

SURROUNDING CELL SUPPRESSION FOR UNSUPERVISED REPRESENTATION LEARNING IN HEMATOLOGICAL CELL CLASSIFICATION

*Philipp Gräbel**, *Ina Laube**, *Martina Crysandt†*, *Reinhild Herwartz†*, *Melanie Baumann†*
Barbara M. Klinkhammer◊, *Peter Boor◊*, *Tim H. Brümmendorf†*, *Dorit Merhof**

* Institute of Imaging and Computer Vision, RWTH Aachen University, Germany

† Dept. of Hematology and Oncology, University Hospital RWTH Aachen University, Germany

◊ Institute of Pathology, University Hospital RWTH Aachen University, Germany

ABSTRACT

Analysis of hematopoietic cells in bone marrow images is a newly emerging field in computer vision. Deep neural networks provide promising approaches for detection and classification tasks in this field. However, labelling a sufficiently large amount of images by medical experts is infeasible in practice. This can be resolved by semi-supervised methods that use image reconstruction as a way to incorporate images without labelled cells. However, this inevitably leads to an inclusion of surrounding cells into the learned representation. We propose and analyze several techniques for reducing their influence and show that this improves classification results of unsupervisedly learned cell representations.

Index Terms— Representation Learning, Cell Classification, Noise Reduction

1. INTRODUCTION

The analysis of bone marrow samples is particularly important for the diagnosis of hematopoietic diseases such as leukemia. In contrast to peripheral blood, bone marrow shows cells of the entire hematopoiesis including immature forms of most cell types. For diagnosis, trained hematologist need to analyze bone marrow microscopy images by identifying and counting the different cell types and maturity stages. This is tedious work, which could be made more efficient and objective by automating these steps.

As shown in [1], blood cells can be detected with high accuracy while the classification [2] performs worse particularly for cell types with low number of labelled images. Overcoming this issue is particularly challenging due to the large number of classes and an inherent class-imbalance. Other researchers work with a reduced number of classes by summarizing similar cell types [3, 4], which is insufficient for a reliable diagnosis of many hematopoietic diseases.

Due to a high detection accuracy [1], patches with a cell in the center can be easily extracted. These, however, are unlabeled with respect to the classification task as the cell type is

unknown. Because of the significant manual effort and medical expertise required for annotations, the amount of labelled cell images is limited.

Many semi-supervised approaches apply self-supervised learning techniques or use side-objectives such as image reconstruction to incorporate images without labels. Auto-encoders, which comprise encoder-decoder structures for image reconstruction, can be used to learn a meaningful representation based on images alone. Particularly residual networks [5] have been proven to be effective candidates [6, 7, 8] for the encoder network. A representation can either be directly the learned latent space feature vector or a sample of a learned stochastic distribution as with variational auto-encoders [9]. This representation can be used as the input for a classifier, e.g. a shallow fully connected neural network. However, the density of blood cells in bone marrow images results in many cells surrounding the centered cell. For a successful image reconstruction, these need to be part of the representation in the auto-encoder’s latent space even though they are irrelevant for classification of the cell in the center. Ideally, the representation should ignore those surrounding cells.

Existing approaches solve this problem mainly through unsupervised segmentation as pre-processing. Despite promising results [10, 11], these methods highly depend on the complexity of the data.

In this paper we therefore introduce and evaluate three methods to reduce the influence of the area around the cells on the representation. Firstly, we add Gaussian noise to the outer regions of the image. Secondly, we use loss masking to weight the reconstruction loss of pixels in the image center. Thirdly, we use smaller, randomly cropped patches with higher focus on the center. We show that these methods result in a representation that is better suited for classification purposes.

Contributions We propose and evaluate several techniques for suppressing the influence of surrounding cells in unsupervised representation learning for the classification of hematological cells. We find that by suppressing the influence

of surrounding cells, the descriptiveness of learned embeddings increases, resulting in better classification results.

2. IMAGE DATA

In this work, we analyze samples from human bone marrow with Pappenheim staining. Each sample is digitized using a whole slide scanner with $63\times$ magnification and an automated immersion oiling. From each whole slide image, useful regions are extracted and annotated by medical experts.

For the unsupervised representation learning, we use patches centered around individual cells. The positions are automatically determined by a detection network and manually reviewed, which is very efficient and requires no medical expertise. While it can be argued that methods using this dataset should be called "supervised" due to the known positions of cells, we argue that the data is unsupervised with respect to representation learning and classification tasks. Furthermore, this data can be acquired more easily without medical experts and to a large extent automatically. The experiments in this paper are performed with 11 522 samples of cells and cell-like artifacts centered in 256×256 px² patches unless stated otherwise. Cell sizes vary between $8 \mu\text{m}$ and $25 \mu\text{m}$ (92–287 px) in diameter [12], cell-like artifacts can be smaller or larger.

For the evaluation of learned representations, we employ 6 085 unseen patches with known cell type. These include polychromatic, orthochromatic and basophilic erythroblasts, lymphocytes, eosinophilic cells and five neutrophilic cells in different maturity stages (promyelocyte, myelocyte, metamyelocyte as well as band and segmented granulocyte). Examples for two cells are shown in Figure 1.

3. METHODS

As the difference in size varies greatly between cell types, the size of a cell in a given image cannot be easily determined without knowing the cell type. Consequently, we have to either use probabilistic methods, accept a loss in information from the actual cell or risk an influence of surrounding cells. Our approaches follow three different strategies: adding Gaussian noise to the image (Section 3.1), masking the loss (Section 3.2) and using smaller patches (Section 3.3). All methods assume that the cell is roughly centered such that the probability of a pixel showing the given cell decreases with its distance from the center.

3.1. Gaussian Noise

Applying noise in the outer regions of the image makes the network focus on the inner pixels: as there is no consistent gradient in the outer regions, the network is not able to reconstruct these regions. Therefore, they are less likely to become a significant part of the representation.

As the actual size of the cell is unknown, we apply Gaussian noise which increases in variance with increasing distance from the center:

$$I(x, y) = I(x, y) + n(x, y) \quad (1)$$

$$n(x, y) \sim \mathcal{N}\left(0, \sigma^2 \cdot \min\left(1, \frac{d(x, y)}{r_{\max}}\right)^2\right) \quad (2)$$

with $d(x, y) = \sqrt{(x - x_c)^2 + (y - y_c)^2}$ the distance of point (x, y) from the center (x_c, y_c) and $r_{\max} = 128$ as half the patch size. Figure 1 shows examples for different values of σ .

3.2. Loss Masking

To increase the focus on pixels close to the center, loss masking can be applied. Instead of treating each pixel equally when aggregating pixel-wise loss values, they can be weighted based on the distances from the center by element-wise multiplication. For a pixel-wise loss function $L(x, y)$ that is usually aggregated by computing the mean to obtain the final loss, this equates to

$$L_{\text{masked}}(x, y) = L(x, y) \cdot \max\left(0, 1 - \sqrt{\frac{d(x, y)}{r_{\max}}}\right). \quad (3)$$

We propose three different ways of determining the unknown radius r_{\max} .

- **fixed:** $r_{\max} = r_{\text{fix}}$ is set to a pre-defined value.
- **sampled:** $r_{\max} \sim \mathcal{N}(\mu_r, \sigma_r^2)$ is sampled from a Gaussian distribution.
- **learned:** r_{\max} is determined by a small sub-network which is trained simultaneously to find an optimal radius.

The sub-network for learning r_{\max} consists of two fully connected layers, which convert the representation into two scalars. These represent μ_{learned} and σ_{learned} of a Gaussian distribution, a sample of which is used as r_{\max} . As there is a trivial local optimum for $\mu_{\text{learned}} = \sigma_{\text{learned}} = 0$, we further introduce an additional term to the loss function that penalizes small r_{\max} . Figure 1 shows the loss masking gradient overlaid onto the original image.

3.3. Smaller Patches

The usage of a smaller patch size has the advantage of removing surrounding cells, possibly surpassing the disadvantage of removing parts of the cell. To this end, we apply random crop data augmentation. For each image, a random angle $\alpha \in [0, 2\pi)$ is chosen from a uniform distribution and

a radius r_{offset} is sampled from a folded Gaussian distribution $|\mathcal{N}(0, \sigma_{\text{offset}}^2)|$. α and r_{offset} determine how far the cropped patch is shifted from the center. Due to the Gaussian distribution, this random crop includes patches with decreasing probability for increasing distance to the center. In this way, the effect of the cell in the center on the representation increases without completely excluding information from other parts of the image.

3.4. Experimental Setup

In all cases, a variational auto-encoder with a ResNet-18 encoder is used to find a Gaussian distribution of dimension 256. A sample of this Gaussian distribution $z \in \mathbb{R}^{256}$ corresponding to an input image is interpreted as its representation. A shallow decoder network with four layers of transposed convolution with bilinear upsampling and ReLU activation between each layer, followed by a Sigmoid activation function is used to reconstruct an image from the encoded representation. We use the L_1 -distance $L(I_1, I_2) = |I_1 - I_2|_1$ between reconstructed and original image in combination with the variational lower bound as the loss function.

The following parameters are used for the experiments:

Gaussian Noise: $\sigma \in [50, 100, 200]$

Loss Masking (fixed): $r_{\text{fix}} \in [50, 90, 126]$

Loss Masking (sampled): $\mu_r \in [50, 90, 126], \sigma_r = 20$

Loss Masking (learned): penalty for small radii (yes/no)

Smaller Patches (small): size $88^2 \text{ px}^2, \sigma_{\text{offset}} \in [0, 5, 15]$

Smaller Patches (medium): size $172^2 \text{ px}^2, \sigma_{\text{offset}} \in [0, 5, 15]$

Firstly, we evaluate each method with respect to the reconstruction SSIM loss on unseen test-data in five-fold cross-validation. Secondly, we evaluate the best and the worst performing networks from this cross-validation per method in terms of descriptiveness for classification. To this end, we obtain the representations of the labelled cells in the classification dataset from these networks. These are used in five-fold cross-validation to train and evaluate a shallow classifier with three fully connected layers and ReLU activation functions.

For classification, we further compare with two baselines: (1) representations from an identical network without any surrounding cell suppression method (VAE-256), and (2) four times larger representations from a simple residual auto-encoder with a ResNet-18 encoder (RAE-1024). We found in preliminary experiments that residual auto-encoders perform better than variational auto-encoders for larger latent space sizes.

4. RESULTS

Figure 2 shows the mean L_1 -distance between original and reconstruction within the cell contour. The reconstruction quality for smaller patches leads to the lowest L_1 -distance. For Loss Masking, it decreases with decreasing maximum radius. The L_1 -distance for the learned radius and for a small

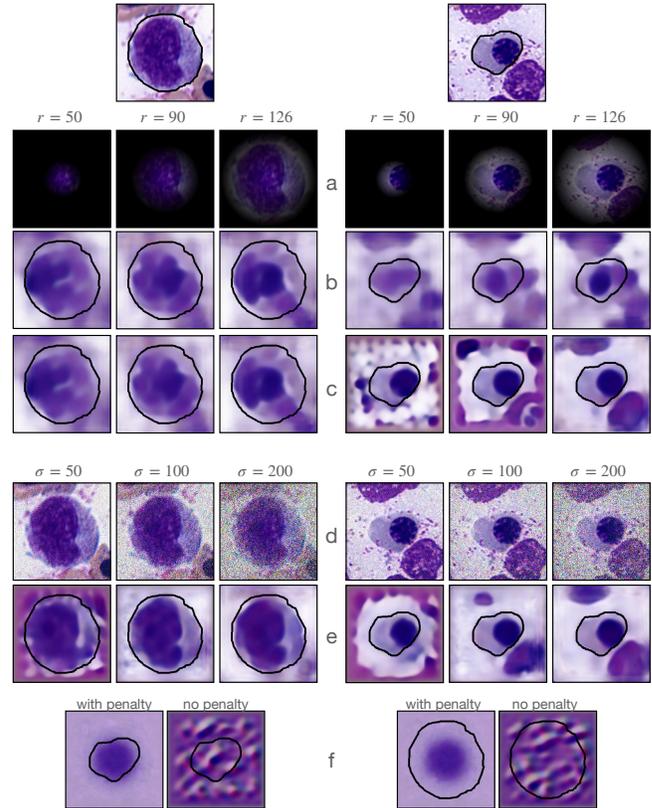


Fig. 1. Visualization for two different input images (left and right block). a) Loss Mask overlaid over image. b) Reconstruction using Loss Masking with fixed radius. c) Reconstruction using Loss Masking with sampled radius. d) Gaussian noise overlaid over image. e) Reconstruction using Gaussian Noise method. f) Reconstruction using Loss Masking with learnt radius.

fixed radius is higher than for other methods. Figure 1 shows reconstruction results for Loss Masking and Gaussian Noise methods.

In Figure 3, the classification results based on the representations of each method are shown. Using smaller patches or Gaussian noise does not improve the results in comparison to the same network without any of these methods (VAE-256). Loss Masking with sampled or fixed maximum radii leads to higher F_1 -scores, in most cases surpassing a network with a four times longer representation (RAE-1024). The highest F_1 -score of 0.63 is obtained by Loss Masking with a radius sampled around $\mu = 50$.

5. DISCUSSION

The results show that Loss Masking yields better representations with respect to the classification task: the F_1 -Score increase of 0.2 in comparison to a network without surrounding cell suppression, is a considerable improvement. While

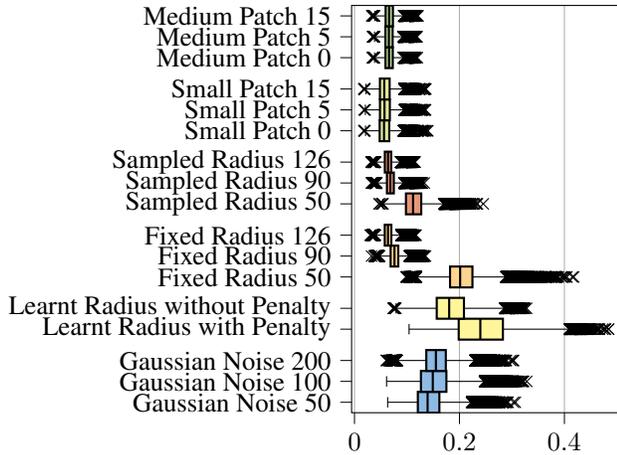


Fig. 2. L_1 -scores of the reconstruction results for each method within the manually annotated cell contour (higher is better).

the choice of the maximum radius r_{\max} does not have a large influence, slightly better results were obtained with $r_{\max} \sim \mathcal{N}(50, 20^2)$. Decreasing the patch size does not lead to a comparable improvement. This could be caused by necessary changes to the network architecture to enable smaller input images.

The intuitively more meaningful approach of letting the network learn the radius itself does not lead to useful results. Further analysis shows that the network without a penalty for small radii often chooses very small radii resulting in non-meaningful representation. If preventing small radii with a penalty, the network optimizes towards local minima of unrealistically large radii, leading to similar problems. More research is needed to find an improved unsupervised learning method for more suitable radius estimation.

We expect that these methods can also be successfully applied to other datasets with centered circular objects. However, it should be considered that the optimal radius r_{\max} highly depends on the input data and has to be determined by hyper-parameter optimization.

A possible downside of these methods is that they work best for objects of approximately similar sizes. For datasets with a larger variance in object size, these methods might not be applicable. Additionally, these methods can only be applied for datasets with centralized objects. The position of the objects needs to be known, i.e. a robust object detection is required.

This paper focuses on unsupervised representation learning. These methods could also result in improvements for semi-supervised classification. We expect that a more relevant inner representation, which focuses on the cells while ignoring the surrounding pixels, is beneficial in these cases as well. Further research is needed to evaluate the effectiveness of these methods for semi-supervised approaches. Addition-

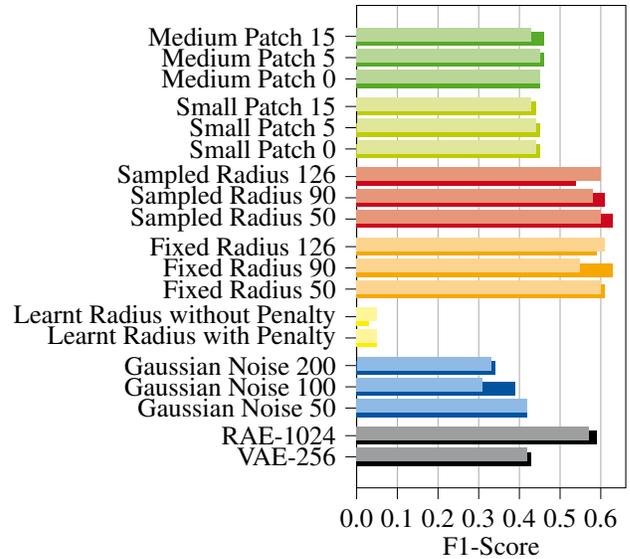


Fig. 3. F1-Scores of classification result using each of the method (higher is better). For each method, the two models with highest (upper, brighter bar) and the lowest (lower bar) reconstruction are used to generate the features used by a shallow classification network. As baselines, a residual auto-encoder (RAE) with a large latent space (size 1024) and a variational auto-encoder with a small latent space (size 256) are shown.

ally, different kinds of gradients for the loss masking (i.e., non-linear pixel weighting) and a larger number of different maximum radii should be considered in future research.

6. CONCLUSION

In this paper, we propose three different methods to reduce the influence of surrounding areas on the image representation. They focus on prior knowledge, particularly the position of the cell within sample images and the range of possible cell sizes. We show that masking the L_1 -loss with linearly decreasing weights from the middle to a chosen maximum radius leads to considerably better representations with respect to a classification task. This considerable improvement cannot be reached with smaller patch sizes or Gaussian noise. A maximum radius sampled stochastically for each batch from a Gaussian distribution with a small mean $\mu_r = 50$ px yields the best results. Although we focused on learning a representation without any class labels, the methods are applicable to semi-supervised methods with similar architectures as well.

7. COMPLIANCE WITH ETHICAL STANDARDS

This work is a purely retrospective, pseudonymized analysis of bone marrow samples under the Helsinki Declaration of 1975/2000 with written informed consent of all patients.

8. ACKNOWLEDGEMENTS

This study was supported by the following grants: DFG: SFB/TRR57, SFB/TRR219, BO3755/6-1, BMBF: STOP-FSGS-01GM1901A, BMWi: EMPAIA project to PB.

9. REFERENCES

- [1] Philipp Gräbel, Özcan Özkan, Martina Crysandt, Reinhild Herwartz, Melanie Hoffmann, Barbara M. Klinkhammer, Peter Boor, Tim H. Brümmendorf, and Dorit Merhof, “Circular anchors for the detection of hematopoietic cells using retinanet,” 2020.
- [2] Philipp Gräbel, Martina Crysandt, Reinhild Herwartz, Melanie Hoffmann, Barbara M. Klinkhammer, Peter Boor, Tim H. Brümmendorf, and Dorit Merhof, “Evaluating out-of-the-box methods for the classification of hematopoietic cells in images of stained bone marrow,” 2018.
- [3] T T. Song, V. Sanchez, H. ElDaly, and N. Rajpoot., “Simultaneous cell detection and classification in bone marrow histology images.,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2018.
- [4] T T. Song, V. Sanchez, H. ElDaly, and N. Rajpoot., “Hybrid deep autoencoder with curvature gaussian for detection of various types of cells in bone marrow trephine biopsy images.,” *IEEE 14th International Symposium on Biomedical Imagings*, p. 1040–1043, 4 2017.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015.
- [6] Lianfa Li, Ying Fang, Jun Wu, and Jinfeng Wang, “Autoencoder based residual deep networks for robust regression prediction and spatiotemporal estimation,” *CoRR*, vol. abs/1812.11262, 2018.
- [7] Simone Zini, Simone Bianco, and Raimondo Schettini, “Deep residual autoencoder for quality independent JPEG restoration,” *CoRR*, vol. abs/1903.06117, 2019.
- [8] Akshay Sethi, Maneet Singh, Richa Singh, and Mayank Vatsa, “Residual codean autoencoder for facial attribute analysis,” *CoRR*, vol. abs/1803.07386, 2018.
- [9] Diederik Kingma and Max Welling, “Auto-encoding variational bayes,” *ICLR*, 12 2013.
- [10] I. Aganj, M.G. Harisinghani, and R. Weissleder, “Unsupervised medical image segmentation based on the local center of mass.,” 05 2018, vol. 8, p. 13012.
- [11] Umaseh Sivanesan, Luis Braga, Ranil Sonnadara, and Kiret Dhindsa, “Unsupervised medical image segmentation with adversarial networks: From edge diagrams to segmentation maps,” 11 2019.
- [12] Gillian Rozenberg, *Microscopic Haematology: A Practical Guide for the Laboratory.*, Elsevier Australia, 2011.