Entropy Role on Patch-Based Binary Classification for Skin Melanoma

Guillaume Lachaud[1,2], Patricia Conde-Cespedes[1,3], and Maria Trocan[1,4]

[1] ISEP - Institut Suprieur dlectronique de Paris.
10 rue de Vanves, Issy les Moulineaux, 92130-France
[2] glachaud@isep.fr
[3] pconde@isep.fr
[4] maria.trocan@isep.fr

Abstract. In this paper, we split the region of interest of dermoscopic images of skin lesions in patches of different size and we analyze the impact of the entropy of the patches on patch-based binary classification using a convolutional neural network (CNN). Specifically, we analyze the distribution of entropy amongst the patches and we compare the training time of a classifier on subsets of the data with varying entropy. We find that the classifier converges faster on patches with higher entropy. Our entropy-based analysis is performed on skin lesion images from the ISIC archive.

Keywords: Entropy · Skin Melanoma · Patch-Based Classification · Resnet

1 Introduction

Convolutional Neural Networks (CNNs) have become one of the most effective machine learning solutions for computer vision problems such as classification, object detection, face recognition, etc. More specifically, CNNs are extensively used for medical image processing tasks and their use in medical research care. The main goal of this field is to extract relevant clinical information or knowledge from medical images. One can mention, for instance, computer-aided diagnosis of cancer using classification methods [1]. Cancer is one of the leading causes of deaths worldwide [14]. However, if the cancer is diagnosed early, when the cancer has not spread, chances of survival are far greater than for later stages [19]. For this reason, there has been a lot of research focused on leveraging deep learning to improve cancer diagnosis and prognosis, especially in breast cancer [22], skin cancer [2], and lung cancer [11].

Typically, image classification tasks take as input the entire image. However, in some situations training an image patch, that is, a subset of the entire image, might be preferable. Not only is this less time consuming, but it can also improve the classifier performance in some particular situations. For instance, in [9], the authors claimed that in cancer subtypes classification, the decision is mostly based on cellular-level visual features

observed on image patch scale. Another example where patch based classification was preferred over pixel based classification is presented in [17] where this approach was used for classification of breast histology. One can find other applications of patch-based classification in [16] and [24].

In information theory, entropy is a measure used to quantify the level of information contained in an object. The higher the entropy, the higher the information content of the image is. For example, an image of random noise will have a higher entropy that a unicolor image. Entropy is indicative of the minimum amount of storage that is required to preserve the full information of an object, which makes this measure particularly useful in data compression to estimate whether the compression algorithm is close to the best possible results [18].

Entropy has been successfully applied to a wide variety of tasks, including image reconstruction where we choose the image with the highest entropy out of all the possible images [20]. Furthermore, applications of maximum entropy are not restricted to images but also extends to other types of data such as text data, in which entropy is used to produce the most uniform probability distribution given the training data [12]. Additionally, entropy has been effectively studied for image texture analysis [25], which can also be used for texture synthesis.

In this paper, we study the role of entropy on the training time of a neural network. We use the region of interest of the image to maximize the relevance of the patches for the classification task. The use of patches instead of the whole image allows us to study the influence of entropy at different scales. To the best of our knowledge, we are the first to analyse the role of entropy on patch-based binary classification. However, it is relevant to mention that in [13], the authors have already focused on entropy for brain tumor patch based classification. Indeed, the authors resized MRI (Magnetic Resonance Imaging) images, split them in patches, and used the entropy of each patch as a feature for the image as well as the image moments.

The dataset we used in this study comes from the ISIC [5] archive (International Skin Imaging Collaboration). Because of the lethality of melanoma cancer, the ISIC project was created to help improve skin cancer diagnosis via imaging data. They started an annual challenge in 2016 [7], and from 2019 onwards, the challenges have focused on dermoscopic image classification, with multiple diagnostic categories [15]. The researchers who had the best results on the 2019 ISIC challenge [4] studied patch-based classification on the HAM10000 dataset [21] in [5]. They took multiple patches from each image and used an attention-based approach to combine the information from the patches and classify the image.

The paper is organized as follows: in Section 2 we describe the datasets and data pre-processing, analyze their entropy and we introduce the

[5] The data is publicly available at `https://www.isic-archive.com`

network architecture we used. Next, Section 3 shows the experimental results. Finally, Section 4 presents the conclusion and perspectives of this work.

2 Proposed method

2.1 Dataset description and pre-processing

The ISIC archive database (see [15]) contains images of skin lesions which can be benign or malignant; other images can also have an unknown status. The image resolution varies across the datasets. The archive also has an API [6] which can be used to get information about images or to retrieve lesion masks created by expert users. Our goal is to perform binary classification using patches of images. Our target variable has two labels[7] indicating whether the lesion is *benign* or *malignant*.

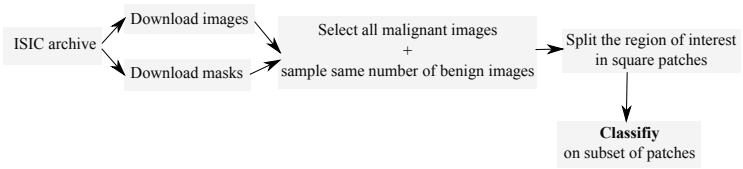


Fig. 1. Data pre-processing workflow

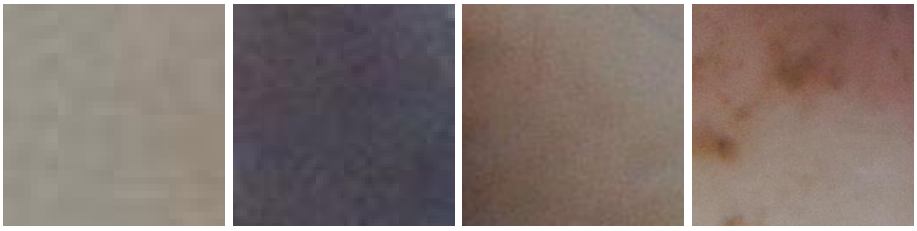
(a) Image (b) Mask

Fig. 2. Example of a malignant skin lesion

All the data pre-processing steps are described in Figure 1:

[6] `https://isic-archive.com/api/v1`

[7] Originally, the ISIC challenge had more refined categories. In this paper we use only 2.



(a) 32×32 (b) 64×64 (c) 128×128 (d) 256×256

Fig. 3. Example of patches of size (a) 32×32, (b) 64×64, (c) 128×128, (d) 256×256, from Figure 2

1. First, using the API from the ISIC archive, we download all the images which have a mask.
2. Second, we take all the malignant images and we sample the same quantity of benign images from all the benign images with a mask.
3. Next, for each image, we take the region of interest, that is, the part where the lesion is, which is obtained from the mask, and we split this region in square patches of size 32×32, 64×64, 128×128 and 256×256. Figure 2 shows an example of an image and its mask, with patches of different size taken from the same image in Figure 3. Table 1 indicates the total number of patches for each patch size. Since we took the same number of malignant and benign images, we have an imbalanced dataset with more malignant patches than benign ones. This is due to the fact that malignant lesions are often captured in higher resolution, because it is more important to get the best image quality when analyzing cancerous lesions than it is for benign ones.
4. Finally, we perform binary classification on patches.

Table 1. Number of patches for different patch sizes

patch size	number of patches
32×32	4,886,969
64×64	1,173,052
128×128	270,821
256×256	58,253

2.2 Entropy

We are interested in the study of the behavior of the Shannon entropy [18] of the images. The formula used for the calculation of entropy is the

following:

$$H = -\sum_{k=0}^{M} p_k \log_2(p_k) \tag{1}$$

where M is the highest intensity of a pixel (in our case, 255), and p_k is the probability associated with the pixel intensity k in the grayscale image. In practice, the entropy is computed using histograms to estimate the probabilities. The entropy can take values between 0 and $\log_2(255) \approx 8$. Although the images in the dataset are in the RGB format, the entropy is computed on the grayscale version of the images. Our choice was motivated by the fact that there is no consensus on how to compute the entropy of an RGB image: Equation 1 does not have a canonical generalization to RGB images, while RGB conversion to grayscale is standardized in the ITU-R Recommendation BT.601-2.

Figure 4 shows the distribution of entropy amongst the patches for different patch sizes.

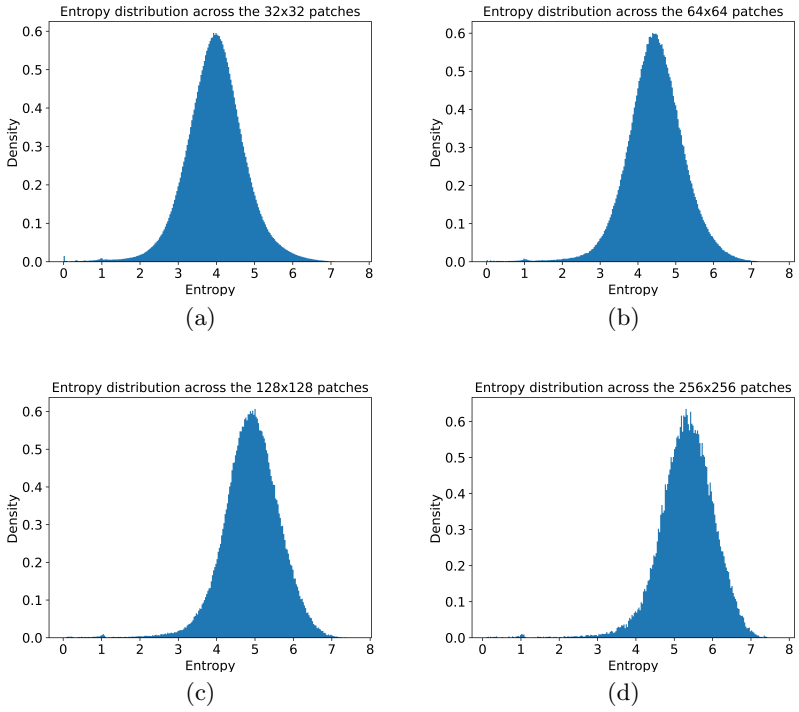


Fig. 4. Distribution of patch entropy. (a)-(d) are taken for square patches of size 32, 64, 128 and 256 pixels.

Table 2 shows the mean, standard deviation and some quantiles of entropy. We observe that, as the patch size grows, so does the entropy. This is expected because the more pixels we have, the more likely they are to have different intensities, which lead to a higher entropy. Also, the entropy for bigger patch sizes is slightly more centered around the mean, which may be due to the fact that bigger patch sizes will average some of the more extreme patches of smaller size. For example, instead of having multiple small patches of low and high entropy, a bigger patch containing all the small patches will have a more average entropy.

Table 2. Entropy statistics

			quantile			
patch size	mean	standard deviation	15	42.5	57.5	85
32	3.974	0.779	3.247	3.85	4.104	4.71
64	4.456	0.765	3.75	4.335	4.588	5.191
128	4.903	0.747	4.223	4.795	5.047	5.633
256	5.319	0.735	4.66	5.229	5.475	6.029

We are interested in the impact of the entropy behaviour on the training of a classifier: whether it is faster to train on a dataset with low entropy than with a dataset with standard entropy; and whether a dataset with higher entropy is harder to train on. We split the created patches in three groups for the four groups of patches:

- one containing the patches with entropy below the 15-th quantile, referred to as *low*.
- one with the patches entropy above the 85-th quantile, referred to as *high* and
- the last one with patches having entropy between the 42.5-th and 57.5-th quantiles, referred to as *intermediate*. Our choice for the quantile values is motivated by having the entropy be equally distant from the other groups, and keeping the same number of samples to make time comparisons meaningful.

We do this for each patch size, e.g. 32×32, 64×64, 128×128, 256×256.

2.3 Network architecture and tuning parameters

Following [23] and [3] who compared classifiers for the same task and dataset, we use a ResNet50 for the classification. ResNet50 [8], is a 50-layer convolutional neural network, which contains *residual units* between convolutional blocks (stacks of convolutional layers) with identity mappings interspersed, to help propagate the gradient and mitigate the problem of vanishing and exploding gradients [6].

Though ResNets can be arbitrary deep, provided we have the computing resources to train the model, e.g. using 101 or 152 layers, we followed [23] and used the 50-layer version. Since we are interested in binary classification, e.g. whether the lesion is benign or malignant, we remove the last layer of the network, designed for multiclass classification, and replace it with a max pooling layer followed by a Dense layer with a *sigmoid* activation.

The optimizer used for the model is the Adam optimizer [10] with a learning rate of 0.001. We use a *binary cross-entropy loss* for the training.

The model is trained for 10 epochs, with early stopping if the validation loss stops decreasing after 3 consecutive epochs.

Each dataset is split in the following way for training: 90% for training, of which 20% goes to validation, and 10% for testing.

3 Results

All the experiments were performed on a device with a 3.60 GhZ Intel CPU, 32Gb of RAM and an NVidia Titan XP, running on Ubuntu. The code was written in Python and Tensorflow. The computation of the entropy was done using Pillow.
To account for the fact that a neural network may take more time to converge based on the random initialization of the parameters, we train 10 instances of a ResNet50 on each dataset. We display the 30-th quantile, the median and the 70-th quantile of the training time of the instances in Table 3.
We see that that the dataset with the highest entropy tends to be the fastest to converge. Since a higher entropy usually indicates that more information is present in the patch, we could expect the neural network to take longer to train. Conversely, a dataset with lower entropy would train faster because the patches would have less discriminating features, and the network would quickly classify them.
A possible explanation for this discrepancy is that patches with higher entropy might share a similar structure or have patterns not present for other patches, and thus are more recognizable by the network, while patches with lower entropy might have less salient features, which makes it harder for the classifier to classify them.

Concerning the training for the dataset with intermediate entropy, it seems to take longer to converge for smaller patch sizes compared to training on datasets with more extreme entropy, but reaches similar speeds in comparison with the other datasets when we increase the patch size. A reason for this could be that, for lower patch sizes, patches with average entropy might be more diverse than patches with lower or higher entropy, and the network will require more time to analyze the patterns.

Table 3. Quantiles for the training time on datasets with varying entropy

patch size	entropy	Quantile of training time (in seconds) 30	50 (median)	70
32	high	1350.7	2013.2	2781.4
32	intermediate	2000.4	2854.0	2855.3
32	low	1534.9	2906.7	3078.5
64	high	291.0	382.9	441.9
64	intermediate	331.0	402.3	498.8
64	low	290.6	338.3	414.2
128	high	155.0	204.6	220.0
128	intermediate	174.6	235.2	281.3
128	low	204.8	255.0	255.4
256	high	142.4	152.2	189.7
256	intermediate	171.3	171.8	204.8
256	low	189.6	226.4	226.5

When the patches are bigger, a patch can be composed of smaller zones which vary greatly in entropy, but have an average entropy when we look at the entirety of the patch. Therefore, these patches would be easier to classify, which would lead to a faster training time.

4 Conclusion and future works

We studied the influence of entropy on the training time of a convolutional neural network applied to patch-based classification. We found that the CNN converges faster when using datasets with higher entropy, which might be due to the presence of patterns on these patches the network can detect. We also observe that performance of datasets with average entropy tend to improve, in comparison with the other datasets, when the patch size increases.

Some perspectives to this work can be to explore the use of segmentation to obtain the regions of interest, increasing the number of images we can work with, and see if the results are comparable. Another possibility can be to analyze the effects of resizing images on their entropy to quantify the loss of information, and the impact it can have on classification using resized images.

References

1. Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K.: Medical Image Analysis using Convolutional Neural

Networks: A Review. Journal of Medical Systems **42**(11), 226 (Nov 2018). https://doi.org/10.1007/s10916-018-1088-1
2. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (Feb 2017). https://doi.org/10.1038/nature21056
3. Favole, F., Trocan, M., Yilmaz, E.: Melanoma Detection Using Deep Learning. In: Nguyen, N.T., Hoang, B.H., Huynh, C.P., Hwang, D., Trawiński, B., Vossen, G. (eds.) Computational Collective Intelligence, vol. 12496, pp. 816–824. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-63007-2_64
4. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A.: Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX **7**, 100864 (2020). https://doi.org/10.1016/j.mex.2020.100864
5. Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., Schlaefer, A.: Skin Lesion Classification Using CNNs With Patch-Based Attention and Diagnosis-Guided Loss Weighting. IEEE Transactions on Biomedical Engineering **67**(2), 495–503 (Feb 2020). https://doi.org/10.1109/TBME.2019.2915839
6. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, D.M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010. JMLR Proceedings, vol. 9, pp. 249–256. JMLR.org (2010)
7. Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1605.01397 [cs] (May 2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.90
9. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2424–2433 (2016). https://doi.org/10.1109/CVPR.2016.266
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
11. Marentakis, P., Karaiskos, P., Kouloulias, V., Kelekis, N., Argentos, S., Oikonomopoulos, N., Loukas, C.: Lung cancer histology classification from CT images based on radiomics and deep learning models. Medical & Biological Engineering & Computing **59**(1), 215–226 (Jan 2021). https://doi.org/10.1007/s11517-020-02302-w
12. Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text classification. In: IJCAI-99 workshop on machine learning for information filtering. vol. 1, pp. 61–67. Stockholom, Sweden (1999)

13. Ouchtati, S., Chergui, A., Mavromatis, S., Aissa, B., Rafik, D., Sequeira, J.: Novel Method for Brain Tumor Classification Based on Use of Image Entropy and Seven Hu's Invariant Moments. Traitement du Signal **36**(6), 483–491 (Dec 2019). https://doi.org/10.18280/ts.360602
14. Ritchie, H., Roser, M.: Causes of death. Our World in Data (2018)
15. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., Soyer, H.P.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data **8**(1), 34 (2021). https://doi.org/10.1038/s41597-021-00815-z
16. Rousseau, F., Habas, P.A., Studholme, C.: A supervised patch-based approach for human brain labeling. IEEE Trans. Medical Imaging **30**(10), 1852–1862 (2011). https://doi.org/10.1109/TMI.2011.2156806
17. Roy, K., Banik, D., Bhattacharjee, D., Nasipuri, M.: Patch-based system for Classification of Breast Histology images using deep learning. Comput. Medical Imaging Graph. **71**, 90–103 (2019). https://doi.org/10.1016/j.compmedimag.2018.11.003
18. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
19. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer Statistics, 2021. CA: A Cancer Journal for Clinicians **71**(1), 7–33 (Jan 2021). https://doi.org/10.3322/caac.21654
20. Skilling, J., Bryan, R.: Maximum entropy image reconstruction-general algorithm. Monthly notices of the royal astronomical society **211**, 111 (1984)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 180161 (Dec 2018). https://doi.org/10.1038/sdata.2018.161
22. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology **292**(1), 60–66 (Jul 2019). https://doi.org/10.1148/radiol.2019182716
23. Yilmaz, E., Trocan, M.: Benign and Malignant Skin Lesion Classification Comparison for Three Deep-Learning Architectures. In: Nguyen, N.T., Jearanaitanakij, K., Selamat, A., Trawiński, B., Chittayasothorn, S. (eds.) Intelligent Information and Database Systems, vol. 12033, pp. 514–524. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-41964-6_44
24. Zhang, F., Song, Y., Cai, W., Lee, M.Z., Zhou, Y., Huang, H., Shan, S., Fulham, M.J., Feng, D.D.: Lung nodule classification with multilevel patch-based context analysis. IEEE Transactions on Biomedical Engineering **61**(4), 1155–1166 (2014). https://doi.org/10.1109/TBME.2013.2295593

25. Zhu, S.C., Wu, Y.N., Mumford, D.: Minimax entropy principle and its application to texture modeling. Neural Comput. **9**(8), 1627–1660 (1997). https://doi.org/10.1162/neco.1997.9.8.1627, https://doi.org/10.1162/neco.1997.9.8.1627