HIGH SCHOOL EDITION
Journal of Student Research

Dataset-Agnostic Vessel Segmentation of Retinal Fundus Images by a Vector Quantized Variational Autoencoder

Tejas Prabhune[1] and David Walz[#]

[1]Evergreen Valley High School, San Jose, CA, USA
[#]Advisor

ABSTRACT

The use of retinal fundus images plays a major role in the diagnosis of various diseases such as diabetic retinopathy. Doctors frequently perform vessel segmentation as a key step for retinal image analysis. This is laborious and time-consuming; AI researchers are developing the U-Net model to automate this process. However, the U-Net model struggles to generalize its predictions across datasets due to variability in fundus images. To overcome these limitations, I propose a cross-domain Vector Quantized Variational Autoencoder (VQ-VAE) that is dataset-agnostic - regardless of the training dataset, the VQ-VAE can accurately classify vessel segmentations. The model does not have to be retrained for each different target dataset, eliminating the need for new data, resources, and time. The VQ-VAE consists of an encoder-decoder network with a custom discrete embedding space. The encoder's result is quantized through this embedding space then decoded to produce a segmentation mask. Both this VQ-VAE and a U-Net model were trained on the DRIVE dataset and tested on the DRIVE, IOSTAR, and CHASE_DB1 datasets. Both models were successful on the dataset they were trained on - DRIVE. However, the U-Net failed to generate vessel segmentation masks when tested with other datasets while the VQ-VAE performed with high accuracy. Quantitatively, the VQ-VAE performed well, having F1 scores from 0.758 to 0.767 across datasets. My model can produce convincing segmentation masks for new retinal image datasets without additional data, time, and resources. Applications include using the VQ-VAE after fundus image is taken to streamline the vessel segmentation process.

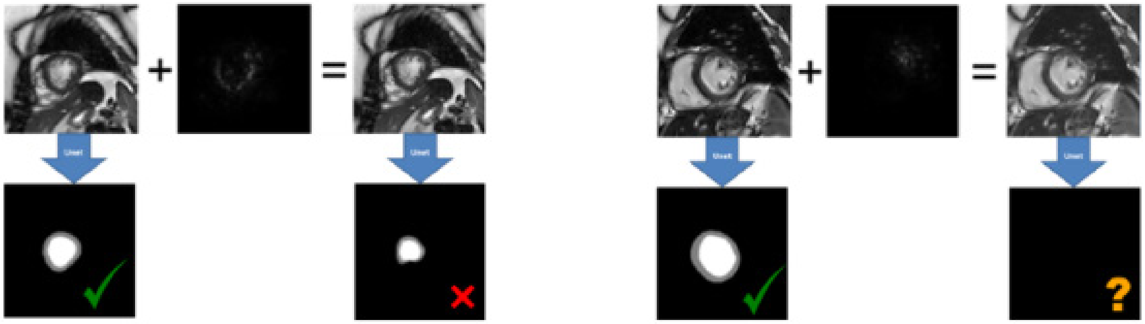
Introduction

The use of retinal fundus images plays a major role in the diagnosis of various diseases such as diabetic retinopathy (Abramoff et al. 2010). Doctors frequently perform vessel segmentation, a method of highlighting the blood vessels in the retina, as it is a key step for retinal image analysis due to its ability to provide valuable information of the blood vessels, including the width, tortuosity, and branching patterns. This segmentation process is very laborious and time-consuming, leading to AI researchers developing models to automate this process. The state-of-the-art method for vessel segmentation is based on the U-Net architecture and has been very successful in replicating the target segmentation masks (Guo et al. 2020).

However, fundus image textures can greatly vary between cameras and datasets. Even for the same camera/dataset, textures may appear different due to changes in clinics and subjects (Zhao et al. 2019). These variabilities make it so a U-Net model trained on one dataset cannot be used on another. It must be retrained with the new data to properly predict segmentation masks. This requires heavy computational resources, a large dataset with textures from the target image, and the necessary time for retraining.

To overcome these limitations, I propose a cross-domain Vector Quantized Variational Autoencoder (VQ-VAE) that is dataset-agnostic - regardless of the training dataset used, the VQ-VAE can accurately classify vessel segmentations. The model does not have to be retrained for each different target dataset, eliminating the need for new images, resources, and time.

Yan et al. observe a similar generalizability problem with the U-Net for cardiac cine MRI images (Figure 1) (Yan et al. 2019). Even small changes in image pixel values lead to the U-Net generating incorrect segmentations or completely blank images, revealing the U-Net's inability to generalize well. To solve this problem, they use a CycleGAN model, which can convert images between two texture "styles", to translate the texture of the target MRI image to match the textures from the image domain the U-Net was trained on. However, this requires the training of a new CycleGAN for each target domain, which needs additional time and computational power.



(a) example 1

(b) example 2

Previously, an R-sGAN has also been used tackle this issue, which generates new images of the target retinal fundus image domain and retrains the segmentation U-Net with this synthesized dataset (Zhao et al. 2019). While this does handle the problem of needing a large dataset, this does not address the issue of the additional time and computational resources required for retraining. The model I propose solves the problem of creating segmentation masks of a target image domain without the need of a separate GAN for image synthesis, time for retraining, or other computational resources.

Methods

U-Net

The U-Net architecture is illustrated in Figure 2. Following the architecture from Ronneberger 2015, it consists of a contracting path and expansive path, with skip connections transferring information between the two. The contracting path is a typical convolutional network, with four repeated units of 3x3 convolution (to gather feature information) and 2x2 max-pooling (for downsampling) layers to encode the retinal fundus image down to a one-dimensional vector of size 1024. The expansive path, similarly, decodes the image using four repeated units of 2x2 up-convolution layers (reduces feature channels) and 3x3 convolution layers, which brings the vector to the original image size. However, the image produced using the feature information is a corresponding vessel segmentation of the original retinal fundus image.

This model was used as a control, to both show the U-Net's inability to generalize across datasets and compare the results of the two models.

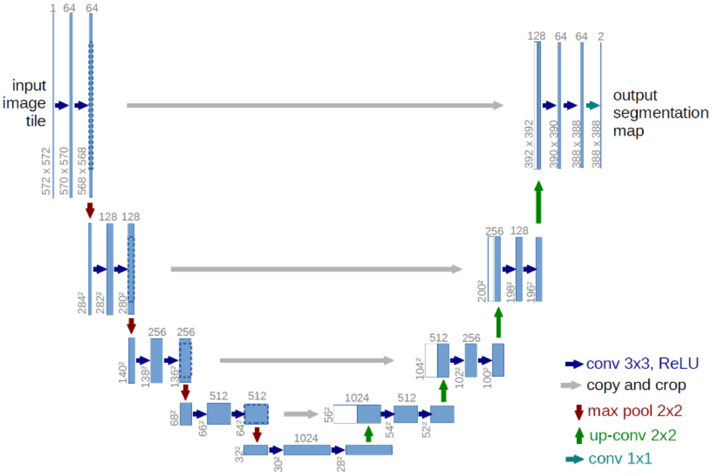


Figure 2. Structure of the conventional U-Net (taken from Ronneberger 2015)

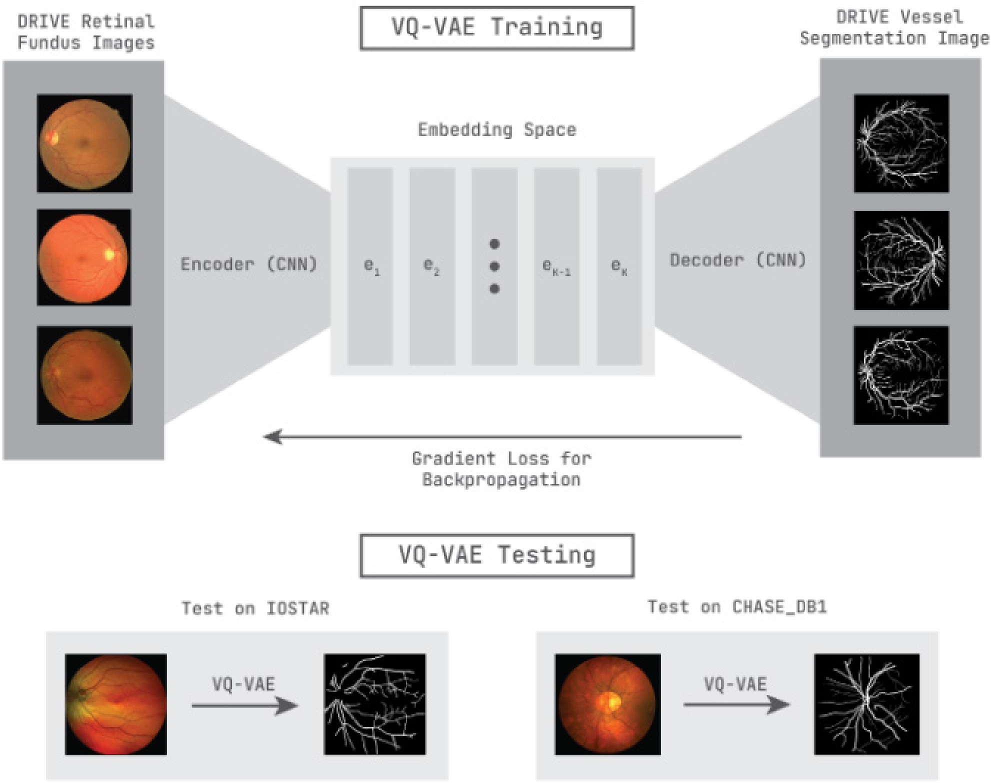
VQ-VAE

Traditionally, the VQ-VAE is set up as a generation model rather than a translation model between two images. For the purposes of vessel segmentation, the base architecture from Oord et al. 2018 is used, but modified to accommodate translation between retinal fundus images and their vessel segmentations.

The base VQ-VAE (Figure 2) uses three repeating units of 3x3 convolution and 2x2 max-pooling layers to encode the retinal fundus image to a vector representation z_e of size (124, 124, 32). A latent embedding space with K embedding vectors e_i is defined. This is a low-dimensional space where vectors representing similar images are placed close together and allows models to learn the general features of the data. The vectors with the shortest $z_e - e_i$ distances (how close image representations are to the defined embedding vectors) became part of the discrete representation variable $z_q(x)$, or the quantized latent variable. This process of quantization is necessary for the model to learn macro-level features of the image, which is how it can generalize across datasets. This new vector representation is then decoded with three repeating units of 3x3 up-convolution layers and two upsampling layers to convert the feature information to a segmentation mask.

Three losses are used for the VQ-VAE to incentivize certain learnings: a reconstruction loss to maximize the similarity between the true and predicted segmentations, an l_2 loss to minimize the distance between z_q and e_i (the initial and quantized vector representations), and a commitment loss to make sure the defined latent embedding space does not increase in volume without bound (this would cause overtraining, since the embedding space would accommodate individually to each image).

$$L = \log p\left(x \middle| z_q(x)\right) + \left|\left|sg[z_e(x)] - e\right|\right|_2^2 + \beta\left|\left|\mathrm{z}_\mathrm{e}(x) - sg[e]\right|\right|_2^2$$



Both models were trained on the public DRIVE dataset made of 40 pairs of retinal fundus images/segmentation masks captured on a Canon CR5 non-mydriatic 3CCD camera (Staal et al. 2004). After training, each model was tested on the DRIVE dataset, the public IOSTAR dataset (EasyScan SLO (scanning laser ophthalmoscopy); 30 pairs), and the public CHASE_DB1 dataset (Nidek NM-200-D camera; 28 pairs) (Jiong 2016, Abbasi-Sureshjani 2016, Fraz et al. 2012). By testing on three different datasets with distinct fundus image textures, the generalization capability can be highlighted.

Results and Discussion

Qualitative Analysis

As seen in Figure 3, the U-Net performed best on the test images from the dataset/camera it was trained on - DRIVE. The predicted vessels have the same structure as the true segmentation masks including finer details. However, when the model attempted to predict retinal images from the IOSTAR or the CHASE_DB1 datasets, it produces a black screen with no clear depiction of retinal vessels.

The VQ-VAE has similar results to the U-Net for the DRIVE dataset. Both the main vessels and surrounding thin branches are picked up by the VQ-VAE. In contrast to the U-Net predictions, the VQ-VAE performed predictions of other datasets successfully without any additional training. Even though the textures of the IOSTAR and CHASE_DB1 cameras are different from the DRIVE textures, the predicted segmentations by my model have as much detail in the main vessel trunks and the branches as in the DRIVE predictions. The vasculatures are all present for predicted segmentations in every dataset.

For the IOSTAR images, the VQ-VAE has a few false positive and negative issues. In the predicted IOSTAR images shown in Figure 3, the VQ-VAE detects false positives near the edges of the image, and some vessel structures near the center of the image are not detected (false negatives).

There are similar issues with the predicted CHASE_DB1 images, where the optic disc at the center of the images is mistaken for vessels by the VQ-VAE. In addition, the vessel structures near the top right are not fully detected for the second CHASE_DB1 image.

While the U-Net was only able to produce segmentation masks for the training dataset, DRIVE, while failing to generalize to the IOSTAR and CHASE datasets, the VQ-VAE was able to predict accurate segmentations for every dataset, regardless of the changing textures.

Quantitative Analysis

The F1 score is a tool used to measure the differences between the true and generated images and provides a quantitative measure of how accurate a model's segmentation method is (Staal et al. 2004). This score was used to compare the images predicted by VQ-VAE to the true segmentation masks:

$$F_1 = 2 * \frac{precision*recall}{precision+recall} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

where TP is true positive, FP is false positive, and FN is false negative. As seen in Table 1, the F1 scores for the VQ-VAE on DRIVE are slightly behind state-of-the-art models such as the SA-UNet, which have F1 scores over 0.8 (Guo et al. 2020). However, the VQ-VAE has high F1 scores across multiple datasets, while the U-Net has scores of 0 for the IOSTAR and CHASE images.

Table 1. F1 Scores for the VQ-VAE.

	DRIVE	IOSTAR	CHASE_DB1
F1 Score	0.767	0.761	0.758

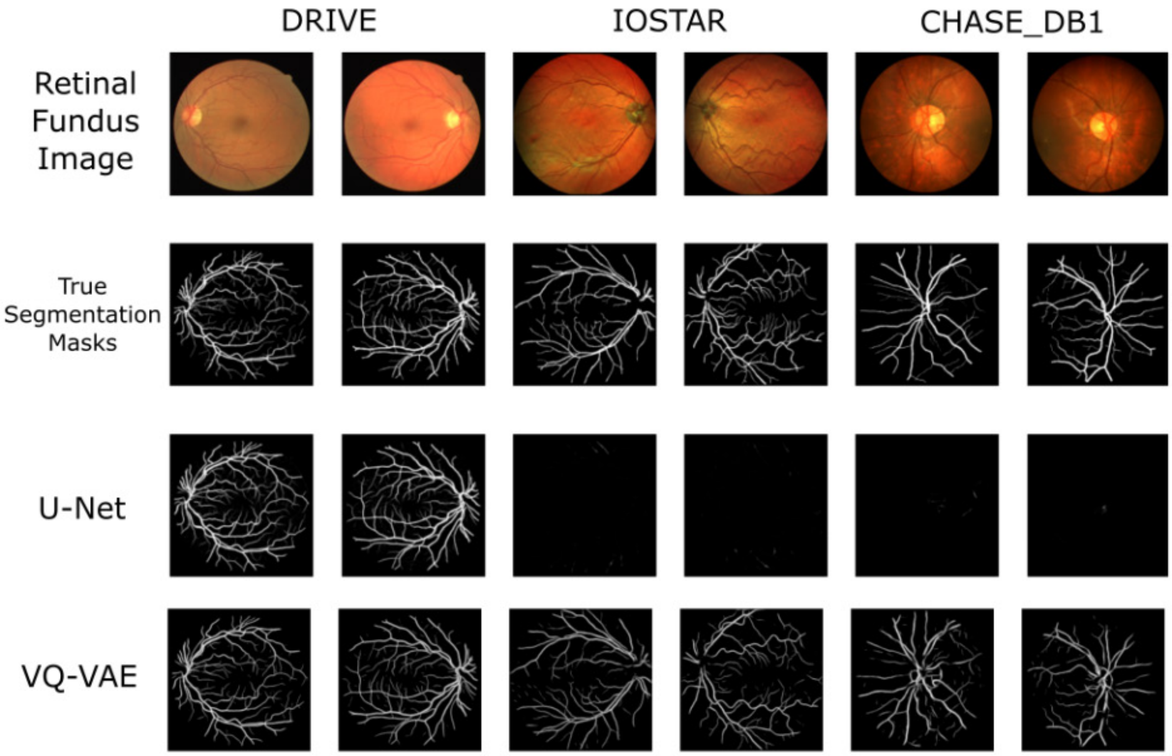


Figure 3. A table comparing true segmentation masks, U-Net predicted masks, and VQ-VAE predicted masks on 3 datasets.

Conclusion

The VQ-VAE consistently outperformed the U-Net on data from other cameras. It was able to generate convincing segmentation masks and maintain high F1 scores across multiple datasets.

While current solutions can employ a CycleGAN or R-sGAN to adapt the U-Net to a target domain, they require additional data, time for retraining, and computational power to train. These extra requirements make these solutions inaccessible to doctors, especially those with low budgets and/or in developing countries.

The cross domain VQ-VAE introduced in my research can produce convincing segmentation masks for target retinal image datasets without the added requirements of current solutions. This can increase accessibility for doctors and streamline the segmentation and diagnosis process through automation.

Future Work

Future work includes adding a self-attention layer similar to the SAGAN and the SA-UNet to lower the amount of false positive and negative instances (Zhang et al. 2019, Guo et al. 2020). The VQ-VAE can also be trained on IOSTAR or CHASE_DB1 datasets and tested on the other two datasets to verify whether the VQ-VAE can be trained on any initial dataset. This model structure can be tested on other medical images as well to see if the generalizability translates well to other image types.

Acknowledgments

Thank you to my family for supporting and sponsoring my project with funding for Google Colab.

References

Abràmoff, Michael D et al. “Retinal imaging and image analysis.” *IEEE reviews in biomedical engineering* vol. 3 (2010): 169-208. doi:10.1109/RBME.2010.2084567

Fraz, Muhammad Moazam et al. “An ensemble classification-based approach applied to retinal blood vessel segmentation.” *IEEE transactions on bio-medical engineering* vol. 59,9 (2012): 2538-48. doi:10.1109/TBME.2012.2205687

Guo, Changlu, et al. “SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation.” *ArXiv:2004.03696 [Cs, Eess]*, (2020). arXiv.org, http://arxiv.org/abs/2004.03696.

Zhang, Jiong et al. “Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 12, pp. 26312644, (2016). DOI: 10.1109/TMI.2016.2587062

Oord, Aaron van den, et al. “Neural Discrete Representation Learning.” *ArXiv:1711.00937 [Cs]*, (2018). arXiv.org, http://arxiv.org/abs/1711.00937.

Ronneberger, Olaf, et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” *ArXiv:1505.04597 [Cs],* (2015). arXiv.org, http://arxiv.org/abs/1505.04597.

Abbasi-Sureshjani, Samaneh et al. “Biologically-inspired supervised vasculature segmentation in SLO retinal fundus images,” in *International Conference Image Analysis and Recognition*, pp. 325334. Springer, (2015). DOI: 10.1007/978-3-319-20801-5_35

Staal, Joes et al. “Ridge-based vessel segmentation in color images of the retina.” *IEEE transactions on medical imaging* vol. 23,4 (2004): 501-9. doi:10.1109/TMI.2004.825627

Yan, Wenjun, et al. “The Domain Shift Problem of Medical Image Segmentation and Vendor-Adaptation by Unet-GAN.” *Medical Image Computing and Computer Assisted Intervention MICCAI 2019*, edited by Dinggang Shen et al., vol. 11765, Springer International Publishing, (2019), pp. 62331. DOI.org (Crossref), doi:10.1007/978-3-030-32245-8_69.

Zhang, Han, et al. “Self-Attention Generative Adversarial Networks.” *ArXiv:1805.08318 [Cs, Stat]*, June 2019. arXiv.org, http://arxiv.org/abs/1805.08318.

Zhao et al. "Supervised Segmentation of Un-Annotated Retinal Fundus Images by Synthesis." *IEEE Transactions on Medical Imaging* vol. 38, no. 1, pp. 46-56, (2019) doi: 10.1109/TMI.2018.2854886.