28.

[15] Alexander Robey, Hamed Hassani and George J. Pappas. *Model-Based Robust Deep Learning: Generalizing to Natural, Out-of-Distribution Data*. 2020. arXiv: 2005.10247 [cs.LG].

[16] Dinggang Shen, Guorong Wu and Heung-Il Suk. 'Deep Learning in Medical Image Analysis'. In: *Annual Review of Biomedical Engineering* 19.1 (2017). PMID: 28301734, pp. 221–248. DOI: 10.1146/annurev-bioeng-071516-044442. eprint: https://doi.org/10.1146/annurev-bioeng-071516-044442. URL: https://doi.org/10.1146/annurev-bioeng-071516-044442.

[17] Sidney J. Winawer et al. 'Prevention of Colorectal Cancer by Colonoscopic Polypectomy'. In: *New England Journal of Medicine* 329.27 (1993). PMID: 8247072, pp. 1977–1981. DOI: 10.1056/NEJM199312303292701. eprint: https://doi.org/10.1056/NEJM199312303292701. URL: https://doi.org/10.1056/NEJM199312303292701.

[18] John R. Zech et al. 'Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study'. In: *PLOS Medicine* 15.11 (Nov. 2018), pp. 1–17. DOI: 10.1371/journal.pmed.1002683. URL: https://doi.org/10.1371/journal.pmed.1002683.