Improving CNN Training using Disentanglement for Liver Lesion Classification in CT

Avi Ben-Cohen1 Roey Mechrez2 Noa Yedidia1 Hayit Greenspan1

Abstract—Training data is the key component in designing algorithms for medical image analysis and in many cases it is the main bottleneck in achieving good results. Recent progress in image generation has enabled the training of neural network based solutions using synthetic data. A key factor in the generation of new samples is controlling the important appearance features and potentially being able to generate a new sample of a specific class with different variants. In this work we suggest the synthesis of new data by mixing the class specified and unspecified representation of different factors in the training data which are separated using a disentanglement based scheme. Our experiments on liver lesion classification in CT show an average improvement of 7.4% in accuracy over the baseline training scheme.

Index Terms—Disentanglement, Medical, Synthesis, Liver lesions.

I. INTRODUCTION

A neural network trained for image classification is considered to be a black-box model to the user, as one can not identify what are the factors that contribute to a specific classification. Access to such information can reveal new insights and assists in classical medical practice of observing images.

The concept of disentangling in the context of deep representation refers to separation of factors in the representations [1]. The key idea is to separate the representation into complementary code which is associated with distinctiveness, i.e. distinguish the specified factors and the unspecified factors of the data. For example, if we would like to distinguish between different faces, the shape might be an unspecified factor where the appearance would be a specified factor. The main goal of disentanglement is to allow the learning process to represent the information as two complementary vectors without constraining the latent space. As a result, the specified and unspecified factors definition is not accessible. Mathieu et al. [1] suggested to solve it by adversarial training, where label information is removed from the unspecified part of the encoding. In Hadad et al. [2] a simpler two-step disentanglement method was suggested: an encoder is first trained to satisfy a classifier, then this encoder is used in order to synthesize a new image with the same classification features but with different unspecified factors.

In this paper we follow [2] and use it as a pipeline for disentangling factors in the image representation of liver lesions. The separation is based on the relevance of a specific feature to the classification process of the image. For example, the contrast and texture can be significant factors for classification while the discriminative power of the general shape can be less important. We next use it in the training process by adding synthesized examples that include a mixture of the specified factors taken from different liver lesion types. By mixing the specified factors we can synthesize “difficult samples” which may look like a mixture of two classes or more, the label is updated accordingly taking into account the portion of each class in the mixture. Figure 1 shows an example by combining the unspecified representation of a cyst and mixing it with the specified representation of an hemangioma to create a new sample, in this case the specified factors of the cyst were completely replaced by these of the hemangioma.

We focus our experiments on liver lesions classification in Computed tomography (CT) images. The liver can include multiple lesions from several types which makes it a complex task. Several steps in automating the liver analysis can be found in recent literature, including liver segmentation, lesion detection, lesion categorization and follow-up. In the lesion categorization task it is important to separate between malignant and benign lesions. Frid-Adar et al. [3] showed that generating synthetic medical data using generative adversarial networks [4] can improve liver lesion classification results. In this study we present an alternative approach to generate mixed samples within our dataset that
challenge the training process and improve the classification system overall accuracy.

The dataset used includes 4 categories: cyst, metastasis, hemangioma and healthy liver parenchyma. After training our disentangling network, we used the synthetic samples in order to augment the training data. Our experiments show that the proposed synthetic data augmentation improves the baseline results in terms of classification accuracy.

The contributions we present are two-fold:
1) To the best of our knowledge this is the first attempt to disentangle factors in medical images.
2) We present a novel pipeline for synthesizing new data for training by swapping the specified and unspecified components of different liver lesion types.

II. METHOD

Our proposed method includes two major components: the first one is the disentanglement process and the second is the robust training process that makes use of the disentanglement to synthesize new examples with corresponding categorical vectors. In the following we elaborate on both components.

A. Disentanglement of Liver Lesions

Figure 2 shows the general disentanglement pipeline used in this study. Similar to [2] our training process consists of two steps. In the first step an encoder $E_C$ is trained using the classic cross entropy classification loss. In the second step, we keep $E_C$ fixed and train an additional encoder termed $E_R$. $E_R$’s output is concatenated with $E_C$’s output. The concatenated vector, $z = (E_C(x), E_R(x))$, is then used as input to a decoder termed $G_R$ that is trained to reconstruct the original image using $L_2$ loss:

$$L_{rec} = \|G_R(z) - x\|_2^2$$  \hspace{1cm} (1)

In addition, an adversarial classifier is used to make sure the code derived from $E_R$ is not discriminative in terms of liver lesion classification. Hence, our overall loss function used in the second training step is as follows:

$$\mathcal{L}(E_R, G_R) = L_{rec} + \lambda \cdot L_{adv}$$ \hspace{1cm} (2)

where $\lambda$ is set to be negative in order to achieve the desired separation of the representation and $L_{adv}$ is the cross entropy loss.

B. Mixture of Specified Factors for Improved Classification

Using the encoders from the disentanglement process we can represent each image using the specified vector ($c$) and the unspecified vector ($r$). One possible manipulation would be to swap the vector $c$ of one image with the vector $c$ of another image. Figure 3 shows a swap example using the $r$ vector of a region of interest (ROI) of a cyst and swapping its $c$ vector with vectors that were extracted from different lesion types. Next, the $c$ and $r$ vectors are concatenated and $G_R$ is used to construct a new image.
Similar to the swapping example, we suggest a mixture of specified factors for training. For each image in the training process we extract its \( r \) vector and create a new \( c \) vector that includes a mixture of the original \( c \) with other vectors from the different classes. The mixture is a sum of all vectors with different random proportion of each one. This process is used to challenge the model training with “difficult samples” which are a linear interpolation of the different classes over the factors used for classification. In addition, we wanted to mix samples that are taken from different classes but have a relatively similar specified factors so we can focus on the harder samples in the dataset and generate a mixture of those. To achieve that the different \( c \) vectors are chosen as those that are most similar (in terms of \( L_2 \)) to the original one to include a mixture \( c \) vectors that are hard to distinguish but derive from different classes. The target classification vector is based on the proportion of each class in the mixture. For example, if a portion of 0.8, 0.1, 0.1, and 0.0 was used to interpolate the specified factors of a cyst, hemangioma, metastasis, and healthy image respectively, the target label vector will be \([0.8, 0.1, 0.1, 0.0]\). Figure 4 illustrates the proposed classification training scheme. Note that since the new generated samples are a mixture of different lesion classes we include the mixture in the target class probability vectors.

**C. Networks Architecture**

Three network architectures were used in this study: 1) for both encoders; 2) for the decoder; 3) for the classifier and the adversarial classifier. Table I provides a description of the models.

**III. EXPERIMENTS AND RESULTS**

**A. Dataset**

The dataset used in this work contains cases of liver lesions collected from the Sheba Medical Center by searching medical records for cases of cysts, metastases and hemangiomas. Cases were acquired from 2009 to 2014 using four CT scanners: a General Electric (GE) Healthcare scanner, a Siemens Medical System scanner, a Philips Healthcare scanner and Toshiba Healthcare scanner, with the following parameters: 120\( kVp \), 120 – 630\( mAs \) and 0.9 – 5.0\( mm \) slice thickness. Cases were collected with the approval of the institutions Institutional Review Board. The input to our system are ROIs of lesions cropped from CT scans using the radiologists annotations. The ROIs are extracted to capture the lesion and its surrounding tissue relative to its size. Due to the large variability in lesion sizes, this results in varying size ROIs, afterwards all ROIs are resized to 128 \( \times \) 128. The dataset was made up of 239 portal phase 2-D CT scans: 66 cysts, 81 metastases, 65 hemangiomas, 27 healthy. An expert radiologist marked the margin of each lesion and determined its corresponding diagnosis which was established by biopsy.

**Table I**

| Encoder       | Classifier   | Decoder       |
|---------------|--------------|---------------|
| C64K3S2       | Dense 256    | Dense 262144  |
| C256K3S2      | BatchNorm    | Reshape (16,16,1024) |
| C1024K3S2     | Dense 256    | Dropout 0.5   |
| Dense 256     | BatchNorm    | C1024K3       |
|               | Dense 4      | C256K3        |
|               | Softmax      | C64K3         |
|               |              | C64K3         |
|               |              | C1K3          |
|               |              | Sigmoid       |

**Network Architecture** for the encoders, the classifiers, and the decoder. We used ReLU activation for all convolution layers and dense layers. We used the following notation: C64K3S2 is a Conv layer with 64 filters, kernel size \( 3 \times 3 \) and stride 2. Each Conv layer in the decoder is followed by a \( x2 \) upsampling.
or a clinical follow-up. This serves as our ground truth. Fig. 5 shows a set of data samples from the different categories. Perceptual similarities often exist across the class samples.

B. Evaluation

We used a 3-fold cross validation (splitting subjects) to evaluate the classification performance with and without our proposed training process (the baseline). The following settings were used for both the baseline and the proposed training schemes. In each fold 30% of the training set was used as a validation set during the training process. Online data augmentation was performed showing each case 10 times in each epoch with different affine transformations including scale, rotation, and translation. Five images were synthesized for each training sample. The results are presented in Table II. Using our training scheme we achieved an improvement of 7.4% on average with a similar trend in each fold. Table III shows the confusion matrices using the classic and proposed training scheme. Using the proposed training scheme the network was able to improve the overall classification performance but had more mistakes distinguishing between metastasis and hemangioma cases. These cases can be very similar as can be seen in Fig. 5 and additional work should be conducted to tackle this problem.

![Fig. 5. Real data samples used for training including cyst, metastasis, hemangioma, and healthy liver parenchyma.](image)

### TABLE II

| Fold | Baseline [Acc.%] | Proposed [Acc.%] |
|------|------------------|------------------|
| 1    | 64.8             | 67.8             |
| 2    | 69.5             | 75.6             |
| 3    | 68.6             | 81.9             |
| Avg. | 67.6 ± 2.0       | **75.0 ± 5.8**   |

**ACCURACY IN CLASSIFICATION COMPARISON WITH AND WITHOUT THE PROPOSED TRAINING SCHEME.**

### TABLE III

| GT\Predicted | Cyst | Met. | Hem. | Healthy |
|--------------|-----|-----|-----|---------|
| Cyst         | 60  | 5   | 1   | 0       |
| Metastasis   | 13  | 43  | 19  | 6       |
| Hemangioma   | 3   | 19  | 34  | 9       |
| Healthy      | 0   | 0   | 2   | 25      |

**CLASSIFICATION CONFUSION MATRIX. TOP: USING THE CLASSIC TRAINING SCHEME (I.E. THE BASELINE). AND BOTTOM: USING THE PROPOSED TRAINING SCHEME. THE PROPOSED METHOD IMPROVE THE ACCURACY ALONG ALL CLASSES.**

### IV. Conclusions

In this study we showed that using disentanglement of factors we can synthesize new mixtures of samples from different classes. Using the proposed classification target vector, we showed superior accuracy for the liver lesion classification task. Future work includes comparison to other synthetic data generation schemes, as well as extend to more complex data augmentation techniques. In addition, we plan to test the generalization and robustness of our proposed approach on different datasets as well as better understand the different disentanglement factors that contribute to the classification decision using retrieval or attention maps.

### ACKNOWLEDGMENT

We thank Prof. Michal Amitai and Dr. Eyal Klang from the abdominal imaging unit at the Sheba Medical Center for providing the data used in this study.

### REFERENCES

[1] M.F. Mathieu, Junbo J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, “Disentangling factors of variation in deep representation using adversarial training,” in **NIPS**, 2016.

[2] N. Hadad, L. Wolf, and M. Shahar, “Two-step disentanglement for financial data,” in **CVPR**, 2018.

[3] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification,” **arXiv preprint arXiv:1803.01229**, 2018.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in **NIPS**, 2014.