Modifying Variational Auto-encoder Reparameterize Function to Generate Malignant Blood Cell Images

Nathan Bodger
Department of Computer Science
University of Wisconsin - Eau Claire
Eau Claire, United States
bodgernc0166@uwec.edu

Chris Brantner
Department of Computer Science
University of Wisconsin - Eau Claire
Eau Claire, United States
brantncj6345@uwec.edu

Dylan Simanski
Department of Computer Science
University of Wisconsin - Eau Claire
Eau Claire, United States
simansdr4863@uwec.edu

***Abstract*—Variational Auto-encoders (VAE's) have become a popular deep learning model to extract features from images, as well as to transfer the style of one image onto other images. Recent work with VAE's have included the extraction of relevant features from cancer transcriptomes, in addition to predicting what cancer cells may look like in response to chemotherapy. We utilized the Blood Cell Cancer data set from M Amir Eshraghi in an attempt to extract features from malignant preB blood cell cancer images and transfer it onto benign blood cell cancer images to generate images that predict the progression of blood cell cancer in the long-run. The hyper parameter temperature was introduced to the reparameterize function of our VAE to influence the distribution of its latent space and generate more probable images. Various values for the temperature were evaluated to determine which one would perform the best for our model. Our findings could help pave the way for effective cancer progression tools, as well as generating new image data that could be used in future work related to blood cell cancer and deep learning.**

I. Introduction

Variational Auto-Encoders (VAE's) are a probabilistic form of auto-encoders, which work by reducing features of an image down to a latent space and then constructing images based on those latent features. In this, we will investigate the usage of VAE's on blood cell images, both benign and malignant. The variational autoencoder model is shown in 1

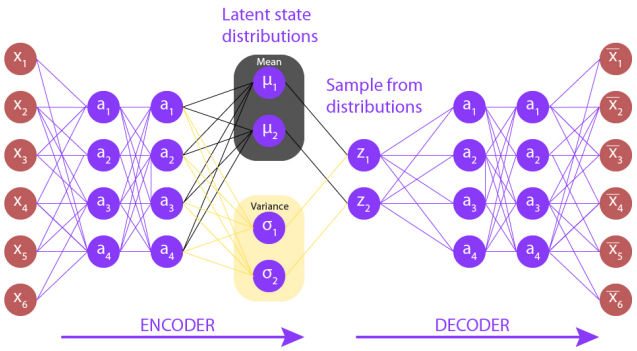


Fig. 1: Model of the autoencoder courtesy of [1]

The probability spaces in the encoder are Gaussian distributions, and estimates are constructed through training of these spaces. An important feature of the system is the *global noise*, or variation across the distribution. In [2], the authors introduce the *"local reparameterization trick"* to model this instead as local noise, allowing for better performance during backpropogation. The vanishing gradient problem is prevelant in this model, and the reparameterization helps minimize this by introducing an artificial noise parameter, denoted ϵ.

II. Previous Work

Cancer prediction in the medical field is of immense significance because of the implications of the prediction to the patient. The ability to accurately predict the progression of

cancer in a patient is directly correlated to the early detection and timely treatment of said patient. Due to the significance of the implications, cancer prediction has seen an increase of interest in recent years. Researchers are applying deep learning models such as VAEs to develop accurate models for predicting cancer. Since the VAE is so adept at selecting prominent features from a given data set, there have been several studies which link the VAE with cancer datasets. One type of study has been used to capture biological features from the TCGA Pan-cancer Project RNA-sequence data set to model gene expression [3]. In this paper the learned features of the authors VAE found tissue specific patterns that were able to distinguish between the 33 cancer types in the dataset. The pattern proved that an unsupervised VAE is able to construct a feature that described a clearly biological source of variance in the data. Our VAE will similarly learn the features of our dataset, however our VAE will not be differentiating between cancer types. We will train our model with just blood cell cancer, and aim to have our biological source of variance be related to the shape of the blood cell. Another type of study was to predict the growth of a lung tumor over time using a Conditional recurrent VAE [4]. In this paper the authors weight time and patient conditions as well as extracted features to create the generated tumor image. By doing this, the VAE will become more personalized to the patient and is able to bypass the issue of having a small dataset. This is important to note for our model because we could incorporate time as a factor for our generated image in future use of our model. Another type of study was to predict the chemotherapy resistance of the cells using VAE [5]. In this paper the authors trained a VAE on a dataset of ovarion cancer cells, and experimented with several different hyperparameters. The results of the paper showed that their VAE is highly robust to cancer data contaminated with large amounts of Gaussian and dropout noise noise through reparameterization, however they did not have enough data to statistically show that the reparameterization was consistently improving the VAE. Our model will be utilizing reparameterization in a similar way, however our model will use temperature and bias to reperameterize. Similarly to the efforts of previous works to model the prominent features of the cancer cells and their response to treatment, our model will attempt to predict how benign cells will appear malignantly. Our model will also address the vanishing gradient problem with reparameterization. In this paper [2], the authors proved that using reparameteriation leads to quicker convergence during backpropogation. In our paper we will test how different reparametirizations will affect the features picked up when our VAE turns a benign blood cancer cell malignant. These works have demonstrated the advantages of using VAEs to extract meaningful features from various cancer datasets. Furthermore, the use of VAEs with diverse data sources has given improved predictive performance of cancerous cells. These works have also demonstrated that it is effective to use reparameterization in order to refine your latent space. This is specifically applicable to our model, where we use reparameterization to extract the shape of malignant cell as a feature. That feature will then be applied to the generated malignant image of the benign input.

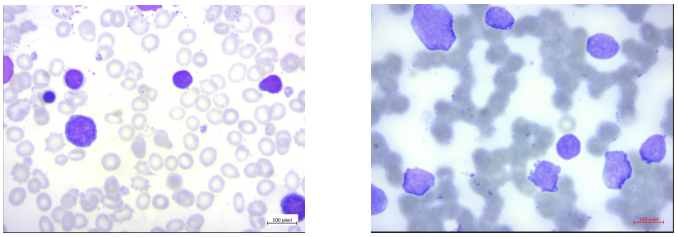
III. Methodology

A. Data Set

The data set [6] comprised of images of peripheral blood smear images of blood cells. The set contains 512 images of benign cells and 979 images of early Pre-B malignant cells. The malignancies are classified as acute lymphoblastic leukemia, and malignancies other than early Pre-B are present in the data set but were not used.In 2, you can see some sample images of the benign and malignant smears in their original format. These images are .jpg filetype with original dimensions of 1024×768.

B. Data Preparation

For preperation of the images for use in the VAE model, we first began by resizing the images to 256×256 using the *BILINEAR* method in python. Images were then inputted into



Benign Malignant early Pre-B

Fig. 2: Peripheral Blood Smear Images

numpy arrays and normalized, and then were split into training and testing sets. The split for this was 70% training and 30% testing. The batch size used for training was 32 images per batch.

C. Architecture

In our work, our constructed VAE had four *Conv2D* layers in the encoder. Each of these convulution layers used a kernel size of 3, strides of 2×2, and *LeakyReLU* activation functions. The filters on each layer were $32, 64, 128,$ and 256 as passing through the encoder. Then, our latent space had dimesion of 128. The decoding layer of our model had four *Conv2DTranspose* layers with the same kernel, stride, and activation function as the encoder. This layer constructed the new 256×256 image from the latent space parameters.

D. Loss Function

The loss function comprised of two parts, the latent loss, or *KL-Divergence Loss* [1] as well as the generative loss. The total loss was these two combined. The *Kullback Leibler (KL)-Divergence* can be calculated as follows: *KL*:=

$$\frac{1}{2}\sum_{j=1}^{J}(1+\log(\sigma_j^{(i)})^2))\frac{1}{2}\sum_{j=1}^{J}(1+\log(\sigma_j^{(i)})^2))-((\mu_j^{(i)}))^2-((\sigma_j^{(i)}))$$

where $X_j \sim N(\mu_j, \sigma_j^2)$ is the standard normal distribution [7] [1]. The *KL-Divergence* plays an important role in calculating the loss. Given a probability distribution z, to generate an observation x from it, we would calculate $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$ from Bayes. To then get $p(x)$ in our case is more difficult, as $p(x) = \int p(x|z)p(z)dz$ which often is intractable either computationally, or has no closed form and is thus also intractable [7]. To fix this problem, we estimate this integral by minimizing the *KL-Divergence* between the two probabilities [7]. Our estimate is given as $\min(KL(q(z|x)||p(z)))$ where q is the output/learned distribution space. Then, if $L(x, \hat{x})$ is the maximum-likelihood estimate for the reconstruction, we have that $Loss = L(x, \hat{x}) + \sum_j KL(q_j(z|x)||p(z))$ [1].

IV. Experiments

A. Temperature

After proposing some select hyper parameters known to effect the distribution of a VAE's latent space during training, evaluation of the trained model's ability to generate images of malignant preB blood cells given an input of benign blood cells. Training was done for 300 epochs during each evaluation, with training data being composed entirely of malignant preB images in order to extract their features. A benign blood cell image was then provided for input, and the output was visually inspected to determine its quality. Among the hyper parameters evaluated, temperature stood out as the one that helped generate the most desirable results. After it was determined that temperature had the greatest effect on our results, different values of temperature were tested to explore temperature's overall effect. The value of 0.001 was the smallest selected value for temperature. Differing temperature increments were tested by increasing the previous temperature by a factor of 10.

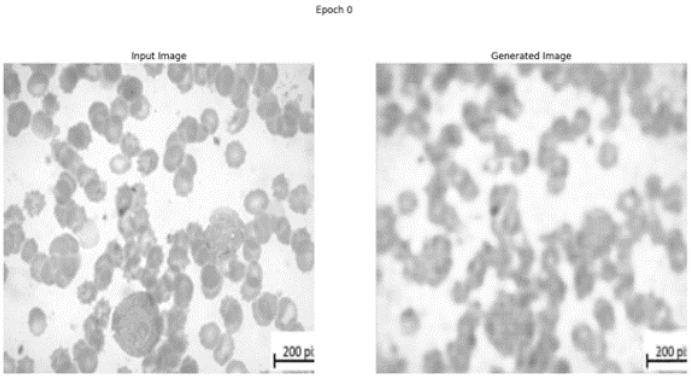


Fig. 3: Input vs. Output for Temperature of 0.001

After observing the input and output of cells based on our initial temperature, we observed that our VAE did not achieve

the desired effect. The general shape of each cell remains relatively smooth and circular, unlike our training data images featuring malignant preB cells. The general appearance of the output indicated that 0.001 was not a sufficient temperature to generate quality results. Although this result doesn't aid in the creation of blood cells that appear to look malignant, this behavior can be explained because of how temperature functions. A temperature closer to 0 signifies that the distribution of latent vectors is closer together, meaning generated images will have a more stable appearance. However, as temperature increases, we expect the output of our images to appear more unstable for our cancer cells.

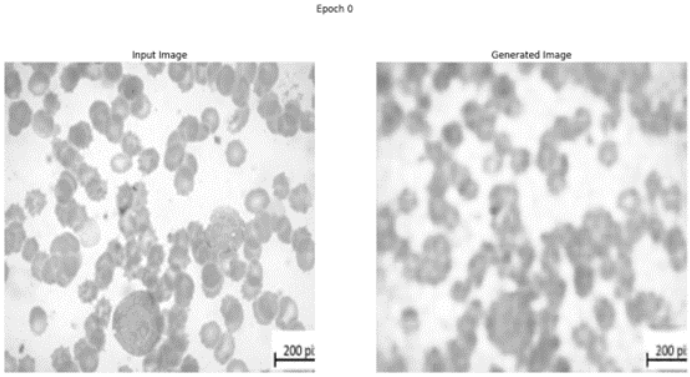


Fig. 4: Input vs. Output for Temperature of 1

By closely examining the appearance of the output in figure 2, slight abnormalities on the edges of cells started being observed at a temperature equal to 1. This temperature displays behavior equivalent to the default behavior of our VAE, because our temperature is a scalar that performs multiplication within our reparameterization function. The result being generated when our temperature is 1 indicates that the VAE model being used is able to extract some of the visual attributes of malignant preB cell images, because the output has some slight cell shape abnormalities. However, the abnormalities being generated were difficult to observe. We predicted that as temperature continued to increase, we would see more significant results.

After evaluating the output of generated images for temperatures greater than 10, such as 100 and 1000, images started

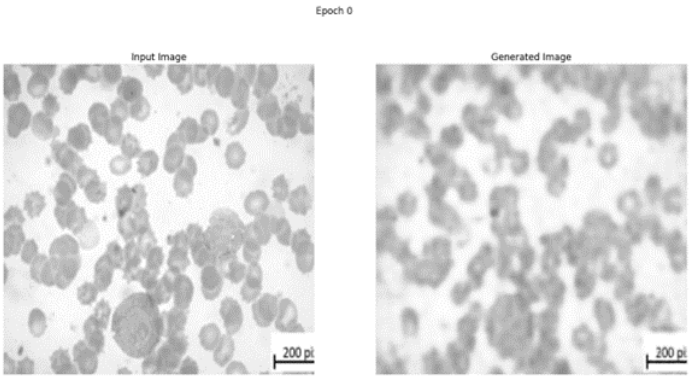


Fig. 5: Input vs. Output for Temperature of 10

displaying results inconsistent with that of malignant preB cells, displaying an appearance similar to that of a benign cell image. One possible explanation for this is because the added randomness at higher temperatures is allowing the model to explore a wider range of latent vectors that are closer to the latent vectors of the original cells. At higher temperatures, the VAE is more likely to generate samples that are further away from the mean of the learned distribution in the latent space. This increased variability can allow the model to explore latent vectors that are closer to the latent vectors of the original cells that it was trained on. As a result, the generated images may more closely resemble the original cells.

B. Introducing Bias

In addition to the hyper parameter temperature, bias was another changed considered for modifying the parameterization function. Although bias did not have a significant impact on our results by itself, it was possible that combining it with temperature could yield more notable outputs by influencing the sampling process to bias the generated samples towards characteristics that are consistent with malignant preB cells. In order to evaluate this, temperature was kept constant at a value of 10, and various levels of bias were introduced, starting at 0.1 and tested at various increments by a factor of 10.

Using a bias of 0.1 had little impact on the output, but by observing the differences between figure 3 and figure 4, it can be noted that when a bias of 1 was used, the edges of neighboring cells in the output was generally more sharp

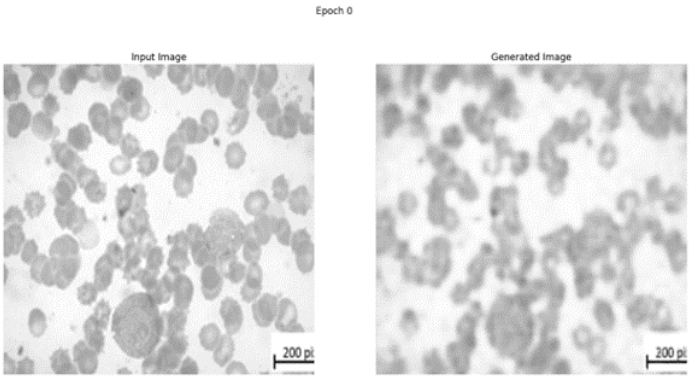


Fig. 6: Input vs. Output for Temperature of 10 + Bias of 1

compared to the output without bias, while maintaining the desired abnormal cell shape. In this case, the bias term of 1 caused the latent vectors to become more biased towards values that result in sharper borders for the generated images. After a bias of 10 was used, results similar to when the bias was equal to 1 were generated. After bias values of 100 and 1000 were used for image generation, the overall quality of the generated images began to decline significantly, which can be seen in Figure 5.

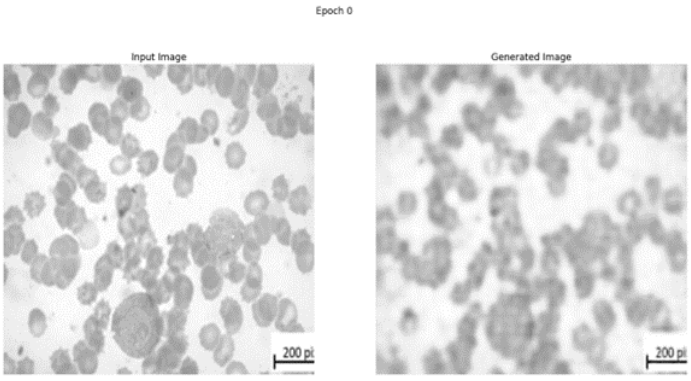


Fig. 7: Input vs. Output for Temperature of 10 + Bias of 1000

From our results, it can be reasonably determined that biases of 1 and 10 led to better image generation because it created a good balance between the diversity of the latent vector distribution and fidelity of our input. However, a bias of 1000 is too large, which led to samples that were too diverse and therefore unfaithful to the true distribution.

C. Evaluating Results

Based on these findings, there is something to be said about the latent distribution of our VAE with respect to both temperature and bias. For this data set, it was determined that a relatively high degree of randomness helped our model utilize a latent space distribution that reflected the characteristics of a malignant preB cell. In nature, the appearance of these cells is often random, which needs to be considered if we are trying to recreate this effect with image generation. Not only will a higher temperature help reflect the biological mechanics of cells, but a high temperature will also help with overfitting by ensuring our VAE will not repeatedly create the same image. Additionally, there are many different characteristics of malignant preB images that could be selected from our latent space, some of which may be unnecessary for our task. Because of this, the optimal bias has been extremely helpful for accessing the features that matter most. Although the impact that temperature and bias has had on image generation has only been explored with the data set we used, it is possible that the values identified as helpful for this data set could be used on similar data sets. Cells are not man made, so they do not possess the predictability of human designs. As it can be seen from the inputs and outputs of these experiments, cells have highly varied sizes, perimeters, and spacial distributions. Because of this, cell image generation will most likely benefit from very particular hyper parameters. The hyper parameters explored in these experiments are a select few from the many possible choices that could have been made, so it is important not to disregard them. It will take time to thoroughly consider which mathematical functions would be relevant for the reparameterization function of a VAE, however, experiments evaluating the effectiveness of VAE image generation should put more focus into what parts of the reparameterization function can be modified, because of the crucial role it plays in creating the output.

V. Conclusion

Our model is promising, however the reparameterization needs to be more fine tuned in order to truly optimize the latent space distribution for feature extraction of the malignant blood cells. We observed that a mixture of both temperature and bias in the parameterization function resulted in the output that was most consistent with the abnormalities of malignant preB blood cells.

A. Future Work

Since the model trained off of the features of malignant cells, it would be beneficial to experiment with multiple different dropout rates in the future to see if we can reduce over fitting of the model. In the future it would be important to train the VAE on several stages of blood cancer malignancies, such as proB and early preB. This would be beneficial because it is more applicable to a real life situation. Someone would be able to compare the generated malignant blood cell to the input blood cell through multiple stages, instead of just the preB we trained the model on.

References

[1] "Geeks For Geeks role of kl-divergence in variational autoencoders," https://www.geeksforgeeks.org/role-of-kl-divergence-in-variational-autoencoders/, accessed: 2023-04-29.

[2] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," *Advances in neural information processing systems*, vol. 28, 2015.

[3] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*. World Scientific, 2018, pp. 80–91.

[4] N. Xiao, Y. Qiang, Z. Zhao, J. Zhao, and J. Lian, "Tumour growth prediction of follow-up lung cancer via conditional recurrent variational autoencoder," *IET Image Processing*, vol. 14, no. 15, pp. 3975–3981, 2020.

[5] T. M. George and P. Lio, "Unsupervised machine learning for data encoding applied to ovarian cancer transcriptomes," *bioRxiv*, p. 855593, 2019.

[6] "Kaggle blood cell cancer (all) dataset."

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022.