A POSTERIORI DEEP LEARNING SEGMENTATION QUALITY ESTIMATION BASED ON PREDICTION ENTROPY

José Márcio Martins da Cruz, Mateus Sangalli, Santiago Velasco-Forero, Etienne Decencière, Thérese Baldeweck

Submitted to Image Anal Stereol, 11 pages
Original Research Paper

A POSTERIORI DEEP LEARNING SEGMENTATION QUALITY ESTIMATION BASED ON PREDICTION ENTROPY

JOSÉ-MARCIO MARTINS DA CRUZ[1], MATEUS SANGALLI[1], ÉTIENNE DECENCIÈRE[1], SANTIAGO VELASCO-FORERO[1] AND THÉRÈSE BALDEWECK[2]

[1]MINES Paris, PSL Research University, Centre for Mathematical Morphology, 35 rue St Honoré, 77300 Fontainebleau, France, [2]L'Oréal Research and Innovation, 1 avenue Eugène Schueller, 93600 Aulnay-sous-Bois, France
(Submitted)

ABSTRACT

Image segmentation is a common intermediate operation in many image processing applications. On automated systems it is important to evaluate how well it, or its subsystems are performing without access to the Ground Truth. In Deep Learning based image segmentation there are very few methods to evaluate the output quality without using a ground truth. Most of them are based on the uncertainty (variance or standard deviation) of the prediction and can be applied to Bayesian Neural Networks, but not to Convolutional Neural Networks. In this research we propose to use the Entropy as a measure of uncertainty applied to the segmented image predicted by the Neural Network and some indicators based on it. The method is tested in a segmentation task of labeled skin images. The entropy based indicators are evaluated without knowing the ground truth and compared with indicators based on the real labels (Jaccard, Dice and Average Symmetrical Surface Distance). This experimentation showed that they are correlated and some Entropy based indicators can predict quite well the ground truth based indicators.

Keywords: Image segmentation, Deep Learning, Segmentation quality estimation, Entropy.

INTRODUCTION

Segmentation is, probably, the most important operation in image processing as it is part of most image applications. During research and development, algorithms are validated against an expected results: the Ground Truth (GT). In real production applications, when the GT is not available, it is desired, and sometimes a strong requirement, to estimate how good the segmentation result is.

Segmentation is an image operation which partitions it into a number of regions after some criteria. Pixels[1] inside each region share a common property. Haralick and Shapiro (1985) proposed, in the 80s, that a good image segmentation should obey four criteria:

1. Regions should be uniform and homogeneous with respect to some characteristic(s) such as grayscale intensity or texture;
2. Adjacent regions should have significant differences with respect to the characteristic on which they are uniform;
3. Region interiors should be simple and without many small holes;
4. Boundaries should be simple, not ragged, and be spatially accurate.

From the Machine Learning point of view, segmentation can be viewed, at pixel level, as a classification problem at which a class is assigned to each pixel depending on its own value and those from its neighbors.

Since then, many kinds of solutions for segmentation evaluation have been proposed (Zhang, 1996; Zhang *et al.*, 2008). Most of them are based on the morphology of regions or evaluating how different are adjacent ones using, most of the time, their intensity. All these solutions are not of particular interest to this research because few of them are related to Deep Learning (DL) and the few of them based on entropy employ this concept with a particularly unusual meaning (Pal and Pal, 1993; Pal and Bhandari, 1993; Zhang *et al.*, 2003; Hao *et al.*, 2009; Rill-García *et al.*, 2022). They use entropy as a criterion to evaluate the intensity homogeneity inside each region (narrow histograms) before checking if the histograms of adjacent regions are disjoint. In this paper we use entropy with its usual Information Theory (IT) meaning: the uncertainty or amount of information still needed to make the right decision. We mention their research here just to avoid confusion.

The arrival of DL to process images provided more efficient segmentation methods, mainly on textured regions where histograms may overlap. In

[1]The word "pixel" is used here but all results are valid also for images with three or more dimensions.

classification problems, the last layer of a neural network (NN) is usually a *Softmax* layer[2], where each component corresponds to the probability of assigning the corresponding class to the pixel. These probabilities gave birth to the concept of "uncertainty" attached to the classification results. The interest of "uncertainty" results from the hypothesis that most errors happen on regions of data where we are less certain on what decision to take. Some research work confirmed this hypothesis (Kampffmeyer *et al.*, 2016; Kendall *et al.*, 2017). A summary will be presented in the next section.

The concept of Entropy from IT is, in our opinion, very interesting with many possible applications. In the theory part of this paper we will show the meaning of entropy as a measure of uncertainty per pixel in image segmentation and its lower and upper bounds. In the experimental part of this paper we show how to use it to identify regions in the image where most errors may be occurring and propose an indicator of the segmentation quality, like the Jaccard, Dice, Average Symmetric Surface Distance (ASSD) and other indexes, but without access to GT.

RELATED WORK

DL brought much more reliable solutions to many problems, including segmentation of images with textured regions. We are mainly interested in two DL NN paradigms: Convolutional Neural Networks (CNN) and Bayesian Neural Networks (BNN). Most research with the theme "Uncertainty" comes from BNNs domain. In BNNs the weights usually found in CNNs are replaced by distributions. Furthermore, the epistemic uncertainty[3] can be deduced from these distributions.

Uncertainty in DL was deeply developed in Gal (2016) PhD Thesis and by Kendall and Gal (2017) for BNNs and extended to CNNs. *"Uncertainty"*, in this research shall be understood as the complement of *"confidence"* and is usually expressed by the *"variance"* (or *standard deviation*) of some result.

Gal (2016) and Gal and Ghahramani (2016) have shown that inserting a random dropout just before every weight layer in CNNs, during both training and prediction, is equivalent to the probabilistic Gaussian process in BNNs. Monte Carlo Sampling (or Monte Carlo Dropout) is done by repeating predictions a number of times for each data object to evaluate the variance on output probabilities and produce an uncertainty map. This method is simple but requires that prediction be repeated a number of times. When superposing the distribution of each class, the less the class softmax distributions overlap, the more certain are the predictions.

Kampffmeyer *et al.* (2016) applied Monte Carlo sampling (10 samples) on CNNs to evaluate the standard deviation over the softmax outputs of samples during prediction. Their research confirms the link between uncertainty and segmentation accuracy.

Hendrycks and Gimpel (2017)[4] have found that *"simple statistics derived from softmax distributions provide a surprisingly effective way to determine whether an example is misclassified or from a different distribution from the training data"*. They explored the idea that smaller values of the maximum of Softmax vector indicate higher error probability. They experimented with classification tasks in various domains, all of them having Softmax as the final layer of the NN. None of them were image segmentation. Roughly speaking, to validate the idea they partitioned samples into two sets, correctly and wrongly classified, based on a threshold set on maximum value of Softmax for each sample. Although they extensively tested the idea they did not explore the theory behind it nor its limitations. Also, we have not found any research indicating how their heuristics are related to uncertainty. Their paper deserves particular attention as, according to SemanticScholar[5], it was cited more than 2000 times. In the next section we will show that this idea can be understood under the light of IT: how it is related to Shannon Entropy (Shannon, 1948) and its limitations.

Using BNNs, DeVries and Taylor (2018a) proposed *"Learned Confidence Estimates"*, a method where the network produces two separate outputs: prediction probabilities and the confidence estimates. In another research paper, DeVries and Taylor (2018b) evaluated their method against four other methods, including the max of Softmax from Hendrycks and Gimpel (2017) and Entropy of Softmax applied to Monte Carlo Dropout. In their setup they had two separated NNs: one to produce prediction probabilities and uncertainty estimation (as a score related to it) from input images and the other to estimate the

[2]The *Softmax* (Bridle, 1989), also known as *normalized exponential*, converts a vector of real numbers (*logits*) into a probability distribution. This function can be seen as a generalization of the *logistic sigmoid function* (Bishop, 2007, p. 198). The elements of the output vector are in the range $(0, 1)$ and sum to 1.

[3]The *epistemic uncertainty* is the uncertainty of the model while the *aleatoric uncertainty* is the one from the input data: noise, out of focus, ...

[4]This article, retrieved from arxiv.org, was accepted as a poster at ICLR 2017.

[5]See: `https://api.semanticscholar.org/CorpusID:13046179`

segmentation quality indicator (Jaccard index[6]) from the first network results. The predicted Jaccard index was then compared against the true Jaccard index obtained from GT. Finally they showed that from the five methods used to predict Jaccard index, the max of Softmax (Hendrycks and Gimpel, 2017) and Monte Carlo Dropout presented very similar results, slightly better than their method. Particularly all confidence (based on variance) methods perform worse than those using direct statistics of the Softmax.

In a recent preprint paper Galil *et al.* (2022) have found that the entropy of the Softmax is slightly better than using just its maximum as a score of uncertainty.

Nair *et al.* (2020) compared, in a two class problem with Monte Carlo Dropout setup, the variance of the Softmax vector, the *Predictive Entropy* and the *Mutual Information*. The *Predictive entropy* is defined as the entropy of the mean of Softmax vectors and the Mutual Information is defined as the difference between the *Predictive Entropy* and the mean entropy of Softmax vectors. The interesting point of this research is establishing the difference in how these values are interpreted: while the variances and mutual information evaluate the confidence in the predicted value (the model's uncertainty), the Predicted Entropy evaluates the uncertainty of the prediction, supposing that the predicted value is correct.

Sometimes, the words "confidence" (or "lack of") or "confidence interval" and "uncertainty" are employed interchangeably. Rigorously, although both are related to the quality of results, they do not represent the same point of view.

Finally, we have found few results encouraging the use of entropy as a measure of uncertainty or quality of results in DL applications. But, effectively, from the IT point of view, Entropy can be seen as a measure of uncertainty or information still needed to make a decision.

ENTROPY AND UNCERTAINTY

IT is about quantifying information: the amount of information contained in, e.g., a file or the information still needed to make some decision without ambiguity. The latter is how entropy is understood, under IT, as a measure of *uncertainty*.

The entropy of a discrete random variable X is defined as (Shannon, 1948; Cover and Thomas, 2006):

$$H(X) = -\sum_{x\in\mathscr{X}} p_x \, log_2(p_x) \tag{1}$$

where $\mathscr{X}$ is the finite set of possible outcomes and p_x is the probability of having x as the outcome. $H(X)$ takes values in the interval $[0, log_2(|\mathscr{X}|)]$, where $|\mathscr{X}|$ is the cardinality of $\mathscr{X}$. The minimum value happens when one of the possible outcomes has probability 1 and all other 0 and the maximum when all outcomes are equally probable ($1/|\mathscr{X}|$).

The final layer in a NN for image segmentation is usually a *Softmax* layer, with one component per class. For each pixel, *Softmax* values sum to one. A class is the label assigned to each region in the segmented image. The same label may be assigned to different unconnected regions having some common characteristics.

In the prediction image, we assign to each pixel the class $\hat{y}$ corresponding to the maximum value (σ_{max}) of the Softmax vector (σ), where c is a class in the set of classes $\mathscr{C}$.

$$\sigma_{max} = \max_{c\in\mathscr{C}} \sigma(c) \tag{2}$$

$$\hat{y} = \arg\max_{c\in\mathscr{C}} \sigma(c) \tag{3}$$

We can notice that the minimum value of σ_{max} is $1/|\mathscr{C}|$. This happens when all classes are equal in probability. So, $\sigma_{max} \in [1/|\mathscr{C}|, 1)$

The prediction entropy can be written as:

$$h = -\sum_{c\in\mathscr{C}} \sigma(c) \, log_2(\sigma(c)) \tag{4}$$

and broken down into:

$$h = -\sigma_{max} \, log_2(\sigma_{max}) - \sum_{c\in\mathscr{C}\setminus\{\hat{y}\}} \sigma(c) \, log_2(\sigma(c)) \tag{5}$$

The first term in the right hand side corresponds to the contribution of the predicted class to the prediction entropy and is referred to as residual entropy. It is the value taken into account by Hendrycks and Gimpel (2017) in their heuristics, which neglects the second term. Now, we can evaluate what could be the result without neglecting the residual entropy and the difference.

The value of the residual entropy ($h_{residual}$) depends on how the residual probability ($1-\sigma_{max}$) is distributed over the remaining classes. We can easily find its upper and lower bounds.

[6]The Jaccard index (Jaccard, 1912), also called *Intersection over Union* or IoU, has been generalized to be used as a multiclass indicator (Ruzicka or MeanIOU indexes) or even as a loss function with real values in the interval $[0,1]$. Most generalizations, even different from the original indicator, are frequently still called just Jaccard.

The upper bound is attained when the residual probability is evenly distributed over the remaining classes:

$$max(h_{residual}) = -(1-\sigma_{max})\, log_2\left(\frac{1-\sigma_{max}}{|\mathscr{C}|-1}\right) \quad (6)$$

On the other hand, the lower bound is attained when the residual probability is concentrated on a minimum number of classes, knowing that the maximum value is σ_{max}. If N_f is the number of classes which can be fulfilled and σ_r is the remainder, we can write:

$$N_f = \left\lfloor \frac{1-\sigma_{max}}{\sigma_{max}} \right\rfloor \quad (7)$$

$$\sigma_r = (1-\sigma_{max}) - N_f\, \sigma_{max} \quad (8)$$

$$min(h_{Residual}) = -N_f.\sigma_{max}\, log_2(\sigma_{max}) - \sigma_r\, log_2(\sigma_r) \quad (9)$$

	Classes					
	0	1	2	3	**Entropy**	**Comments**
1	1.000	0.000	0.000	0.000	0.000	Minimum
2	0.400	0.400	0.200	0.000	1.519	Low bound
3	0.400	0.300	0.200	0.100	1.846	
4	0.400	0.200	0.200	0.200	1.922	Up bound
5	0.250	0.250	0.250	0.250	2.000	Maximum

Table 1: A numeric toy example showing how the distribution of Softmax values impacts prediction entropy in a four class problem. Class 0 is the winning class, as an example. Rows 2 to 4 correspond to the situation where σ_{max} is fixed to 0.4. Notice that the predicted class is not unique in rows 2 and 5.

Table 1 shows how the distribution of Softmax values affects the prediction entropy in a four class toy problem.

Based on the above, Figure 1 presents the upper and lower bounds of the prediction entropy and the contribution of σ_{max} to it against σ_{max} for a problem with 2, 3 and 8 classes. Some conclusions can be drawn:

- the contribution of σ_{max} to the entropy has a maximum at $1/e$. So, for problems where the minimum value allowed for the probability of the winning class is smaller than $1/e$ the contribution of σ_{max} is no longer a monotone function. This happens when the number of classes is greater than two.
- when $\sigma_{max} < 1/e$ its use as a score of uncertainty is qualitatively correct but quantitatively wrong because the contribution of σ_{max} to the entropy is no longer a bijective function;
- in the particular case of two classes, $\sigma_{max} \geq 0.5$ always, so its use as a score of uncertainty is valid.

It is worthwhile to notice that if one sums up the entropy of all pixels one will get the global uncertainty of the segmentation distributed over the whole image. The key point is what is the best way to aggregate uncertainty evaluated on each pixel into an index associated to the whole image. This is the subject of the next section.

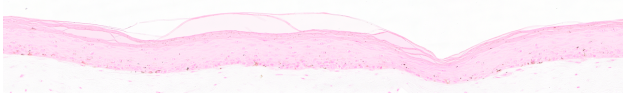
EXPERIMENTS

The experiments here are intended to investigate two points: is there a correlation between entropy based indexes and GT ones and how the relation between indexes varies with the segmentation quality level. Answering these questions is a step to solving the problem of evaluating the segmentation quality without having access to the GT.

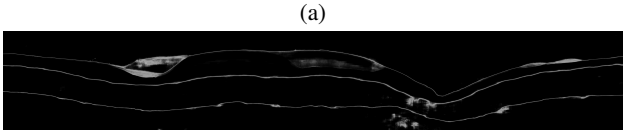
There are many ways to use entropy as a measure of uncertainty in image segmentation. The Softmax output of a NN can be used to generate an uncertainty map, where the value associated to each pixel is its entropy. In this section we will present and compare four entropy based indexes and rank them in order of correlation. At the same time, we compare results from these indexes against those from the heuristics proposed by Hendrycks and Gimpel (2017).

DATASET

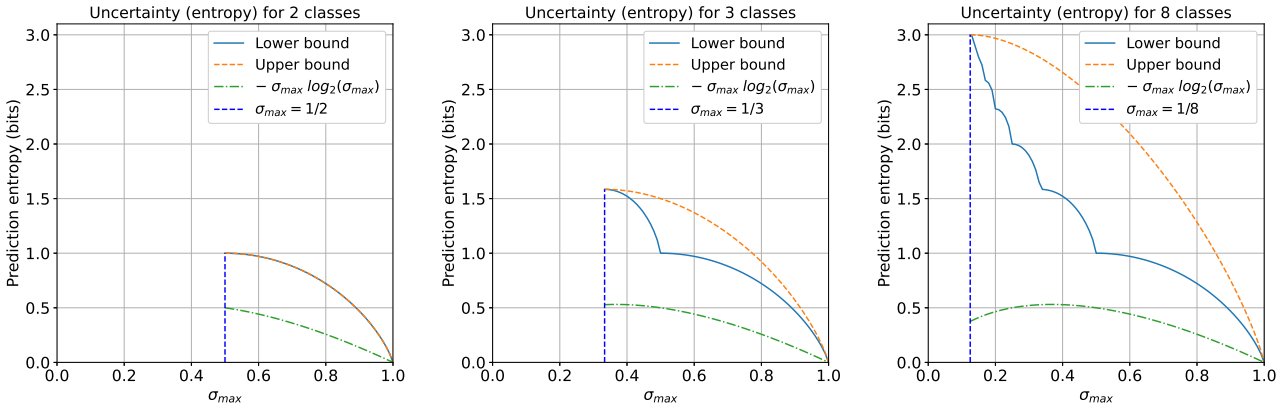
We will be using a private dataset of images coming from microscopy. Images represent reconstructed skin sections stained with the Fontana-Masson method. These images are to be segmented and categorized into three classes: *Stratum Corneum* (SC), *Living Epidermis* (LED) and a third one which is the union of the *tissue corresponding to dermis* and the background. The test data set contains 175 images and was segmented in a UNet CNN Network trained on a set of 215 images. The mean dimensions of the images are 3400x1200 pixels (variable size). An example of these images is shown in Figure 2. Although we are not concerned with details of the NN, its results were chosen from a preliminary version of the project in order to have more less well segmented images.



(a)

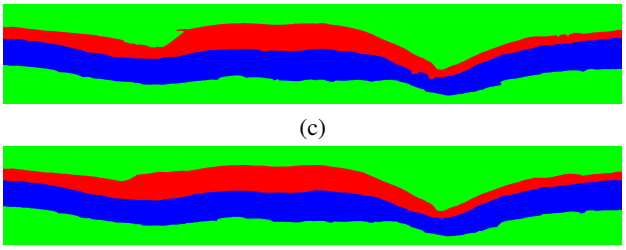


(b)



(a) (b) (c)

Fig. 1: Upper and lower bounds of entropy against the probability of the predicted class and the partial contribution of σ_{max} to entropy for segmentation with 2 (a), 3 (b) and 8 (c) classes



(c)

(d)

Fig. 2: Example of image of reconstructed epidermis: Original (a), Entropy map (b), Segmentation after post-processing (c) and Segmentation GT (d). Colors legend: Red: SC, Blue: LED, Green: background and dermis. Size 5477x874 pixels. Lighting and contrast may change between acquisitions. Processing depends more on texture than on pixel levels.

IMPLEMENTATION DETAILS

A block diagram of our application is shown in Figure 3. "Green" blocks are those handling uncertainty. During initial development we have found that most errors happen in regions with high uncertainty touching or intersecting a neighborhood of interfaces of segmented regions. Regions with high uncertainty, far from interfaces are not relevant as, either way, they will be removed at post-processing. This hypothesis may not be true for every problem. This hypothesis allows us to select only relevant regions with high uncertainty. Notice that we do not care about what the NN does and how: we just need the Softmax output and, whenever possible the predicted segmented image with its post-processing already done.

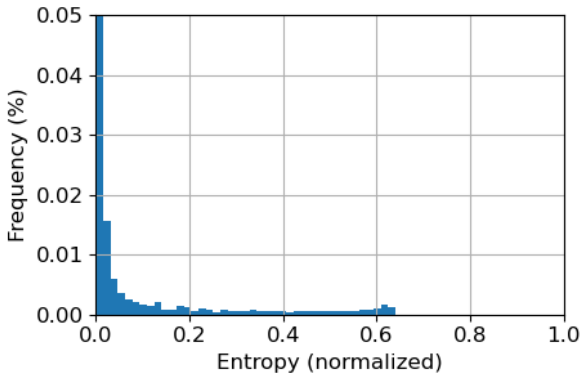


Fig. 4: Typical histogram of the entropy map. The small mode just beyond 0.6 ($1/log_2(3)$) corresponds to maximum uncertainty interfaces neighborhood: Softmax taking values around $[0.5, 0.5, 0]$.

In the upper part of the block diagram, we create the entropy map from the Softmax results. To get values in the interval $[0,1]$, entropy is divided by $log_2(|\mathscr{C}|)$, where $\mathscr{C}$ is the set of classes. An hysteresis threshold (Canny, 1986) is applied to the entropy map to find regions with high uncertainty. Threshold levels are easlily defined. A typical histogram of an entropy map contains a mode at around $1/log_2(|\mathscr{C}|)$ as shown in Figure 4. This value corresponds (but not only) to points where the uncertainty is maximal at interfaces: commonly any permutations of the Softmax configuration $[0.5, 0.5, 0, \ldots, 0]$. In our case, we found that 0.55 and 0.45 as the high and low thresholds are good choices. This corresponds to approximately 0.9 and 0.7 times the critical histogram value in a three class problem. The final block, a morphological opening (Serra, 1982), is needed to remove very thin

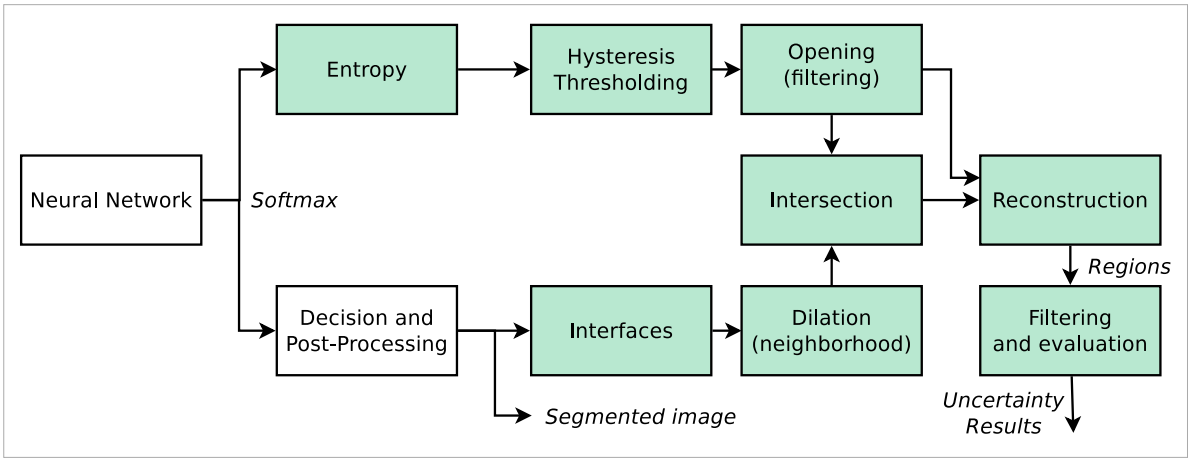


Fig. 3: Block diagram of an application detecting regions with high uncertainty in a segmented image.

regions, mainly around interfaces, which are of no interest.

In the lower part of the diagram, the first block creates the segmented image from σ_{max} (as defined in a previous section) of each pixel and does some post-processing (regularization of interfaces, filling holes, etc). A morphological gradient is applied to the segmented image to find the interfaces, which are dilated by some number of pixels to define the neighborhood. The size of the dilation structuring element does not impact significantly results if chosen in the interval $[5, 15]$ pixels. Results presented hereafter are for a neighborhood of ten pixels.

Finally, the intersection and reconstruction allow to select only regions intersecting or touching the neighborhood of interfaces. The two parameters - size of opening and size of neighborhood - depend on the spatial characteristics of the images and the two threshold values depend on the number of classes and, roughly, the global quality of segmentation.

EVALUATION

For each image in the test set we collected: the sum of the areas of regions with high uncertainty (SAR), the sum of the entropy inside these regions (SER), the area of the biggest region (ABR), the mean entropy in the whole segmented image (MEI) and the mean of the maximum Softmax values[7] in the whole segmented image (MSI). Reference indexes are evaluated thanks to the availability of GTs. We use two overlap based methods - (Jaccard (Jaccard, 1912) and Dice-Sorensen (Dice, 1945)) - and a boundary distance based one (ASSD (Yeghiazaryan and Voiculescu, 2018)). Coucou.

Firstly we verify the statistical correlation between indexes derived from GT (Jaccard, Dice and ASSD) against entropy derived indexes. Pearson and Spearman (Spearman, 1904) methods were used in correlation. Pearson verifies how well one variable can linearly fit the other one. Spearman method uses the rank of the values instead of the values: the goal is to verify the monotonicity of the relation. Moreover, the Spearman method is less sensitive to outliers and accepts non-linear relations. When trying to substitute one variable by another one, it is may be more interesting to verify the Spearman correlation. We choose the Spearman method instead of Kendall's (Kendall, 1938), another widely used rank correlation method, because Spearman and Pearson evaluation is exactly the same, allowing to easily compare both results.

In the second part of the experiments the set of samples are partitioned into Good and Bad subsets based on some threshold set in GT based indexes (Jaccard, Dice and ASSD). The Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) (Fawcett, 2006) is used to verify how well each entropy derived index predict the correct subset of images based on a GT derived index. An AUC with value "1" means that the entropy derived index can predict perfectly the GT derived one. This is done for various levels of threshold on GT based index and entropy based index. While the correlation coefficient allows to verify the compatibility of indexes over the entire range of values partitioning the samples set over allows to verify the indexes compatibility at different levels of segmentation quality.

[7] In fact, from (Hendrycks and Gimpel, 2017) heuristics, pixel uncertainty is estimated by $(1 - \sigma_{max})$

RESULTS

Figure 5 plots the Jaccard index against the SAR with the experimental conditions described above[8]. As it will be shown in the following, SAR is, globally, the index with the highest correlation with GT derived indexes. We can visually identify a group of points suggesting a linear relation between the two variables. Even if the concept of "good segmentation" is quite subjective, we can arbitrarily set a threshold of 0.9 on the Jaccard index and a threshold of 0.4 million pixels on the SAR, as a support to qualitatively interpret these results. This is plotted on Figure 5 and defines four regions on the graphics, from left to right and from top to bottom: true bad (TB), false bad (FB) (Jaccard says segmentation is correct but not SAR), false good (FG) (Jaccard says segmentation is wrong but not SAR) and true good (TG), indicating how well the SAR predicts Jaccard index. FGs and FBs are samples for which SAR fails to predict the right value range of Jaccard index.

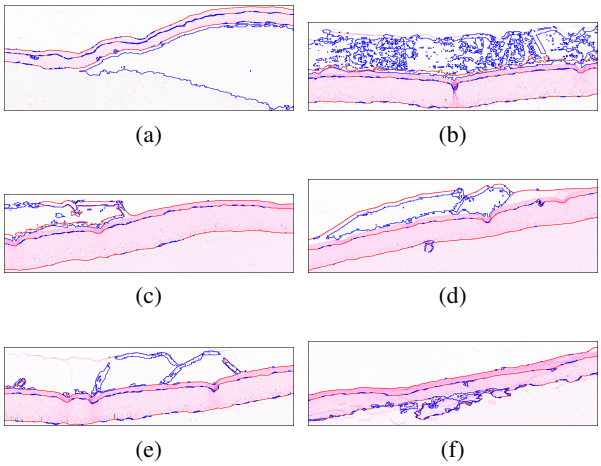


Fig. 6: Outliers (a) and (b), False bad (c) and (d) and False good (e) and (f), after SAR and Jaccard indexes. Predicted interfaces are indicated by red lines and uncertainty regions by blue lines.

The two topmost FBs (Figure 6 (a) and (b)) correspond to outliers with lighting problems resulting in a really bad segmentation. The other two FBs (Figure 6 (c) and (d)) were correctly segmented, after Jaccard, but with large regions with high uncertainty around the interfaces. In a real world application, even if the segmentation of these FB images was done quite well according to the Jaccard index, the operator must be warned about them and the SAR index allows it. About FGs (Figure 6 (e) and (f)), both images are wrongly segmented according to the Jaccard index, with regions with high uncertainty but too far from the predicted interfaces which are far from the GT ones. Increasing the neighborhood parameter will not solve the issue and will increase FBs count.

With outliers					
Index	SAR	SER	ABG	MEI	MSI
Jaccard	**-0.3850**	-0.3805	-0.2565	-0.3287	0.3011
Dice	**-0.3808**	-0.3770	-0.2535	-0.3157	0.2903
ASSD	**0.3865**	0.3793	0.2751	0.3131	-0.2888
Without outliers					
Index	SAR	SER	ABG	MEI	MSI
Jaccard	-0.7338	**-0.7361**	-0.5689	-0.4252	0.4656
Dice	-0.7379	**-0.7409**	-0.5781	-0.4137	0.4559
ASSD	**0.7552**	0.7544	0.6403	0.4137	-0.4618

(a) Pearson

Index	SAR	SER	ABR	MEI	MSI
Jaccard	**-0.6854**	-0.6834	-0.6528	-0.4230	0.4388
Dice	**-0.6850**	-0.6830	-0.6522	-0.4229	0.4387
ASSD	**0.8229**	0.8221	0.7829	0.5392	-0.5530

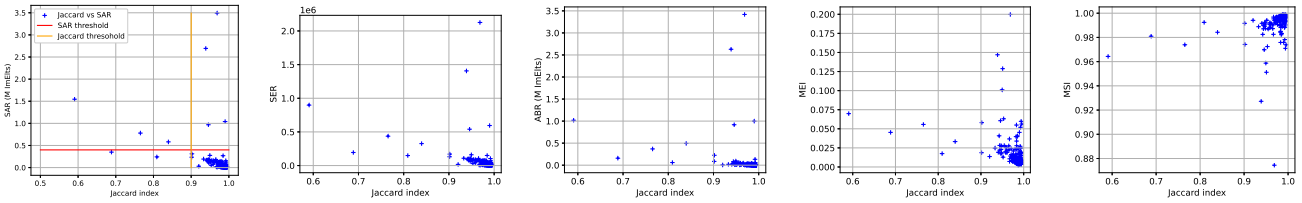
(b) Spearman

Table 2: Pearson (a) and Spearman (b) correlation coefficients between entropy based indexes and GT based ones. Because the Pearson correlation is too sensitive to outliers, results are presented with and without them.

Table 2 shows the correlation between entropy based indexes against Jaccard, Dice and ASSD indexes evaluated when the GT is available. Pearson correlation is shown with and without the two outliers. SAR and SER are, in both correlation methods, the winners. It is true that when comparing just MEI against MSI, the latter is the winner by a small margin, but both have correlation coefficients much smaller than SAR and SER. Finally, we notice that Pearson correlation is greater than Spearman. This indicates that the relationship between GT based indexes and entropy based ones is better represented by a linear function than by a monotonic one.

The next step is to evaluate how well entropy based indexes can predict GT based ones. As explained in Section 4.2, the test set is partitioned in Good and Bad based on a threshold set in the GT index, for each couple of indexes from GT and entropy based ones. AUC is used to verify how well each entropy based index can predict GT one. Results are presented in Figure 7 and clearly show that the prediction quality is not the same over the entire range of the GT based indexes, mainly when the segmentation quality is high. The particular meaning of AUC must be recalled to interpret this result: the AUC is the probability that if we take a couple of samples, one of each class, their ranking is the right one. This is equivalent to the Wilcoxon test of ranks (Hanley and McNeil, 1982).

[8]Dice and ASSD results are qualitatively the same.



(a) (b) (c) (d) (e)

Fig. 5: Graph of Jaccard threshold against indexes: SAR - Sum of areas of all regions (a), SER - sum of entropy in all regions (b), ABR - area of biggest region (c), MEI - mean entropy in image (d) and MSI - mean softmax in image (e). SAR (a) is the best in most cases.

When we change the threshold, up or down, some samples change their class with respect to GT index but not about the entropy based index, so this can decrease the AUC. In other words, the order of ranks between the indexes is not perfectly verified in small neighborhoods, while it is in large ranges. This is expected as while the Spearman correlation is quite good it is smaller than the Pearson as shown in Table 2.

DISCUSSION

We observed that for all indexes, AUC decreases when the segmentation quality approaches "perfection", after GT based index. This is not surprising because there are too many samples at this region and indexes based on different concepts do not reproduce rank order even if placed in equivalent range of values. This effect was predicted by Spearman correlation from Table 2, lower than Pearson correlation, meaning that rank is better preserved on long range of values than in short ones. Also, even among GT based indexes the correlation is not perfect: 0.931 for ASSD against Jaccard or Dice. This conclusion may surely be applied on the opposite extreme of indexes range when segmentation becomes too bad, but more samples having bad segmentation quality are needed. We can also remark than even among GT based indexes the correlation is not perfect: 0.931 for ASSD against Jaccard or Dice.

An important point is that entropy derived indexes evaluated on regions of high uncertainty are more effective than the entropy alone evaluated over the entire image. This is because regions with errors may be too small compared to the whole image and high local entropy values will be diluted over the entire surface of the image. This is important mainly when small regions are to be detected. Other than the theoretical reason we also have shown that entropy based indexes applied to image segmentation are more efficient than the heuristics proposed by Hendrycks and Gimpel (2017).

FUTURE DIRECTIONS

In the experimental part of this research we have shown that it can be possible to use some entropy derived indexes as a replacement of the GT based indexes when the GT is not available. More research using other image databases and similar problems with more classes could be interesting.

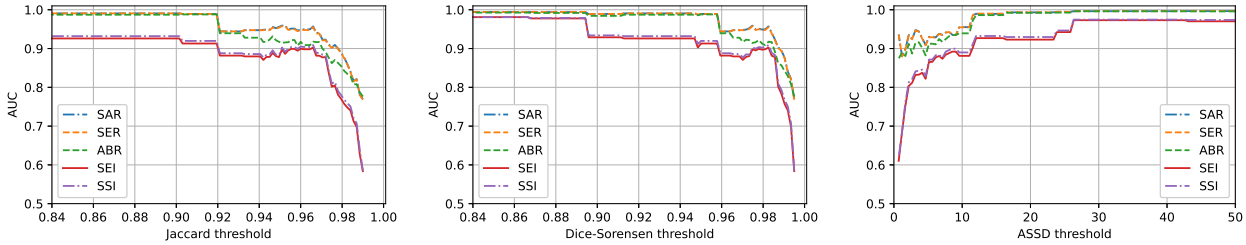
Figure 8 shows image crops from the problem we worked on: an entropy map, the regions with high uncertainty, the resulting segmented image and the original image with superposition of regions interfaces and the boundaries of regions with high uncertainty. To improve segmentation results, these regions can be used as feedback in a previous stage to suggest where to make an additional effort. In a real world application, a visual indicator of possible problems on the process workflow is very interesting.

The idea behind Active Learning (Settles (2010)) is that a good predictive model can be built with a minimum number of samples if the selection of samples to be learned is done by the model and not imposed on it. Some entropy derived index could be used as a criteria to select which samples can be used to incrementally build the model, as done by, e.g., Moon *et al.* (2020).

The entropy is a natural indicator of uncertainty, not specific to image segmentation. Whenever you have a NN with a Softmax as the last layer, so a probability distribution, you can evaluate the associated entropy and associate it to some uncertainty. So, it would be interesting to experiment with entropy based indicators on NN applications other than image segmentation that have Softmax as one of its layers.

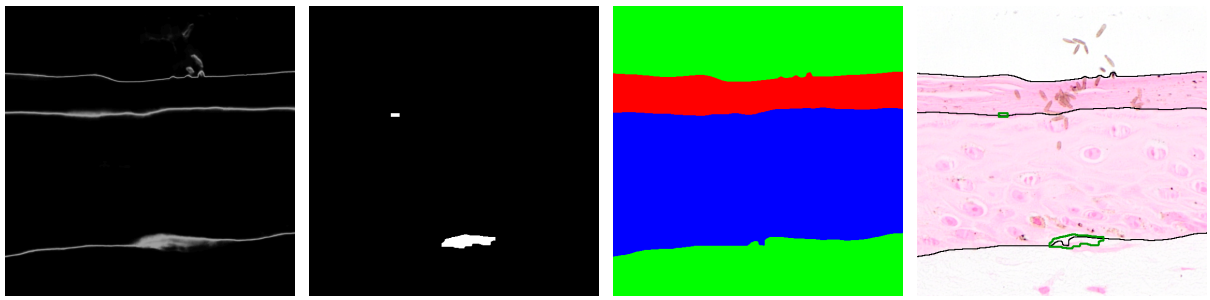
CONCLUSIONS

To be able to evaluate the quality of the results of an application in the real world without access



(a) (b) (c)

Fig. 7: Prediction of GT based indexes by entropy based ones can be estimated by the AUC for different values of thresholds set on GT indexes: Jaccard (a), Dice (b) and ASSD (c). To be remarked the quasi superposition of some results: SAR with SER and SEI with SSI on almost the entire range of values.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 8: Entropy map (a) and (e), regions with high uncertainty (b) and (f), segmented image (c) and (g) and the original image with segmentation interfaces and boundary of regions superimposed on it (d) and (h).

to the expected result is becoming an important requirement in operational automated applications. Entropy is a natural measure of uncertainty (or amount of information) associated to a probability distribution. We demonstrated, in this research, that some entropy derived indexes may be good candidates. This is a flourishing domain and we hope that this research will motivate other works to make use of the entropy concept.

We also presented an application framework which can surely be used as a starting point to be integrated into real world applications.

REFERENCES

Bishop CM (2007). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 1st ed.

Bridle JS (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In: Proceedings of the 2nd International Conference on Neural Information Processing Systems, NIPS'89. Cambridge, MA, USA: MIT Press.

Canny J (1986). A computational approach to edge detection. IEEE Transactions on Pattern Analysis

and Machine Intelligence PAMI-8:679–98.

Cover TM, Thomas JA (2006). Elements of Information Theory (2. ed.). John Wiley & Sons, Ltd.

DeVries T, Taylor GW (2018a). Learning confidence for out-of-distribution detection in neural networks. https://arxiv.org/abs/1802.04865.

DeVries T, Taylor GW (2018b). Leveraging Uncertainty Estimates for Predicting Segmentation Quality. https://arxiv.org/abs/1807.00502.

Dice LR (1945). Measures of the amount of ecologic association between species. Ecology 26:297–302. Full publication date: Jul., 1945.

Fawcett T (2006). An introduction to ROC analysis. Pattern Recognition Letters 27:861–74. ROC Analysis in Pattern Recognition.

Gal Y (2016). Uncertainty in Deep Learning. Ph.D. thesis, University of Cambridge.

Gal Y, Ghahramani Z (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Proceedings of The 33rd International Conference on Machine Learning. PMLR. ISSN: 1938-7228.

Galil I, Dabbah M, El-Yaniv R (2022). Which models are innately best at uncertainty estimation ? https://arxiv.org/abs/2206.02152.

Hanley JA, McNeil BJ (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36. PMID: 7063747.

Hao J, Shen Y, Xu H, Zou J (2009). A region entropy based objective evaluation method for image segmentation. In: 2009 IEEE Instrumentation and Measurement Technology Conference.

Haralick RM, Shapiro LG (1985). Image Segmentation Techniques. Computer Vision Graphics and Image Processing 29:100–32.

Hendrycks D, Gimpel K (2017). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In: 6th International Conference on Learning Representations - ICLR 2017 - (poster), vol. abs/1610.02136.

Jaccard P (1912). The distribution of the flora in the alpine zone.1. New Phytologist 11:37–50.

Kampffmeyer M, Salberg AB, Jenssen R (2016). Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

Kendall A, Badrinarayanan V, Cipolla R (2017). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Proceedings of the British Machine Vision Conference (BMVC). BMVA Press.

Kendall A, Gal Y (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Red Hook, NY, USA: Curran Associates Inc.

Kendall MG (1938). A new measure of rank correlation. Biometrika 30:81–93.

Moon J, Kim J, Shin Y, Hwang S (2020). Confidence-Aware Learning for Deep Neural Networks. https://arxiv.org/abs/2007.01458.

Nair T, Precup D, Arnold DL, Arbel T (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical Image Analysis 59:101557.

Pal NR, Bhandari D (1993). Image thresholding: Some new techniques. Signal Processing 33:139–58.

Pal NR, Pal SK (1993). A review on image segmentation techniques. Pattern Recognition 26:1277–94.

Rill-García R, Dokladalova E, Dokládal P (2022). Syncrack: Improving Pavement and Concrete Crack Detection Through Synthetic Data Generation. In: 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP'22). on-line, France.

Serra J (1982). Image Analysis and Mathematical Morphology, vol. 1. Academic Press.

Settles B (2010). Active learning literature survey - Computer Sciences Technical Report 1648 - University of Wisconsin–Madison. http://burrsettles.com/pub/settles.activelearning.pdf.

Shannon CE (1948). A Mathematical Theory of Communication. The Bell System Technical Journal 27:379–423.

Spearman C (1904). The proof and measurement of association between two things. The American Journal of Psychology 15:72–101. Full publication date: Jan., 1904.

Yeghiazaryan V, Voiculescu ID (2018). Family of boundary overlap metrics for the evaluation of medical image segmentation. Journal of Medical Imaging 5:015006.

Zhang H, Fritts JE, Goldman SA (2003). An Entropy-based objective evaluation method for image segmentation. In: Yeung MM, Lienhart RW, Li CS, eds., Storage and Retrieval Methods and Applications for Multimedia 2004, vol. 5307. International Society for Optics and Photonics, SPIE.

Zhang H, Fritts JE, Goldman SA (2008). Image segmentation evaluation: A survey of unsupervised methods. Computer Vision and Image Understanding 110:260–80.

Zhang Y (1996). A survey on evaluation methods for image segmentation. Pattern Recognition 29:1335–46.