Vesuvius: Ink Detection on 3D X-Ray Scans

Alex Golab
UBC

Daniel Yang
NTU / UBC

Evan Liu
UBC

Abstract

The interpretation of text in fragile historical records remains a challenge today. One of these records is the Herculaneum papyri, a set of fragile scrolls from over 1900 years ago. In this paper, we outline attempts to reconstruct text from carbon ink papyrus by applying deep learning methods. We demonstrate the efficacy of convolutional neural networks (CNNs) and transformers in detecting and reconstructing text from 3D voxel inputs. We propose a variation of transformer models, the UNet Vision Transformer, and compare its advantages and limitations with other baseline models.

1 Introduction

In 79 CE, Mount Vesuvius erupted and buried the cities of Pompeii and Herculaneum in ash and pumice. Among the many buried buildings was the house of Julius Caesar's father-in-law, which contained a massive library of literary works. Over 1,800 scrolls have been identified, making them one of the largest surviving collections of ancient literature.

Attempts to unroll the fragile papyrus scrolls in the 1750s mostly resulted in damage, but recent high-resolution 3D X-ray CT scans have been used to create reconstructions of the scrolls, but due to the ink being radiolucent we cannot discern the content by the human eye alone. Recently, a paper by Parker et al. (2019) demonstrated that by training a 3DCNN on the texture differences and microscopic bumps on the scrolls from the 3D X-ray CT scans, ink location can be inferred. However, this model was trained on a limited dataset for small sections of the scrolls. As a subset of the overall challenge of reconstructing and reading these scrolls, our objective is to develop a model that can extract information about the ink location of the unravelled fragments.

To tackle the binary segmentation problem with voxels as input modality, we proposed a UNet-ViT model and compared it to UNet, 3D Vision Transformer and a 3D-CNN baseline we re-implemented. We argue that the UNet structure should be able to capture detailed local features, while ViT captures global information. However, non of our models were able to generalize on unseen data, and training often becomes unstable. With some experimental results on the validation set, we discovered vanilla UNet is likely the most effective among all models.

2 Background and Data

The Vesuvius dataset provides us with scroll fragments from previous unraveling attempts. These fragments have been scanned using high resolution 3D X-ray tomography, providing us with information about their texture and thickness. The resulting 3D scans are stored as 65 2D .tiff files, with each file representing a slice in the z-direction.

Although the ink is not visible in the X-ray scans, the ink is opaque under infrared light. An infrared photo of the surface and a carefully labeled binary mask is provided as 2D label ground truth and masking of valid data.

Research project for UBC CPSC 440/540, April 2023.

Due to the scarcity of the scrolls, we are only provided with three scroll fragments as the training set. Since each scan has extremely high resolution ($4um$), we downsampled all files by a factor of 2 before training. To expand the miserably small training set, we added random scaling and random flips when training the models.

Figure 1: Image of a Vesuvius dataset fragment. On the left, an infrared image of the fragment. On the right, a hand-labelled mask of the original fragment.

3 Related Work

X-ray image reconstruction is a broad field attempted by Mocella et al. (2015) and Bukreeva et al. (2016) previously. Classical works focus on X-ray computed tomography (XCT) and X-ray phase-contrast tomography (XPCT) techniques. These methods involve projecting the penetrating radiation of an X-ray source onto the papyri and reconstructing the image based on X-ray absorption and refraction. The carbon ink in the papyri can be differentiate via the absorption and refraction properties, creating outlines of how the ink looks.

More recently there has been work done by Parker et al. (2019) to detect and reconstruct the ink text from the papyri into a human-readable format. The authors of the paper apply a 3D convolutional neural network (3DCNN) to reconstruct the text from the tomography data, demonstrating some success with coarse resolution reconstructions. The Vesuvius dataset is a culmination of these works, and provide the dataset for our experiments. In our experiments, we build on the authors' works to reconstruct ink text from the data.

Vision transformers (ViT) are known to be strong alternatives to convolutional neural network (CNN) architectures. Dosovitskiy et al. (2020) have previously demonstrated that ViT can obtain state of the art performance compared to CNNs, while also requiring far less computing resources. Several variations of vision transformers have also found success in their own areas. Examples include Wu et al. (2021) convolutional vision transformers (CvT) and He et al. (2021) masked autoencoders (MAE). These models build on the ViT model by applying certain changes to bits of the architecture. However, none of them have previously been applied to the task of reconstructing ink samples from 3D X-ray tomography data.

U-Nets (UNet) are known for their strong performance in image segmentation tasks. Ronneberger, Fischer, and Brox (2015) demonstrated the architecture consisting of a contracting path and an expanding path allows for strong precision in identifying image boundaries. In addition, UNets are known for quick training, boasting less than 1 second of runtime on a GPU. The UNet Transformer model by Petit et al. 2021 combines UNet and transformer models to track relationships in highly complex image data. We extend the method to reconstructing ink text in 3D X-ray tomography data.

4 Methods

In addition to the baseline 3DCNN model, we explored UNet and Vision Transformer (ViT) model and the combination of both models. On each training step, we sample an chunk of fixed size voxel, centered on a 2D pixel where the mask is valid. On the Z dimension, we only take slices 16 48, since not all 65 slices contain useful information and loading all slices takes substantial amount of memory.

4.1 3DCNN Re-implementation

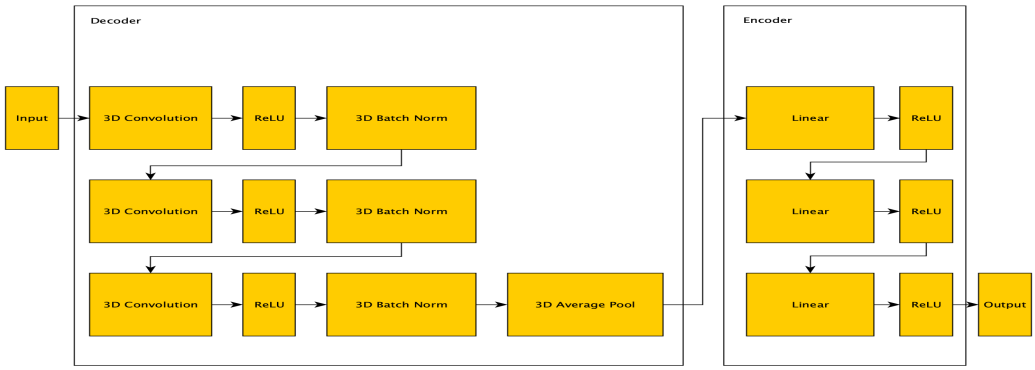

Figure 2: The 3DCNN architecture. The encoder is composed of 3D convolution layers, ReLU activations, and 3D batch norms. The decoder is composed of linear layers and ReLU activations to recreate the image.

To start we utilized an auto-encoder/decoder model based on the 3DCNN used by Parker et al. (2019). The encoder part of the model consists of three 3D convolution layers, which are used to extract high-level features from the input fragments. The extracted features are then passed through a rectified linear unit (ReLU) activation layer to introduce non-linearity and improve model performance. The final layer in the encoder is a 3D batch normalization layer, which helps to normalize the feature maps and improve the convergence of the training process.

The decoder part of the model is designed to reconstruct the input image from the encoded features. It consists of a series of linear layers and ReLU activation layers. The linear layers are fully connected layers that take the encoded features as input and reconstruct the original image. The ReLU activation layers are added to introduce non-linearity and prevent overfitting of the data.

During training, the model takes 32 random patches of the fragments as input. The fragments are split into patches to increase data and the squares are randomly chosen to prevent overfitting. The model is trained using stochastic gradient descent (SGD), which updates the model parameters based on the gradient of the loss function with respect to the parameters.

4.2 3D Vision Transformer (3D ViT)

A naive attempt to encode voxelized data is to employ a 3D vision transformer (3D ViT). A 3D vision transformer is similar to ViT, but it slices patches in three dimensions - height, width, and depth. A $class$ embedding is only added to the Z dimension since binary classification is done only on the Z dimension.

4.3 UNet

We construct a UNet based on the work of Ronneberger, Fischer, and Brox (2015). The downsampling layers are each composed of a max pooling layer, two convolutional layers, two batch norm layers, and two ReLU activations. These outputs are fed into upsampling layers each composed of a transposed convolution layer, two convolutional layers, two batch norm layers, and two ReLU activations. The residual connections between downsampled and upsampled vectors facilitate loss propagation.

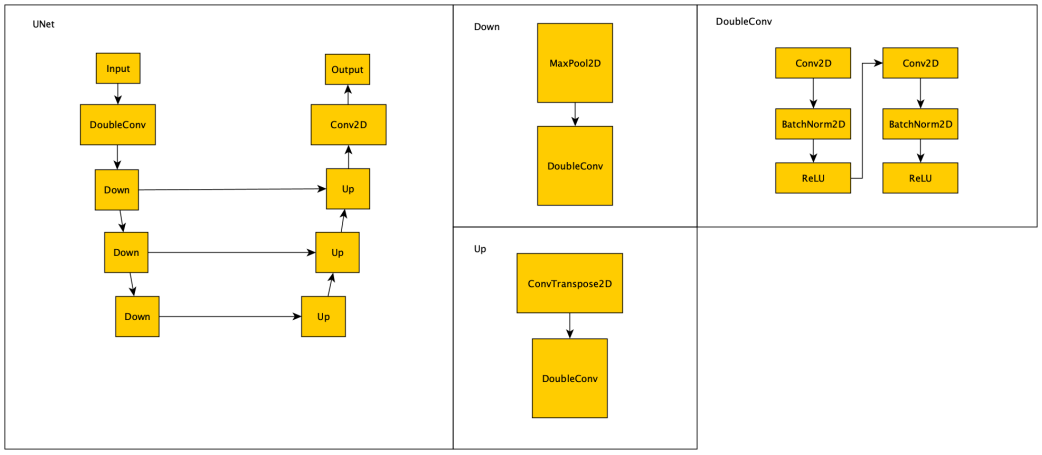

Figure 3: On the left, the architecture for the UNet model. On the center and right, the components used in the model.

4.4 UNet Vision Transformer

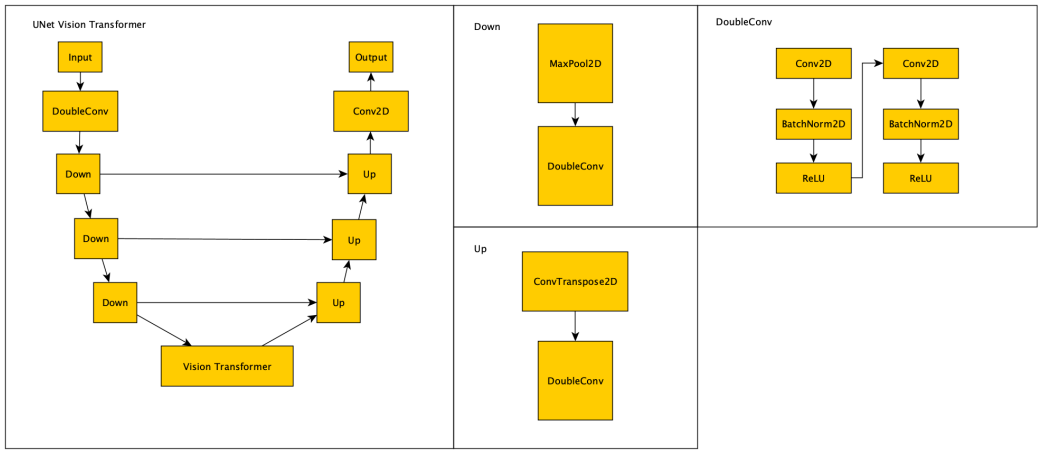

Figure 4: On the left, the architecture for the UNet vision transformer. On the center and right, the components of the model.

UNet has been successful in image segmentation problems, while Vision Transformer is excellent in capturing global dependencies. We attempt to combine the two to take the best out of both worlds. The UNet-ViT has three stages, (i) encode each patch with the convolution layers and max pooling, (ii) employ an Vision transformer network to encode the interactions between patches, (iii) decode each patch an obtain a binary mask with deconvolution layers. The model slices the voxels into patches in the horizontal and vertical dimensions, leaving depth as channel feature. Then convolution is independently applied on each patch. The resulting latent vector of each patch is passed to ViT as input features.

5 Results

Since we have limited access to the public testing set (Kaggle limits each group 5 attempts per day), we trained the model on two of the three fragments, and evaluated the model on the third fragment.

We measure the similarity of our prediction and ground truth by applying the Sorensen-Dice coefficient at a real factor 0.5 to measure precision and recall. This is equivalent to applying a F0.5 score.

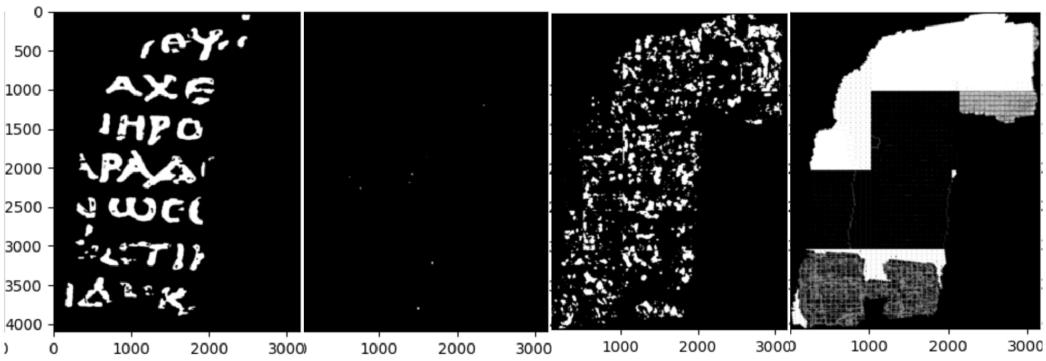

Figure 5: Validation set predictions from different models. From left to right are: ground truth, 3D ViT, UNet, UNet-ViT.

$$\frac{(1+\beta^2)pr}{\beta^2 p + r} \text{ where } p = \frac{tp}{tp+fp}, \; r = \frac{tp}{tp+fn}, \; \beta = 0.5$$

In addition, we apply recall and precision to measure the performance of the model against the unknown data. Some models display non-deterministic results. On some runs, the Sorensen Dice

	training				DSC ↑		recall ↑		precision ↑	
Methods	steps	rand. flip	patch size	sample size	train	val	train	val	train	val
3DCNN	6000	×	64	512	nan	nan	nan	nan	nan	nan
3D ViT	6000	×	64	512	nan	0.00	0.00	0.00	nan	0.22
UNet	6000	×	-	512	0.98	0.21	0.96	0.20	0.98	0.22
UNet-ViT	6000	×	64	1024	0.70	0.21	0.35	0.19	0.86	0.21
UNet-ViT	6000	✓	64	1024	0.00	0.18	0.00	0.37	0.03	0.15

Table 1: Quantitative results

coefficient (DSC), recall, and precision produce a NaN result. Observing our models, the base UNet generally produces the best results, achieving significant performance differences compared to other models. This difference is illustrated in the ink reconstructions.

6 Discussion and Conclusion

We were excited to come up with and try out a variety of models, but sadly, most models experienced severe instability during training. 3D ViT simply failed to converge during overfitting tests. It rapidly degenerated and predicted every pixel as no-ink.

The recreated 3DCNN suffered from very similar issues. In cases where the model was able to generate an output, the image was very noisy. More often than not the model would fail to output a valid image. In these degenerate labellings, none or all of the pixels would be classified as containing ink.

UNet-ViT showed promising results in the overfitting test, even though it failed to generalize well on the validation set. The instability of the model can be further illustrated by comparing its performance with the addition of random flips. While UNet-Vit converged successfully on the training set, it was unable to converge on a slightly more difficult setting with data augmentation. We suspect the instability is caused by the imbalance of positive and negative pixels. We further attempted to combat the data imbalance by incorporating focal loss and dice loss, which sadly yielded similar unstable results.

Most surprisingly, the model that outperforms the rest is vanilla UNet. UNet converged almost instantly and was able to fit to the training set perfectly. Evidently, none of our models were able to generalize to unseen images. We offer two possible reasons: (i) the training dataset is extremely small, containing only three fragments, (ii) the conditions under which the fragments were scanned could be very different for every fragment, meaning that the ink could reside in a varying range of slices.

Acknowledgments

We acknowledge the sponsors of the Vesuvius Challenge for hosting this challenge. We also acknowledge their support for the dataset, sample code, and inspiration from the strong community that tackles this challenge. Additionally, we thank Kaggle for providing many GPU hours to support our research.

References

Bukreeva, I, A Mittone, A Bravin, G Festa, M Alessandrelli, P Coan, V Formoso, R G Agostino, M Giocondo, F Ciuchi, M Fratini, L Massimi, A Lamarra, C Andreani, R Bartolino, G Gigli, G Ranocchia, and A Cedola (June 2016). "Virtual unrolling and deciphering of Herculaneum papyri by X-ray phase-contrast tomography." en. *Sci. Rep.* 6, p. 27227. DOI: https://doi.org/10.1038/srep27227.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (Oct. 2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." DOI: http://arxiv.org/abs/2010.11929. arXiv: 2010.11929 [cs.CV].

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick (Nov. 2021). "Masked Autoencoders Are Scalable Vision Learners." DOI: http://arxiv.org/abs/2111.06377. arXiv: 2111.06377 [cs.CV].

Mocella, Vito, Emmanuel Brun, Claudio Ferrero, and Daniel Delattre (Jan. 2015). "Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging." en. *Nat. Commun.* 6, p. 5895. DOI: https://doi.org/10.1038/ncomms6895.

Parker, Clifford Seth, Stephen Parsons, Jack Bandy, Christy Chapman, Frederik Coppens, and William Brent Seales (May 2019). "From invisibility to readability: Recovering the ink of Herculaneum." en. *PLoS One* 14.5, e0215775. DOI: https://doi.org/10.1371/journal.pone.0215775.

Petit, Olivier, Nicolas Thome, Clément Rambour, and Luc Soler (Mar. 2021). "U-Net Transformer: Self and Cross Attention for Medical Image Segmentation." DOI: https://doi.org/10.48550/arXiv.2103.06104. arXiv: 2103.06104 [eess.IV].

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (May 2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation." DOI: https://doi.org/10.48550/arXiv.1505.04597. arXiv: 1505.04597 [cs.CV].

Wu, Haiping, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang (Mar. 2021). "CvT: Introducing Convolutions to Vision Transformers." DOI: http://arxiv.org/abs/2103.15808. arXiv: 2103.15808 [cs.CV].