Encoding Skin Cancer Images for Case Retrieval using Siamese Networks

Helmut Neher
University of Waterloo
hneher@uwaterloo.ca

***Abstract*—Current challenges in the medical domain, specifically for dermatologists is accurately assessing patient information in an accurately and fast manner. Currently, a doctor's standard method of identifying malignancy in skin cancer is to search a book, which, is prone to human error as well as slow. To improve the accuracy and speed of correctly identifying skin lesion cases, this project explores the possibility of encoding images in an online database for accurate and quick retrieval. Using a siamese neural network, the architecture is trained to learn dissimilarities between malignant and benign skin lesions, via RGb images and encoded in a vector. The method achieves a best mAP score of 0.5 and average mAP score of 0.497 showing that the case retrieval can successfully retrieve related images, however, the PR curve and the t-sne plot both demonstrate the benign and malignant skin lesions are being picked virtually randomly due to the fact that the images, based on the network architecture, are not distinguishable as separate classes, rather, they do not appear to be linear separable. The network does prove that images can be encoded and maintain information.**

I. Introduction

In doctor-patient scenarios, assessing patient information is critically important for diagnosing. For dermatologists, more information may need to be provided. For example, when assessing a skin lesion to determine if it is malignant or benign, the doctor may review past cases, or consult a book of skin cancer images to determine, based on lesion examples, whether the lesion is malignant or benign. Information needs to be assessed on the spot in a fast and accurate way. In most cases, malignant and benign images are very similar in color, symmetry, and border in many instances and can't be correctly identified via visual inspection of the skin surface via a trained professional. In fact, to determine the malignancy of a skin lesion, a patient may experience a lengthy process where eventually the skin lesion is incised to determine malignancy. If incision does not occur, a higher risk of misdiagnoses occurs. To ensure that diagnosis is non-invasive, accurate and fast, this paper addresses a novel method of improving diagnoses by providing a doctor a case retrieval method that provides accurate examples of similar skin lesions in a non-invasive, fast and accurate manner. To improve the image retrieval process of skin cancer images, this work presents a method of categorizing skin cancer images of various pigmentation and degrees of skin cancer using an encoding sequence.

The main ideas of this paper is the use of a siamese network with a contrastive loss function that encodes images into a sequence of values in order to improve image retrieval, specifically for cancerous skin. Using the ISIC dataset [1], a dataset of skin cancer images, the architecture, consisting of convolutional and fully connected layers is used to train a network to compare benign and malignant images in order to learn the disparity between malignant and benign skin cancer images to better encode images in order for faster and accurate retrieval.

II. Background

Earlier techniques prior to convolutional neural networks for image retrieval and hashing usually included bag-of-feature representations with large vocabularies and inverted files [2], [3] or encoding techniques such as Fisher Vectors [4] or VLAD [5]. As convolutional networks research became popular, image retrieval techniques gravitated toward classification via hand-crafted features, and employing cross-matching [6], sum-pooling to improve descriptors for image retrieval [7] or employing region proposal networks to understand global information as well as local information in an image based on a Siamese network using triplet loss [8].

Current image hashing utilizes many techniques that attempt to improve speed of image retrieval, generalization, and precision. Lin *et al*. [9] employs a latent softmax layer that learns features which are optimized for image classification. Xia *et al*. [10] use a two-step function that that learns binary codes for all training data in the first step and then learns hash functions on the basis of the learned codes in a second step. Other research involves learning hashing schemes based on initial learning via stacked RBMs and then fine-tuned using and a Siamese Network [11].

Other networks employ a triplet loss ideal. Nguyen *et al*. [12] implement a triplet loss function to minimize the Hamming distance between the neighbor pairs with a relaxed empirical penalty. Lai *et al*. [13] presents a divide-and-encode module to divide the intermediate image features into multiple branches, where each encoding is in one hash bit, then a triplet loss is fine-tuned to the network. Liu *et al*. [14] implements a Siamese network with a hamming distance loss function and L2 normalization.

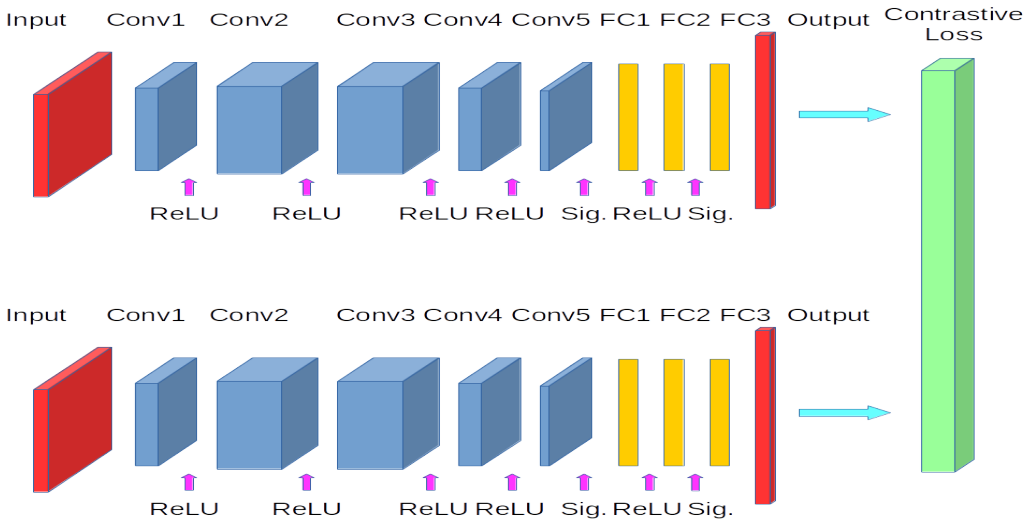

Figure 1: The cancer similarity network employs two streams, known as twins, that share the same weights. In this siamese network, five convolutions are performed with three fully connected layers following. The input is an RGB image and the output is a vector of size *1x5*. The loss function to perform back propagation is a constrastive loss that takes output of each twin and determines based on the output values the similarity between each image, penalizing images that appear to be similar, but are of differing classes.

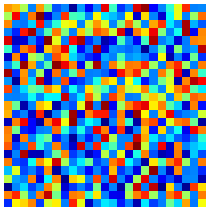

Figure 2: Visual of the encoding sequence. The coding sequence is a *25x25* vector where values range from zero to one.

III. Methodology

A. Architecture

The architecture of this network is inspired by Harshvardhan Gupta in a tutorial [15] that introduced Siamese networks. Siamese networks are where two networks are used to evaluate different images and are joined in some fashion. In Figure 1, the networks are identical, they share the same weights and the information from the output of the network is combined to compute the loss function.

The general structure of each twin composes of five convolutional layers and three fully connected layers. Following each convolutional layer, except the last convolutional layer, a ReLU layer follows with batch normalization. In the fully connected (FC) layer portion of the network, the after the first FC layer, there is a ReLU activation function followed by batch normalization. In between the second and third layer, the activation function is sigmoid. Sigmoid functions are used to provide an estimate between the values of 0 and 1, which, in binary classification and image retrieval, we want the classification and evaluation to be as close to 0 or 1. The input to the network is of *3x100x100* size, while the output of the network is *1x5*.

Once both twins have completed a forward pass, both outputs of vector size *1x5*, we apply a contrastive loss in the form of:

$$loss = (1 - Y)\frac{1}{2}(D_w)^2 + (Y)\frac{1}{2}[max(0, m - D_w)]^2$$

where D_w is the euclidean distance from the output vector from each twin.Y is the label of the image compared and m is the margin of loss to include in the network, and is greater than zero. The metric m is important because it indicates that dissimilar pairs beyond this margin will not contribute to the loss, which makes sense because we only want to optimize the network based on pairs that are actually dissimilar, but that the network thinks are fairly similar. The objective of the contrastive loss function in the siamese architecture is to differentiate between images. Intuitively, the contrastive function evaluates how well the network is distinguishing a given pair of images.

To get the image sequence as shown in Figure 2, the fully connected layers are popped off and the last convolutional layer is pooled to obtain a *25x25* image size. The sequence contains values from zero to one and are color coded to illustrate the values.

B. Evaluation Metrics

There are three forms of evaluation metrics that will be used to understand the efficacy of the method presented. Firstly, to understand the linear separability of the problem, the images will be evaluated based on their conversion to their respective feature map and evaluated using t-distributed stochastic neighbor embedding (t-SNE). Secondly, a precision and recall curve will be used for only 10 samples of the data (anything more would take a very extensive time to evaluate). Thirdly, the mean average precision (mAP) will be evaluated for 10 images and their top 10 similar images will be evaluated.

1) t-sne: T-SNE is an unsupervised machine learning algorithm that is specifically used for visualizing high dimensional data in a low dimensional space, which was first used by van der Maaten *et al*. [16]. More specifically, it models multiple dimensions as a two dimensional point, where related objects are modeled as nearby points and dissimilar objects are modeled as being farther away.

There are two stages of t-SNE. First, t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that dissimilar objects have a lower probability of being picked and similar objects have a higher probability of being picked. Second, t-SNE creates a similar probability distribution over the points in low-dimensionality space and minimizes the Kullback-Liebler divergence between the two distributions.

The idea of t-sne is to visually inspect the two dimensional points to determine if the malignant or benign objects are linearly separable as linear separability shows that each class has identifiable features independent of the other class. T-SNE will evaluate the proposed encoding to determine linear separability.

2) Precision and Recall: A precision and recall curve is the relationship, at certain thresholds, that an algorithm has a precision and recall.

In information retrieval, precision is denoted by:

$$Precision = \frac{\textit{\# of correct results}}{\textit{\# of returned results}}$$

and recall is denoted by:

$$Recall = \frac{\textit{\# of correct results}}{\textit{\# of results that should have been returned}}$$

A good PR curve will have a precision of one with no recall and then will curve downward where recall is equal to one and precision 0; precision and recall tend to be inversely proportional. A poor PR curve will have a relatively flat horizontal curve indicating that precision does not improve at from zero to one recall. Note that the recall and precision evaluated are in fact the micro average recall and the micro average precision, where the PR curve for each case are computed and then averaged.

3) mAP: Mean average precision (mAP) is an important case retrieval metric and is denoted by:

$$mAP = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP + FN)_i}$$

where n is the number of samples to sum, TP is true positives and FN false negatives. The idea is to determine the precision of your top 10 (or however estimated) choices that are closely related, and determine the precision by adding all of the number of correct results (true positives) divided by the total images retrieved (*TP + FN*). In this scenario, 10 random samples, 5 malignant and 5 benign, from the test set will be selected as cases and all other images from the test set will be retrieved and given a dissimilarity score. Then the top 10 images that the algorithm closely associates to the image case in question will be used to create the mAP score. Note that for this metric, since we have an equal amount of images that are malignant and benign, an excellent mAP score will be .5, indicating that 50% of images retrieved were benign whilst the other 50% of images retrieved are malignant. Mean average precision puts a heavier weight on the cases that are more recommended (first place) than the recommended images that are less relevant (last place).

IV. Results

Evaluation can first be demonstrated using the t-SNE visualization as depicted in Figure 4. The plot is used to understand whether the encoded images are linearly separable. T-SNE plots the skin lesions in 2D points and based on how close each point is located on the figure is how 'similar' the network perceives those images to be. From the data, malignant images are depicted with a purple point and benign with a yellow. From the plot, the two classes are not linearly separable. In other words, from the network classifying the images, it was not able to effectively distinguish between malignant and benign images effectively..

The second evaluation is presented by applying the precision and recall by visualizing the PR curve as seen in Figure 5. The curve has two plateau regions, a higher one where recall equals to zero and a lower region where recall equals to 0.25. Below that threshold, the precision hovers at around 0.7, where after the 0.25 threshold, the precision hovers at around 0.5. The plateau, or horizontal PR curve demonstrates that the case retrieval method performs randomly; the network does not effectively correctly labels images in the correct class.

The mAP score is generated by evaluating five malignant and five benign cases and retrieving the top 10 similar cases.

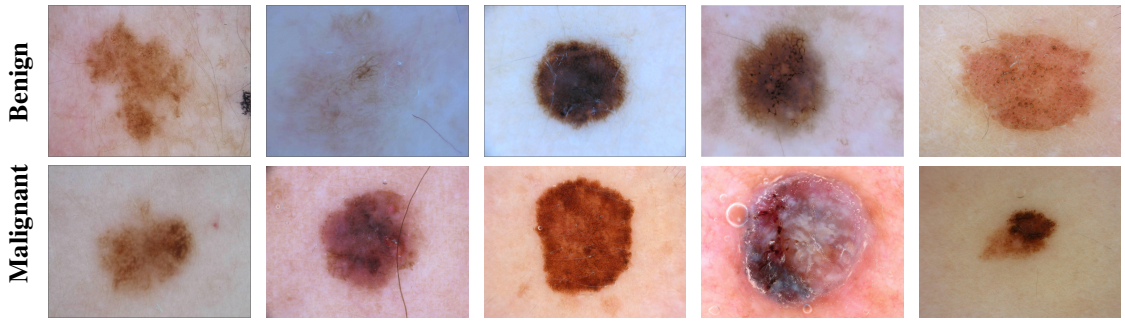

Figure 3: Showing various images in the ISIC Dataset used for training. In some instances, the color, border, symmetry of benign and malignant cases are very similar.

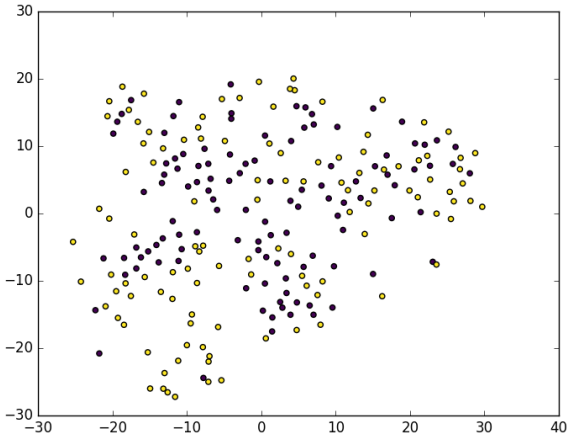

Figure 4: t-SNE plot depicting the benign and malignant skin lesions as 2D points in the plot. The plot demonstrates that the benign(yellow) and the malignant(purple) are not linearly separable.

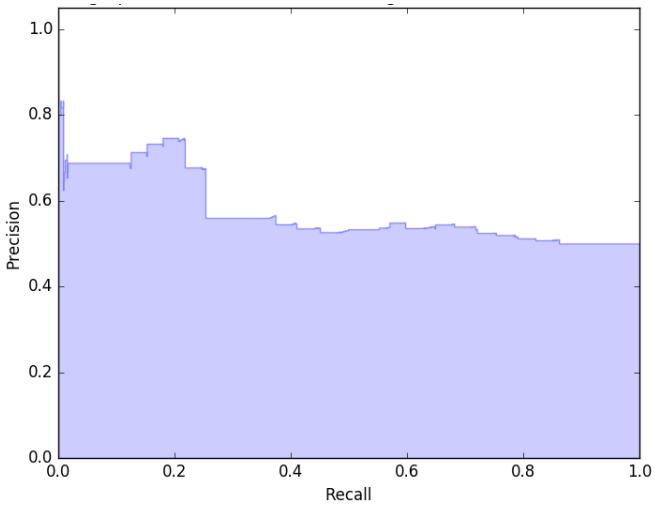

Figure 5: Precision and Recall curve. Overall, the curve is fairly horizontal, however, you do see somewhat of a jump at the beginning of the PR curve where recall equals to one.

This test was evaluated 10 times with the results for this network in Table **??**. From the table, the average mAP score is 0.497 while the best mAP score is 0.5. This means that out of 10 case retrievals, where an equal amount were retrieved for malignant and benign, 50% of those images retrieved were malignant and 50% were benign. This experiment shows the efficacy of the network.

A. Dataset

The dataset is from the International Skin Imaging Collaboration (ISIC) which is a dataset consisting of over 20,000 images from leading clinical centers internationally, acquired from a variety of devices used at each center. The dataset consists of diagnosis, notes and pathology reports of each skin lesion. This paper only utilizes a portion of the dataset. This portion is comprised of a random selection of 1,594 cropped images. A sub sample is shown in Figure 3. The dataset is portioned into train and test sets that are cropped.

Test	mAP
1	.54
2	.61
3	.42
4	**0.5**
5	0.44
6	0.44
7	0.49
8	.48
9	0.61
10	0.44
Avg	**0.497**

Table I: Mean average precision scores for 10 results. The best score is test 4, while the average is 0.497. The score was calculated by randomly sampling 5 benign cases and 5 malignant cases with their top 10 images in the test set evaluated and contributed to the scoring.

The training set consists of 684 benign, and 684 malignant skin cancer images. The test set consists of 113 of benign images and 113 malignant images.

The test and train sets were small because the dataset needed to be pre-processed, which as a result, not all of the images were used. The pre-processing procedure consisted of cropping the images due to different noises within the image. In this dataset, there were 5 different pieces of unneeded information: hair, ink, lens appearance, markers and rulers as shown in Figure 6. The quality of the images were cropped to maintain mole structure and color, however, as much noise was eliminated to reduce the network's reliance on noise information. In addition, the structure and color of the mole were held to a higher priority, so if there was a conflict between cropping noise vs structure and color, structure and color were kept, thereby allowing minimal noise to pervade the dataset. Another reason why the dataset is small is because there was a huge disproportion of benign to malignancy which effectively skewed training. To keep everything consistent, I decided to make the malignant and benign classes have the same amount of images, making the dataset very small.

V. Conclusion

This paper has shown a siamese network made up of identical twins compose of convolutional layers and fully connected layers with a contrastive loss function. From this network, we evaluated the efficacy using three metrics including t-SNE visualization, PR curve visualization and mAP scores. From these results, it can be concluded that the network itself, needs to be improved. Although the algorithm was able to achieve a near perfect case retrieval as indicated in the mAP score, it does so randomly. In other words, the network, thus far, cannot effectively distinguish between what is malignant and what is benign for skin lesions, as indicated in the first two tests. On a positive note, we have shown that by encoding images in a sequence, we can retrieve images that are of the same class as the test case. In addition, encoded images are shown to be a reflection of the original images; coding an image retains its information.

For future works, the following will be tested and evaluate in order to develop a better network that can distinguish binary objects: testing triplet networks, evaluating a better loss function, driving the image sequence to be more binary, while also reducing the image features to at most to a vector of size *1x32*. In addition, more data will need to be used to train a more robust network.

References

[1] D. Gutman, N. Codella, M. E. Celebi, B. Helba, M. Marchetti, N. K. Mishra, and A. Halpern, "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)," 05 2016. [Online]. Available: http://arxiv.org/abs/1605.01397

[2] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 2161–2168.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[4] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[5] H. Jgou, M. Douze, C. Schmid, and P. Prez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3304–3311.

[6] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2014, pp. 512–519.

[7] A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1269–1277.

[8] J. R. D. L. A. Gordo, J. Almazan, "Deep image retrieval: Learning global representations for image search," in *2016 IEEE European Conference on Computer Vision (ECCV)*, Apr 2016.

[9] K. Lin, H. F. Yang, J. H. Hsiao, and C. S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2015, pp. 27–35.

[10] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, ser. AAAI'14. AAAI Press, 2014, pp. 2156–2162. [Online]. Available: http://dl.acm.org/citation.cfm?id=2892753.2892851

[11] J. Lin, O. Morère, A. Veillard, L.-Y. Duan, H. Goh, and V. Chandrasekhar, "Deephash for image instance retrieval: Getting regularization, depth and fine-tuning right," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '17. New York, NY, USA: ACM, 2017, pp. 133–141. [Online]. Available: http://doi.acm.org/10.1145/3078971.3078983

[12] V. A. Nguyen and M. N. Do, "Deep learning based supervised hashing for efficient image retrieval," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, July 2016, pp. 1–6.

[13] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3270–3278.

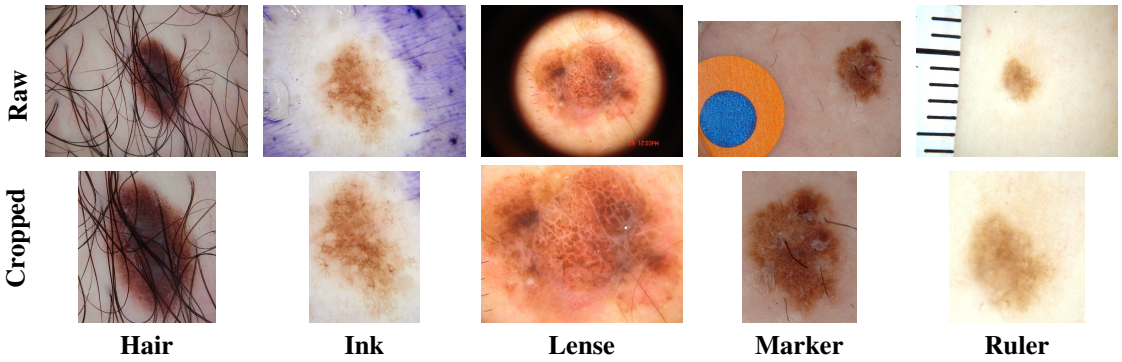

Figure 6: Showing various images that were pre-processed to remove noise in the dataset used. Figure shows various raw images and their cropped counterparts. Typical noise in the dataset consists extensive amounts of hair, ink stains, lens appearances, marker appearances, and ruler appearances.

[14] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2064–2072.

[15] H. Gupta. (2017) Facial similarity with siamese networks in pytorch. [Online]. Available: https://hackernoon.com/facial-similarity-with-siamese-networks-in-pytorch-9642aa9db2f7

[16] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9: 25792605, Nov 2008.