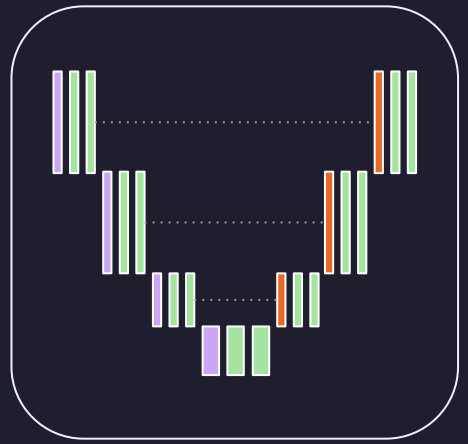
Semantic Segmentation

CNNs, Autoencoders, Skip Connections and Attention Mechanisms

➜ whoami

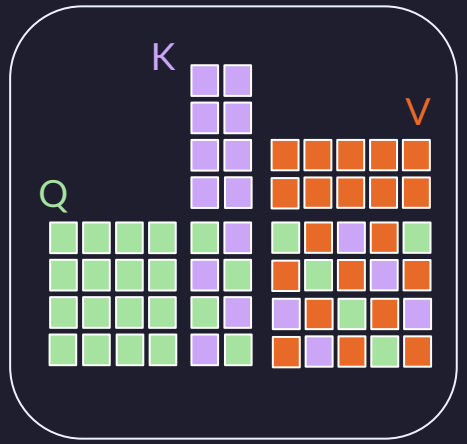
Gabriel Rodríguez de los Reyes
Joaquín Badillo Granillo

High-Level Overview



UNET

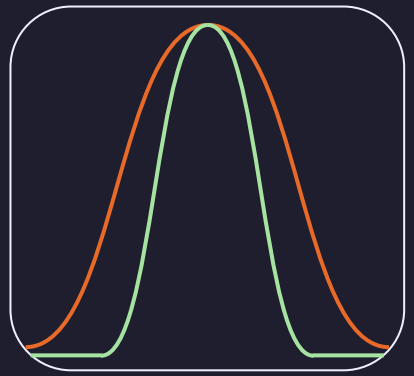
SEGNET



CROSS ATTENTION

Loss Functions

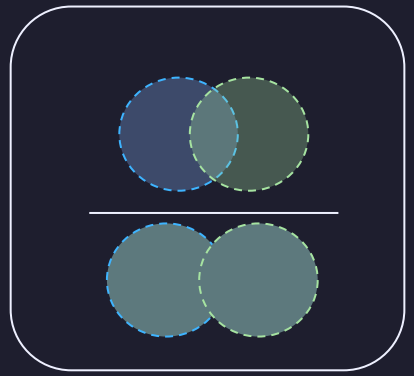
$$\mathcal{L} = H(P, Q)$$



Cross Entropy

In segmentation tasks it measure how well the model's predictions match the target labels [1].

Minimizing H implies minimizing KL Divergence.



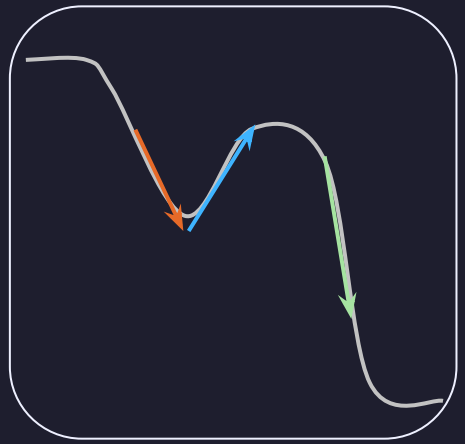
$$\mathcal{L} = d(X, Y)$$

Dice-Sørensen Coefficient

Measures the similarity between 2 samples. Similar to intersection over union

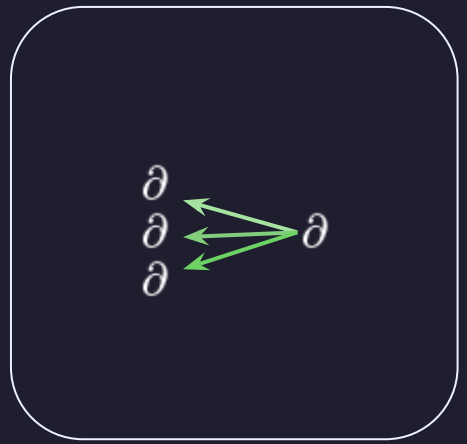
[1] https://arxiv.org/html/2312.05391v1#:~:text=Cross%2Dentropy%20(CE)%20measures,predictions%20match%20the%20target%20labels.

Optimizer



Adam

Stochastic optimization algorithm that uses “momentum” and root mean square propagation to escape local minima and adapt learning rates dynamically.



Backpropagation

[PyTorch] Lightning executes backpropagation after each call to training_step, that’s why loss is returned.

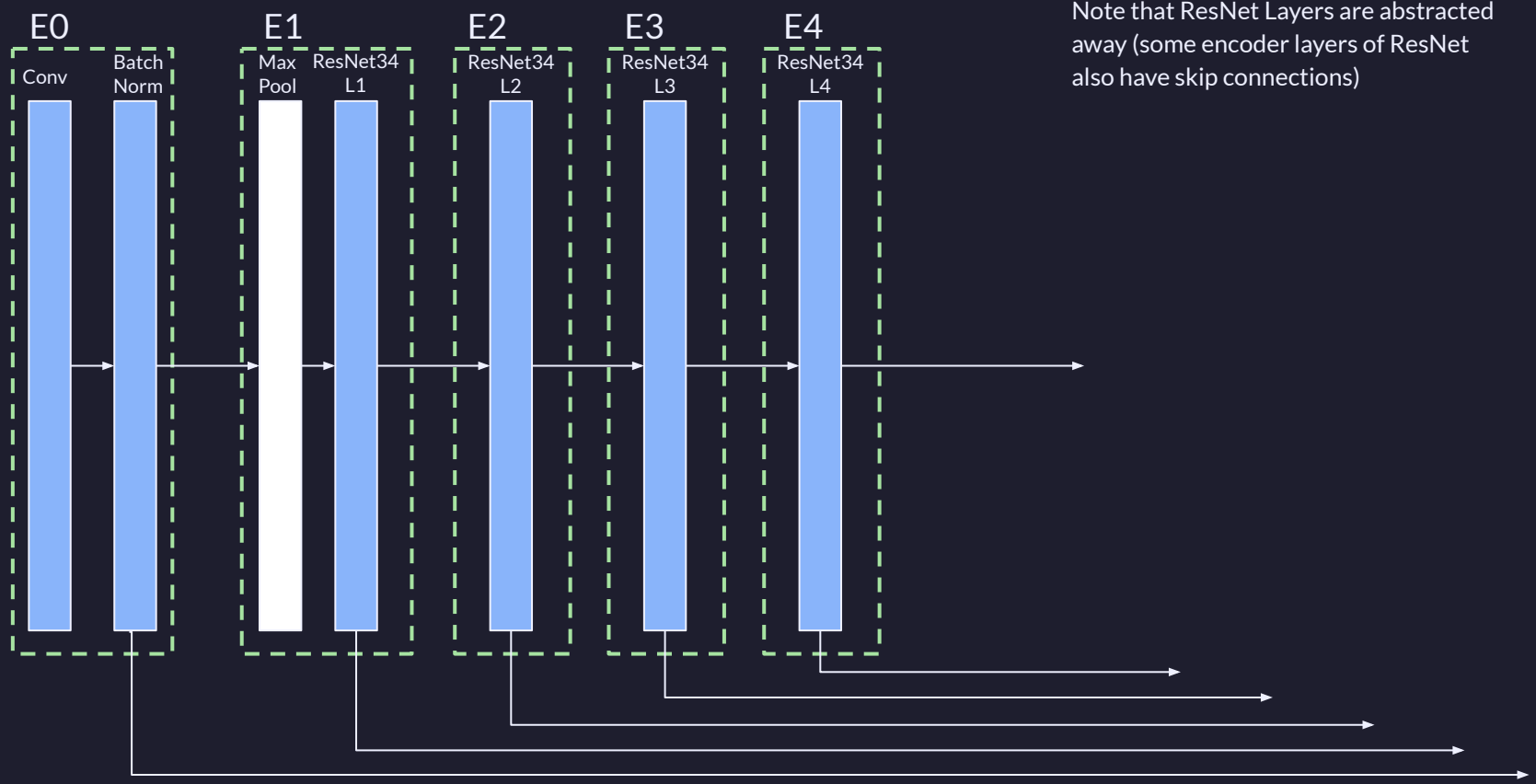
We added a learning rate scheduler so that validation set can also affect LR.

01

UNet

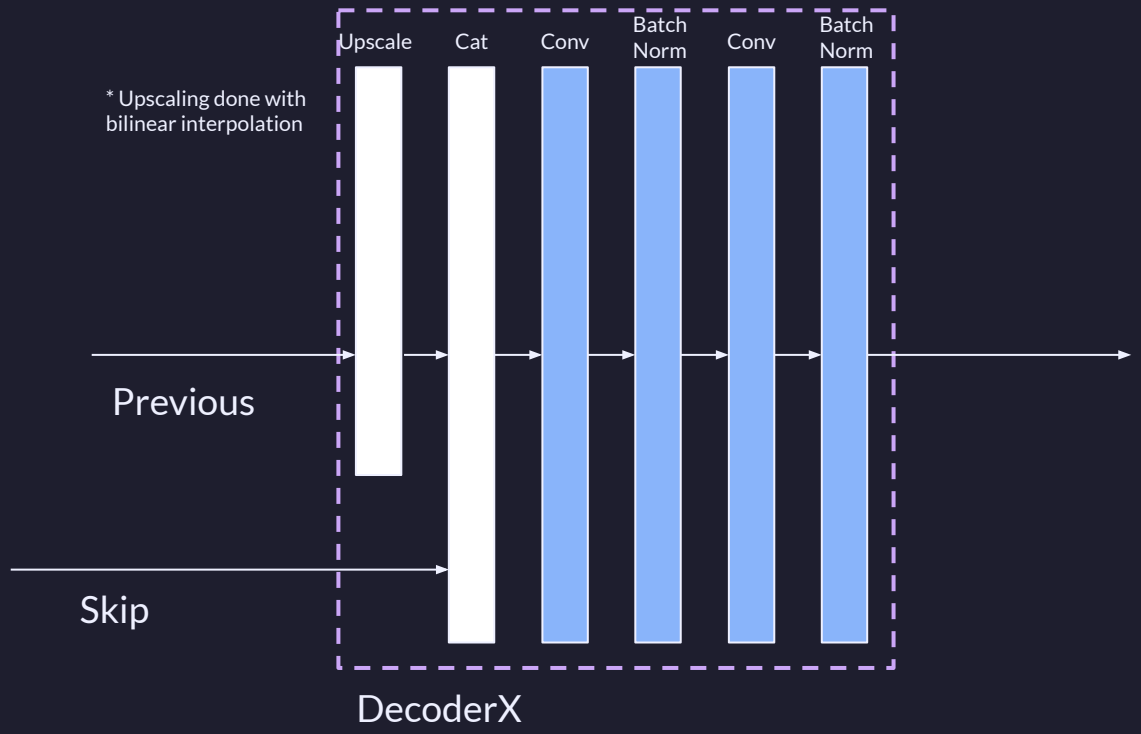
> An idiot admires complexity, a genius admires simplicity
- Terry Davis

Encoder Blocks

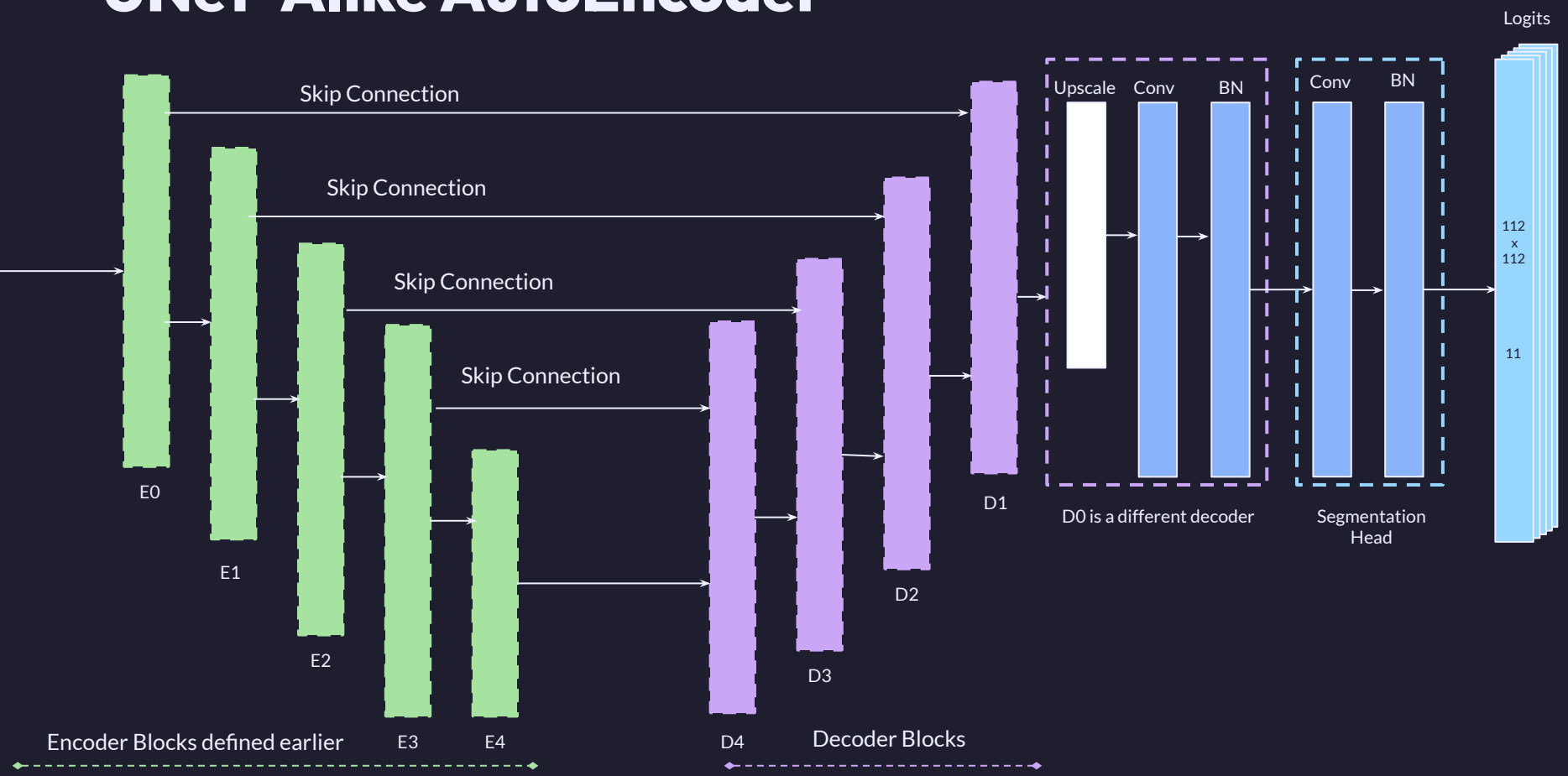


Note that ResNet Layers are abstracted away (some encoder layers of ResNet also have skip connections)

Decoder Block



UNet-Alike AutoEncoder

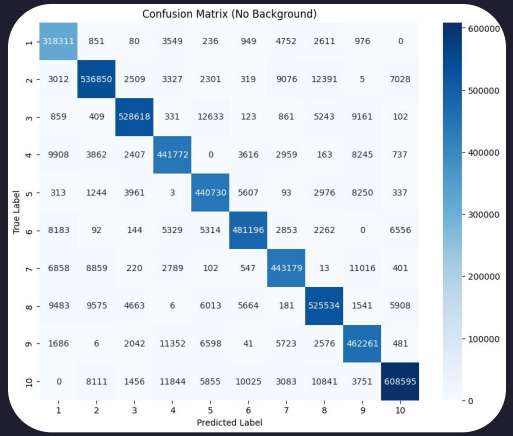


Results: Model Performance

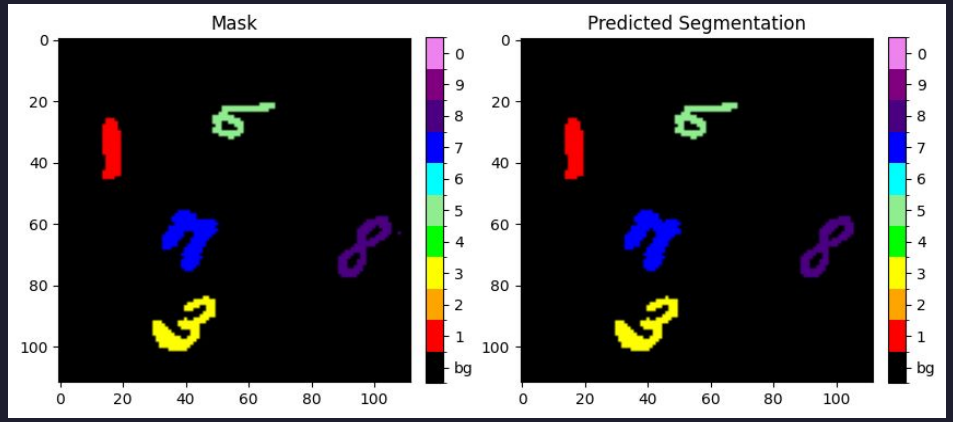
	Prec	Recall	F1	Acc
1	0.920	0.957	0.938	0.957
2	0.930	0.931	0.931	0.931
3	0.957	0.936	0.946	0.936
4	0.922	0.941	0.931	0.941
5	0.926	0.942	0.934	0.942
6	0.942	0.929	0.935	0.929
7	0.948	0.944	0.946	0.944
8	0.927	0.941	0.934	0.941
9	0.942	0.946	0.939	0.936
0	0.954	0.932	0.943	0.932

$$\mathcal{L} = H(P, Q)$$

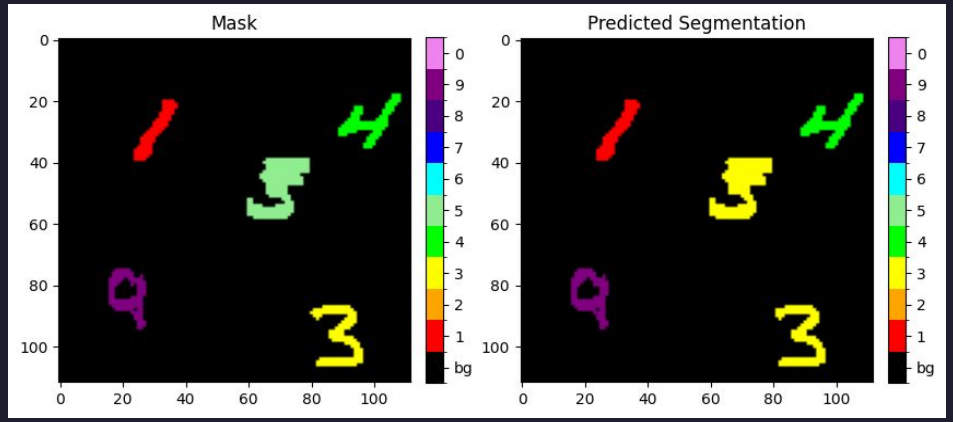
Cross Entropy.



Results: Labels vs Predictions



Good prediction



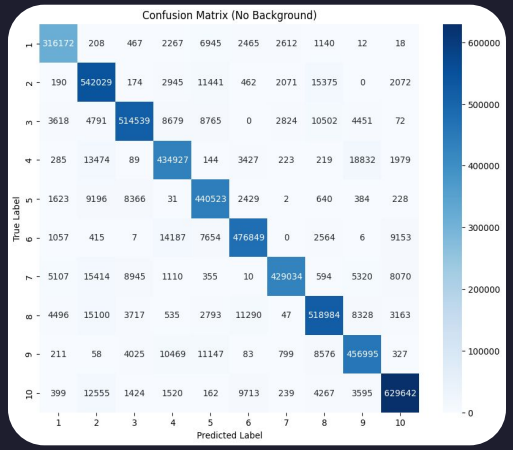
This could suggest the model did not overfit, since the 5 is a really bad sample

Results: Model Performance

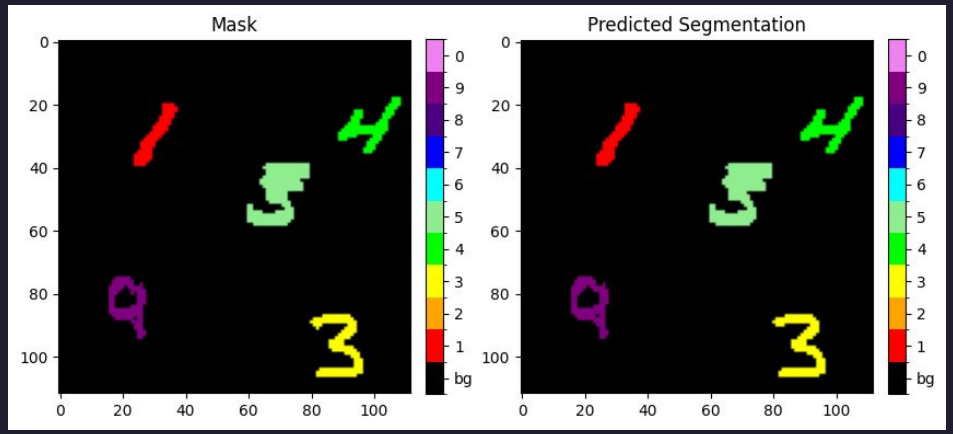
	Prec	Recall	F1	Acc
1	0.949	0.951	0.950	0.951
2	0.884	0.940	0.911	0.940
3	0.950	0.922	0.935	0.922
4	0.912	0.918	0.915	0.918
5	0.899	0.950	0.924	0.950
6	0.941	0.931	0.936	0.931
7	0.980	0.905	0.941	0.905
8	0.922	0.913	0.917	0.913
9	0.918	0.927	0.923	0.927
0	0.962	0.949	0.955	0.949

$$\mathcal{L} = d(X, Y)$$

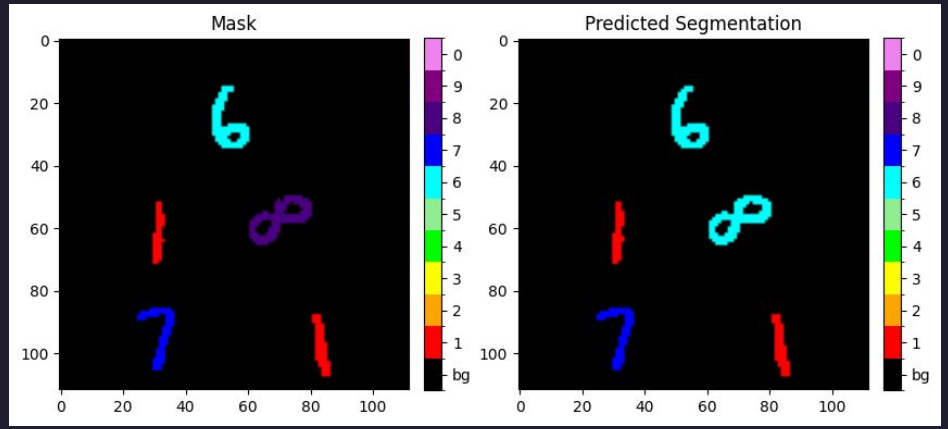
Dice-Sørensen coefficient.



Results



Suspicious result



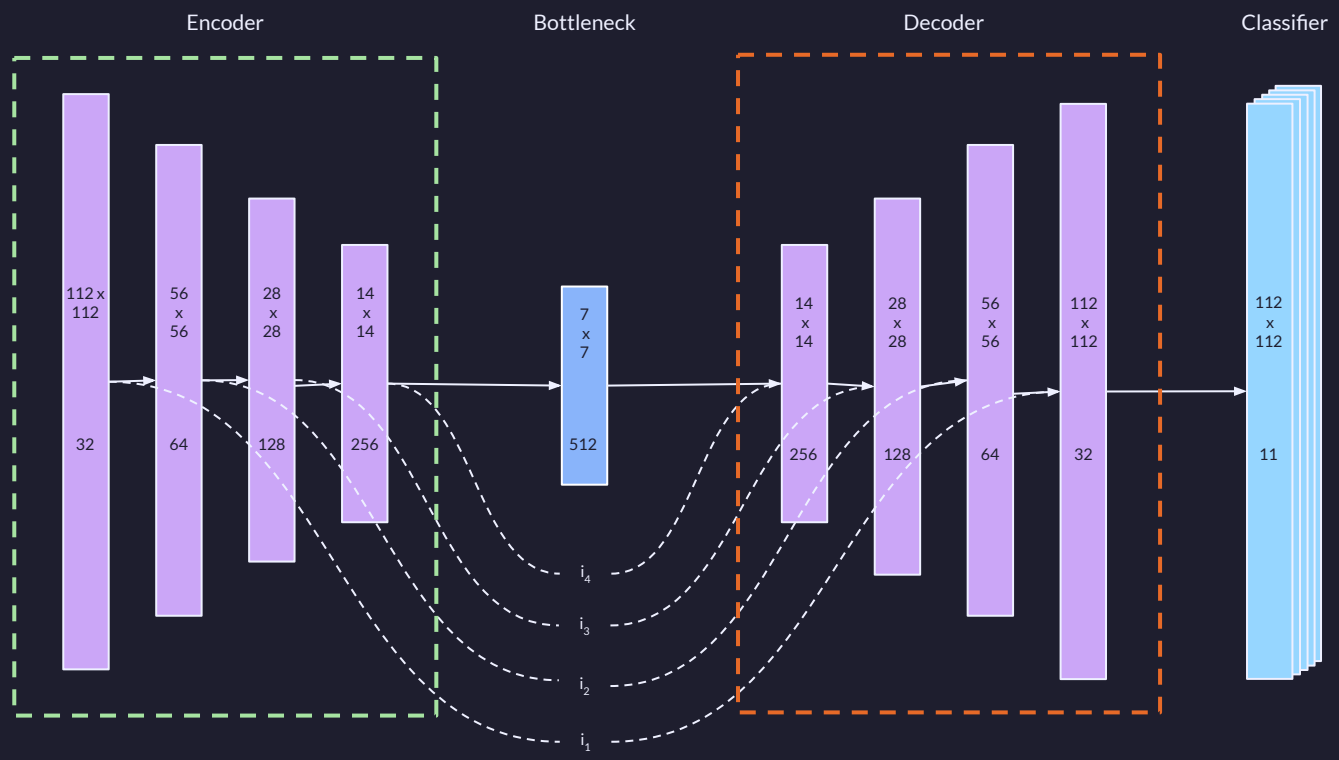
Trouble with rotations?
Or is it overfitting to an image with two 6s?

02

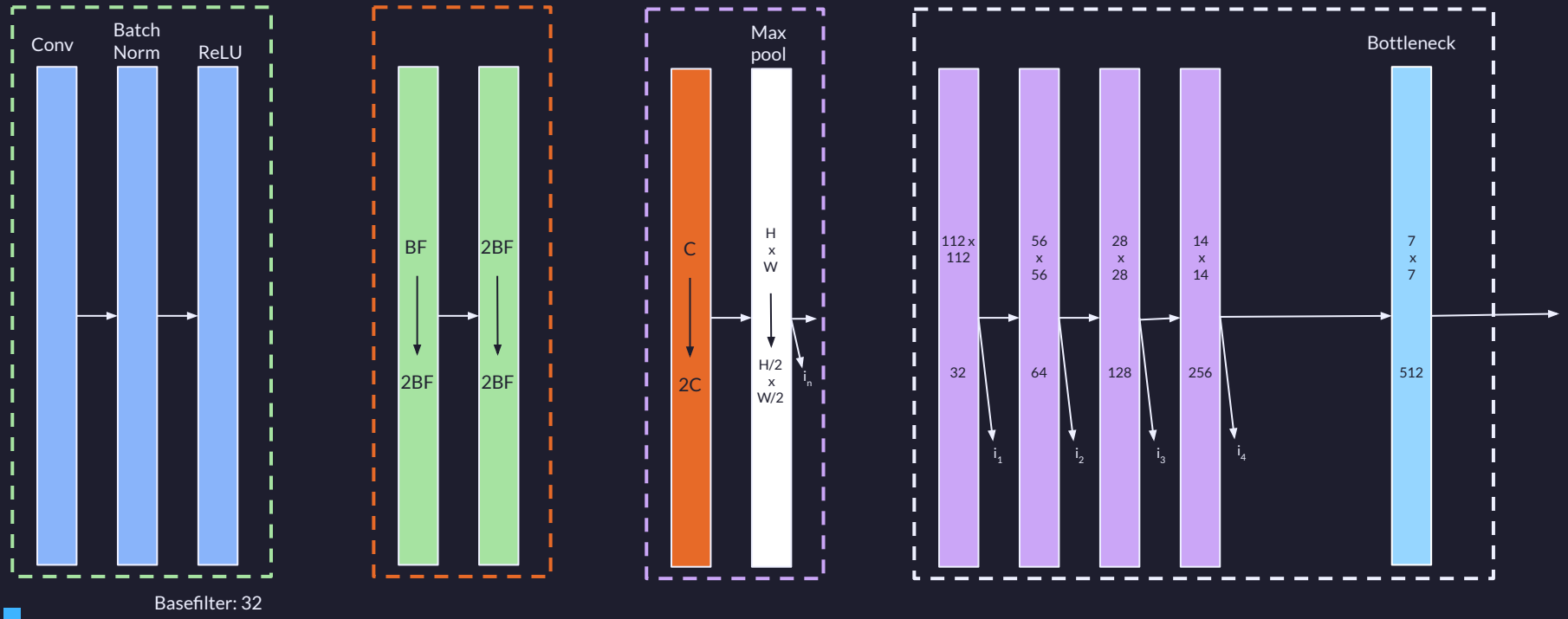
SegNet

> Simplicity is the ultimate sophistication.
- Leonardo da Vinci

Architecture

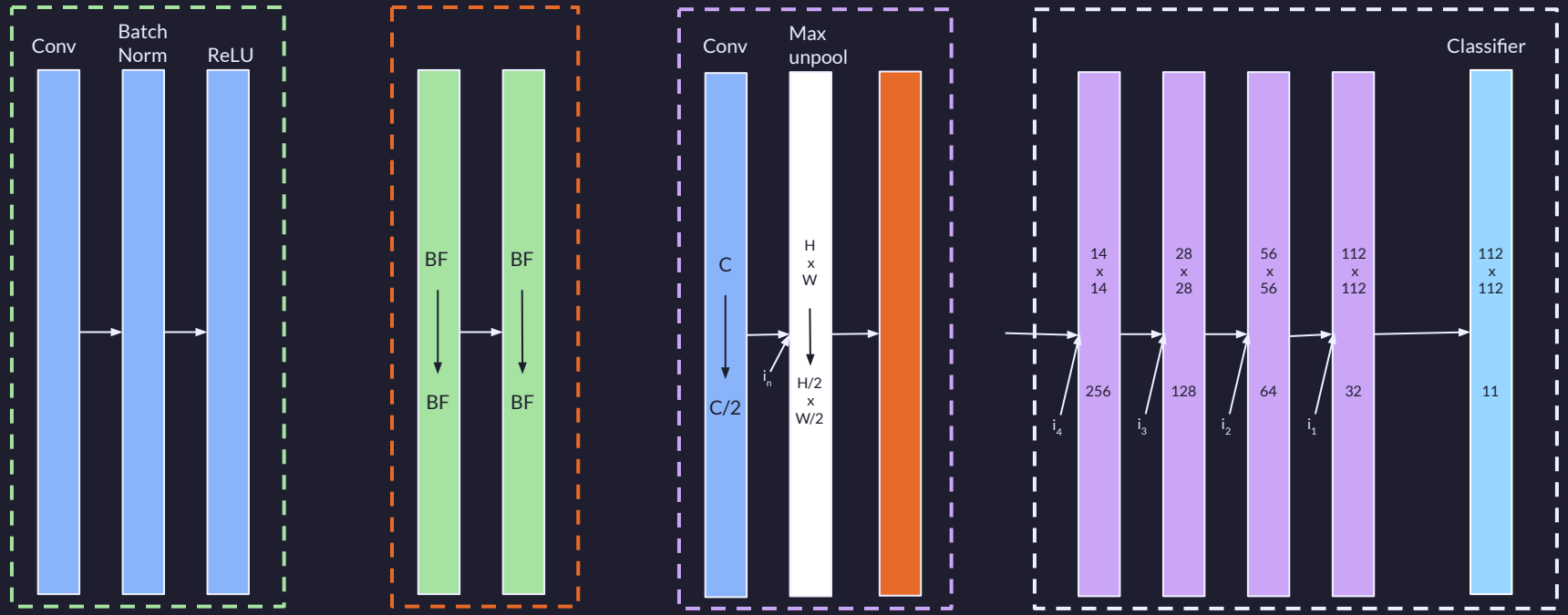


Encoder Blocks



Basefilter: 32

Decoder Blocks

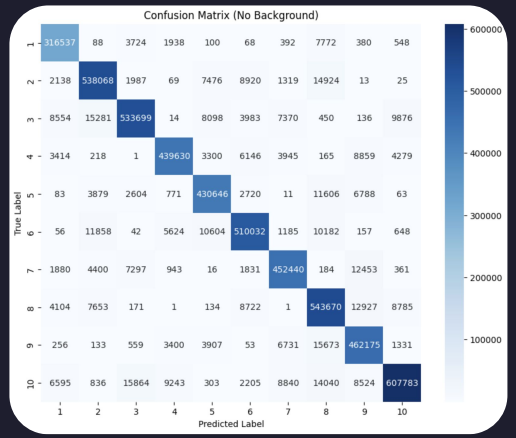


Results: Model Performance

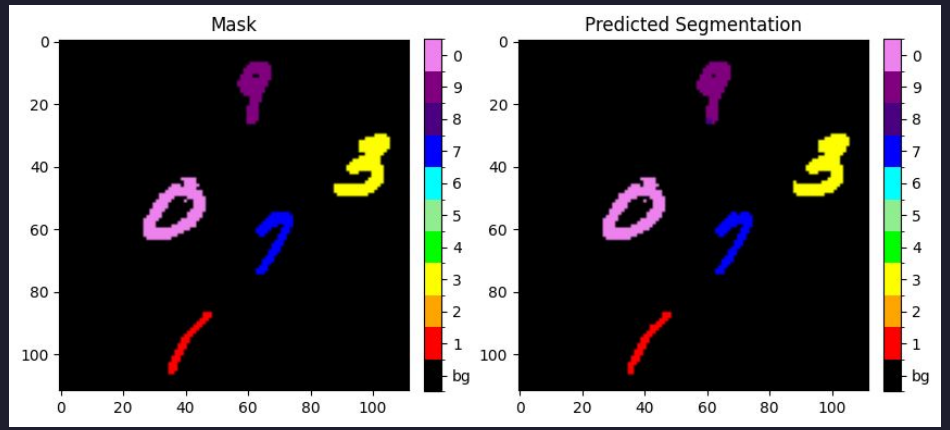
	Prec	Recall	F1	Acc
1	0.921	0.954	0.937	0.954
2	0.924	0.935	0.929	0.935
3	0.943	0.907	0.925	0.907
4	0.952	0.935	0.943	0.935
5	0.927	0.937	0.932	0.937
6	0.936	0.926	0.931	0.926
7	0.938	0.938	0.938	0.938
8	0.879	0.926	0.902	0.926
9	0.902	0.934	0.918	0.934
0	0.959	0.901	0.929	0.901

$$\mathcal{L} = d(X, Y)$$

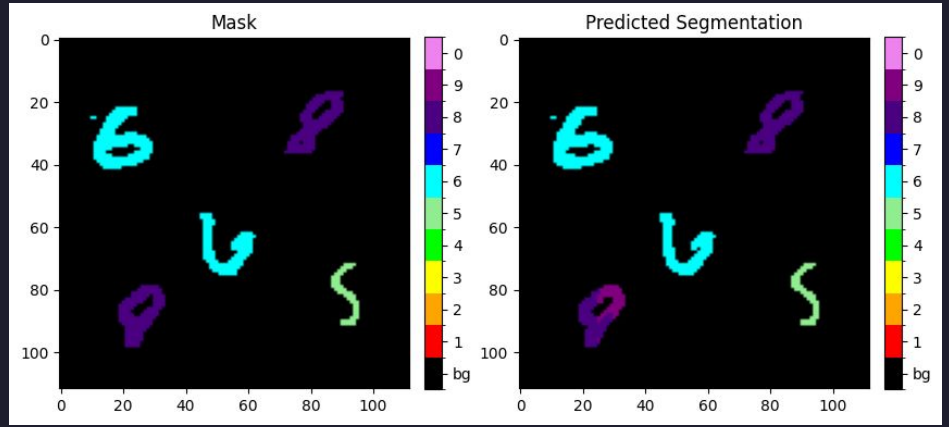
Dice-Sørensen coefficient.



Results



Almost pixel perfect, interesting for the model to match the noise



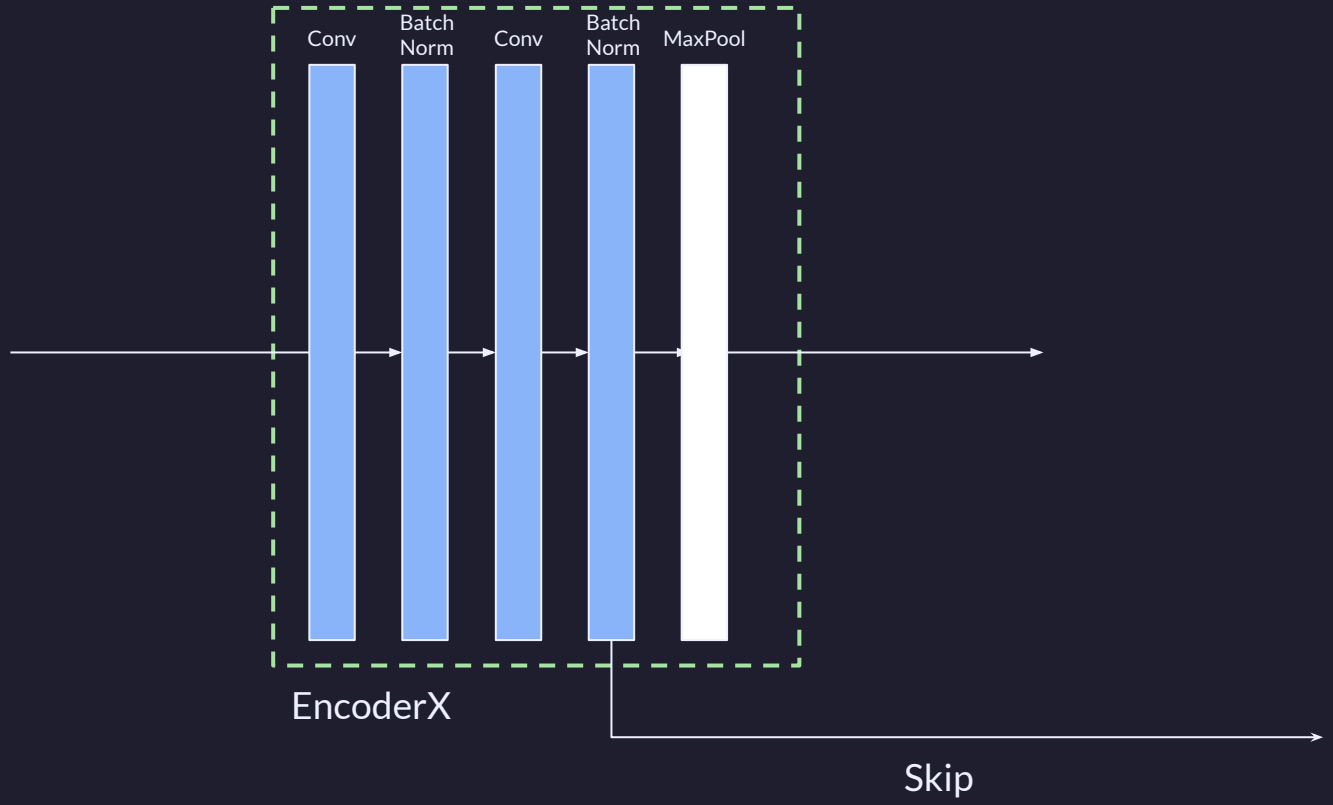
Found both a 9 and an 8 in the bottom left number (got combined after softmax → argmax). Cannot blame the model, the number looks like a 9 but it is labeled like an 8.

03

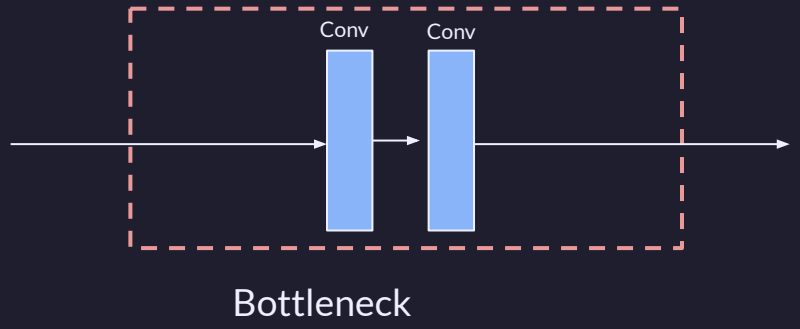
Attention Autoencoder

> Attention is all you need
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

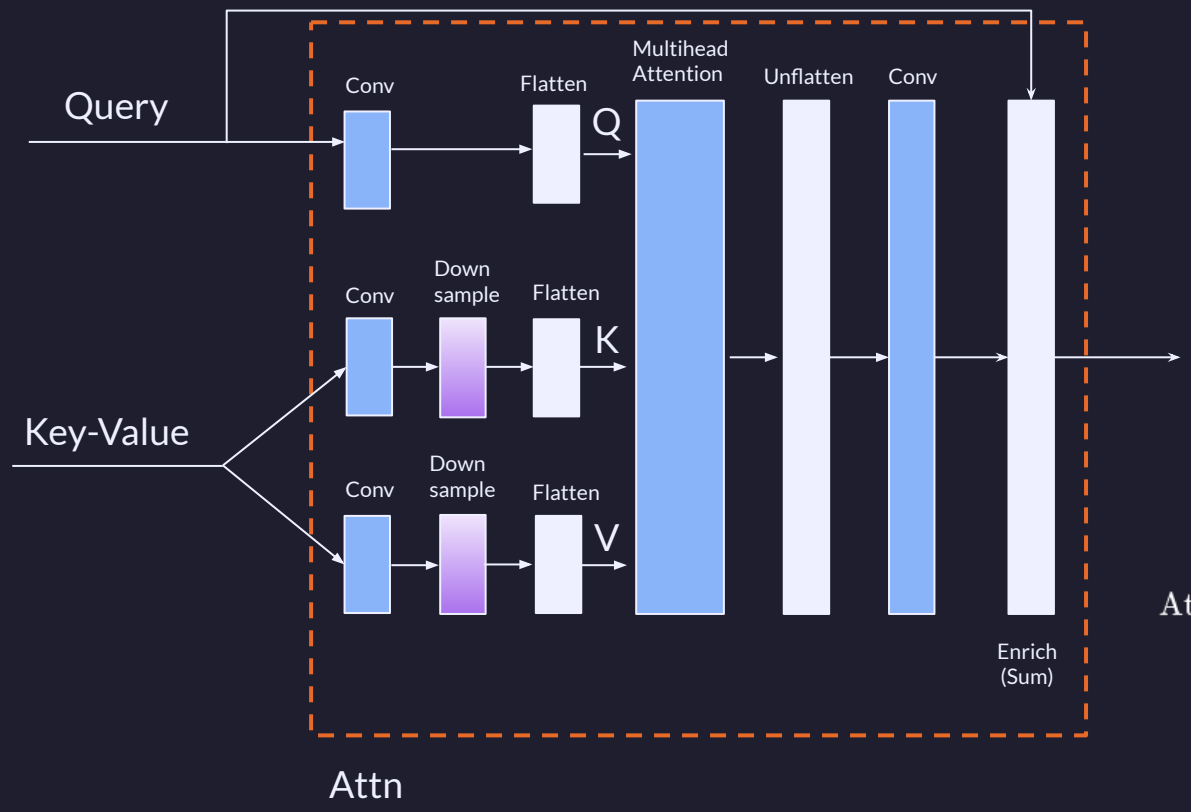
Encoder Block



Bottleneck Block



Custom Attention Block

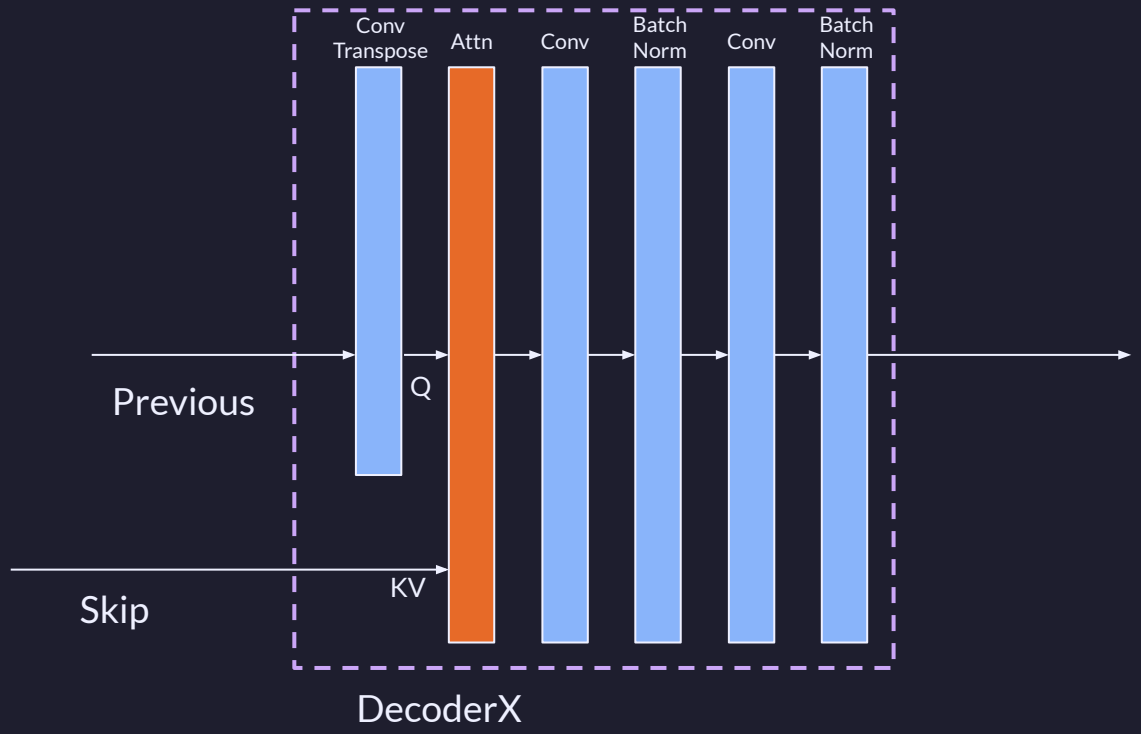


*White blocks are not trainable
- Flatten: view + permute
- Unflatten: permute + view
- Enrich: addition

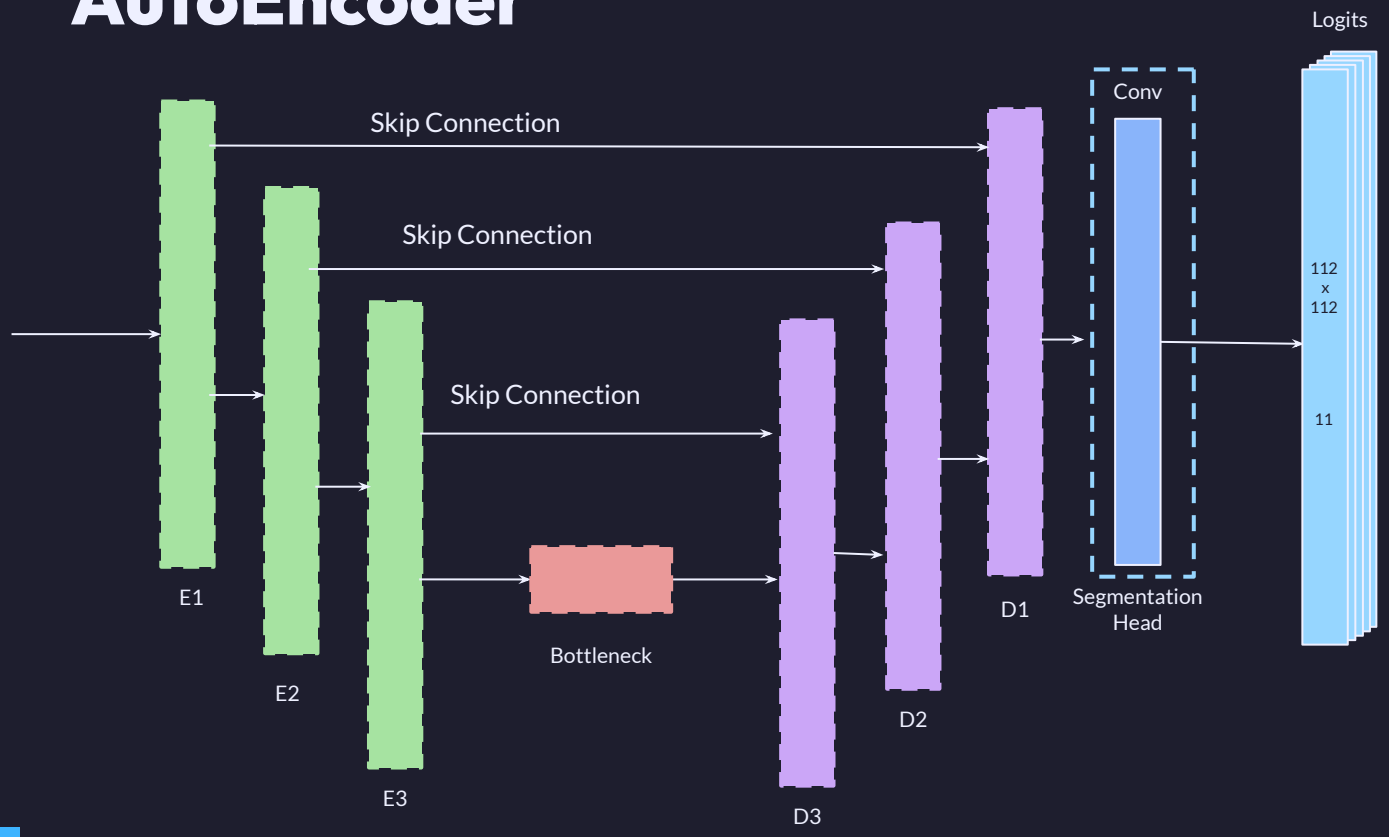
KV downsampling can be a convolutional layer or an identity function if set to 1

$$\text{Attention} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Decoder Block



AutoEncoder



Encoder Blocks

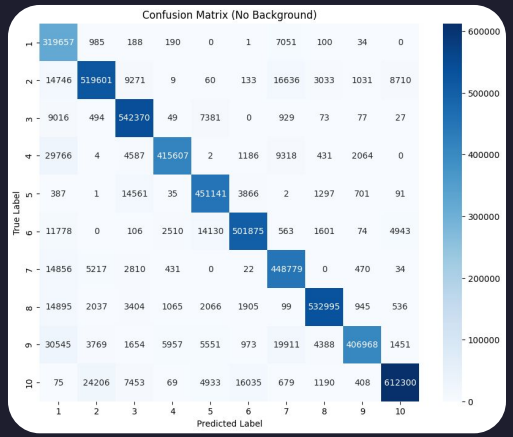
Decoder Blocks

Results: Model Performance

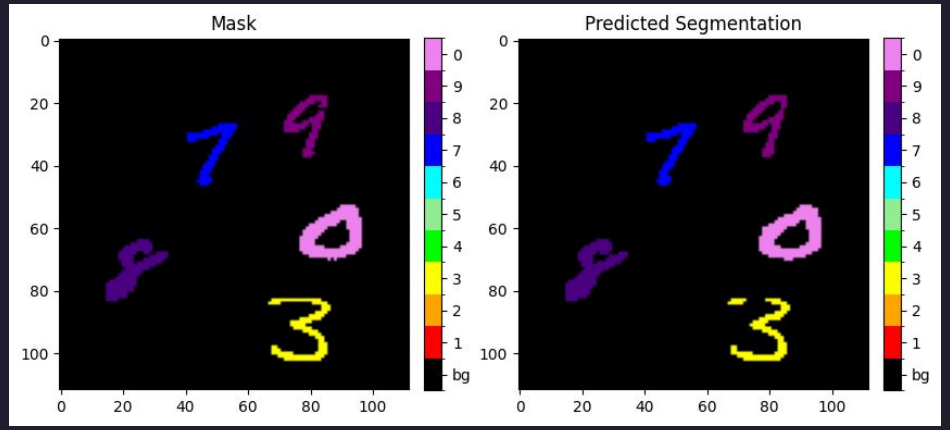
	Prec	Recall	F1	Acc
1	0.717	0.974	0.826	0.974
2	0.934	0.906	0.920	0.906
3	0.925	0.968	0.946	0.968
4	0.976	0.898	0.935	0.898
5	0.930	0.956	0.942	0.956
6	0.954	0.934	0.944	0.934
7	0.890	0.950	0.919	0.950
8	0.978	0.952	0.965	0.952
9	0.986	0.846	0.911	0.846
0	0.975	0.918	0.945	0.918

$$\mathcal{L} = H(P, Q)$$

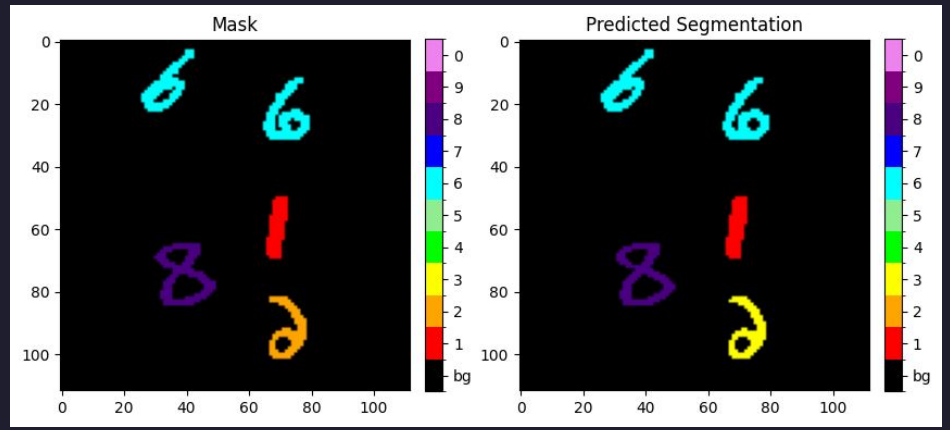
Cross Entropy.



Results



The 3 wasn't drawn continuously. Maybe it found more patterns with luminosity values?



Did the model find patterns suggesting a top and bottom curve usually lead to 3? Or is it overfitting to an image where a 3 was positioned there (with some relativity to other numbers)?

Check the Notebooks

https://drive.google.com/drive/folders/1MtfGj6q7DKcNUrRcE6IEnzLMgPpd4eis?usp=sharing