



# Quick, Count The Fingers!

Can we accurately create a model to distinguish between real and fake images?

# The Problem

- It's no secret that the influx of generative AI models such as Chat-GPT and Gemini has put the future of society in the air. It has spawned questions such as 'Will my job still exist?' and 'What's the point in learning this if some 12-year-old with Chat-GPT can do it in seconds?'
- We will focus on the image creation section of this topic, where AI-created images can be used for poking fun at individuals to a form of aggressive propaganda against others (take a look to the right)
- Today, we'll try and create a model to help prevent this by classifying between real and fake images. But to do this, we must understand what a neural network is and; more importantly, a convoluted neural network!



# Quick look into Neural Networks!



- A neural network can model non-linear relationships you see in real life by using a combination of linear layers (your classic operations in maths) and non-linear layers (think cosine, arg max)
- Essentially how it works is that we have a big data set ( $x$ ) and an output ( $y$ ). We can set up a neural network's layers to mimic the relationship between the variable  $x$  and the outcome  $y$ .
- We then split the dataset into two sets: validation and training. We feed the neural network the training set where it then learns about the details and intricacies. We then feed it the validation set (without  $y$ ) and then compare its results given by the neural network ( $\hat{y}$ ) to what it should be ( $y$ ). This result is called validation error and is our metric for accuracy





# Digging Deeper: Convolutional Neural Networks (CNN)

A CNN is a neural network which has two special features which will help us out with our image problem:

- Translation invariance: Recognises an image no matter where it is on the page.
- Locality principle: Pixels close to each other are related

Do you remember the title where we joked about the fingers, well here if this is learned to be associated with fake images, both features are important in spotting this correctly classifying

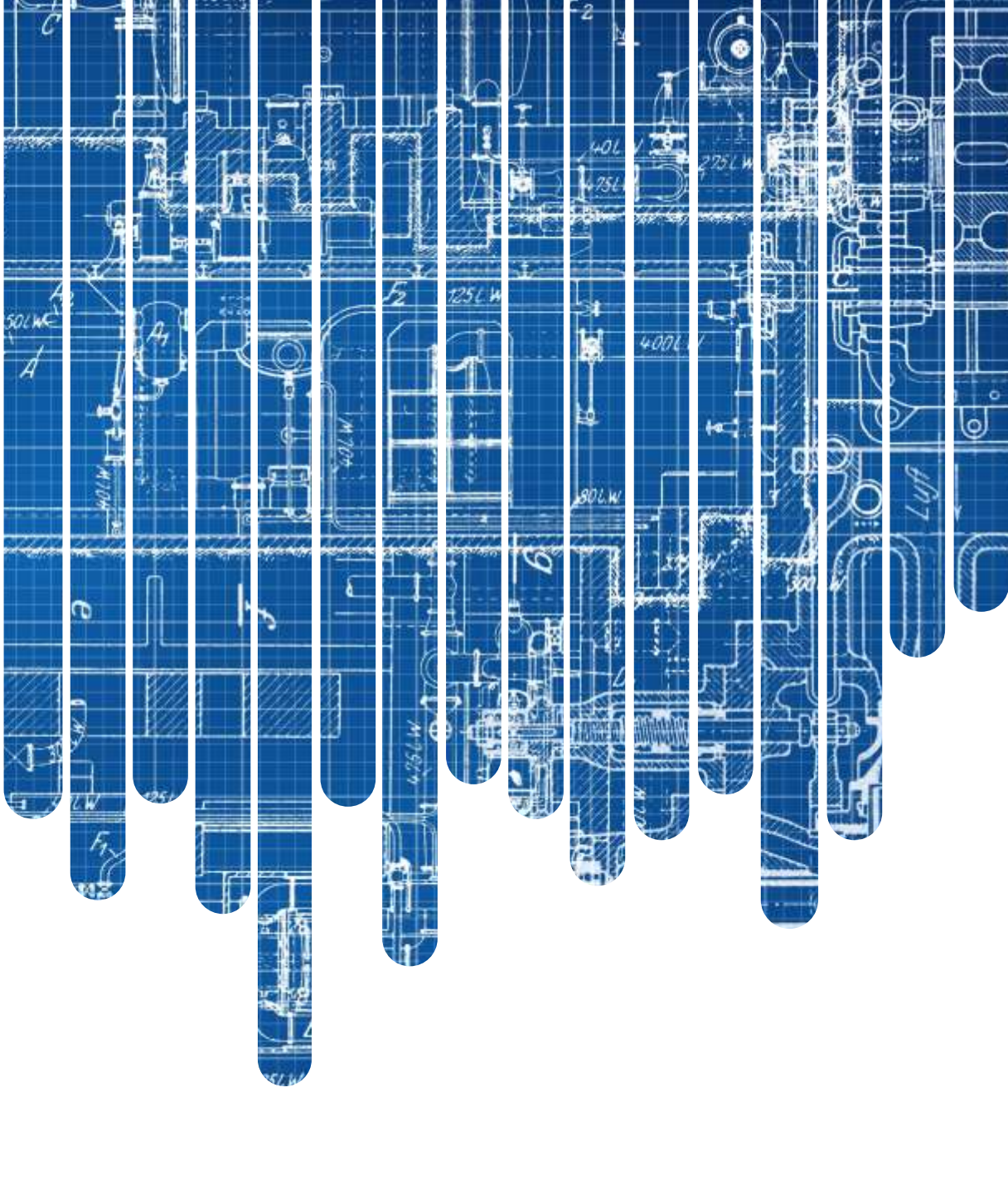


A CNN also has two cool layers which do different things.

1. Convolutional layer: for spotting patterns
2. Pooling layer: for decreasing information as pixels in an image take quite a lot of memory to store compared to tabular data

Each of these has its own parameters (such as learning rate and epoch) that change their behaviour

An important fact to know here is that **as we get deeper into the layers, the more intricate the detail the convolutional layers will pick up and train on.**



# Famous CNN Architectures we'll use!

We'll be using 3 famous CNN architectures for our project:

1. LeNet: Created in the late 80s to identify handwriting
2. AlexNet: Made in 2012 and is renowned for winning an ImageNet competition
3. ResNet-50: The more complicated out of the 3, created in 2015 by Microsoft
  - For us to train a massive model from scratch, it would take a much stronger GPU and database than any of us can afford! (☹)
  - So, we are going to need to use pre-trained models. Lucky for us, ResNet-50 and AlexNet have pre-trained models by AlexNet; LeNet is so small, we can build it ourselves and train from there.
  - Let's go ahead and check out the architecture of the 3 to see what's the hype



# The layer architecture of our models!

## LeNet

- With 5 layers featuring 3 convolutional layers and 2 pooling layer
- This very simple model most likely won't give us the most accurate results, but it will be a good reference point for the other two

## AlexNet

- Having 8 layers with 5 convolutional layers, this model can pick up a good number of details from the dataset's images.
- Unlike LeNet, this model is actually trained on images rather than handwriting so expect stronger accuracy

## ResNet-50

- Now we are hanging with the big boys! This model has 50 layers. Although, It has a complex architecture involving 'residual block'
- The most important feature is the 49 convolutional layers. That's 49 layers of feature spotting and training. We have to be careful of overfitting though (learning the noise of the training set which may not apply to the validation set).

# The Dataset

- We are using a dataset of 120,000 images.
- There is a 50/50 classification split in the images which means there are 60,000 real and 60,000 AI-generated images
- The training set contains 100,000 of the images, and the rest is in the test set.
- This dataset is huge and will take a long time to work with, however, they come from just 2023 so are generally up to date with the capabilities of generated images
- With all of the setup out of the way... let's get to the



# A

First, we'll split the dataset into training and validation sets



# B

Then run them through 3 differently specialised pre-trained CNNs and then choose which one has the most accurate validation accuracy.

[We do it for only 3 epochs (run-throughs) as we have a massive dataset with complicated CNNs]



# C

Zoom in on the winner and then see if we can make further improvements by changing some parameters



# D

Finally, give it a test drive to see how well it does on the whole dataset!

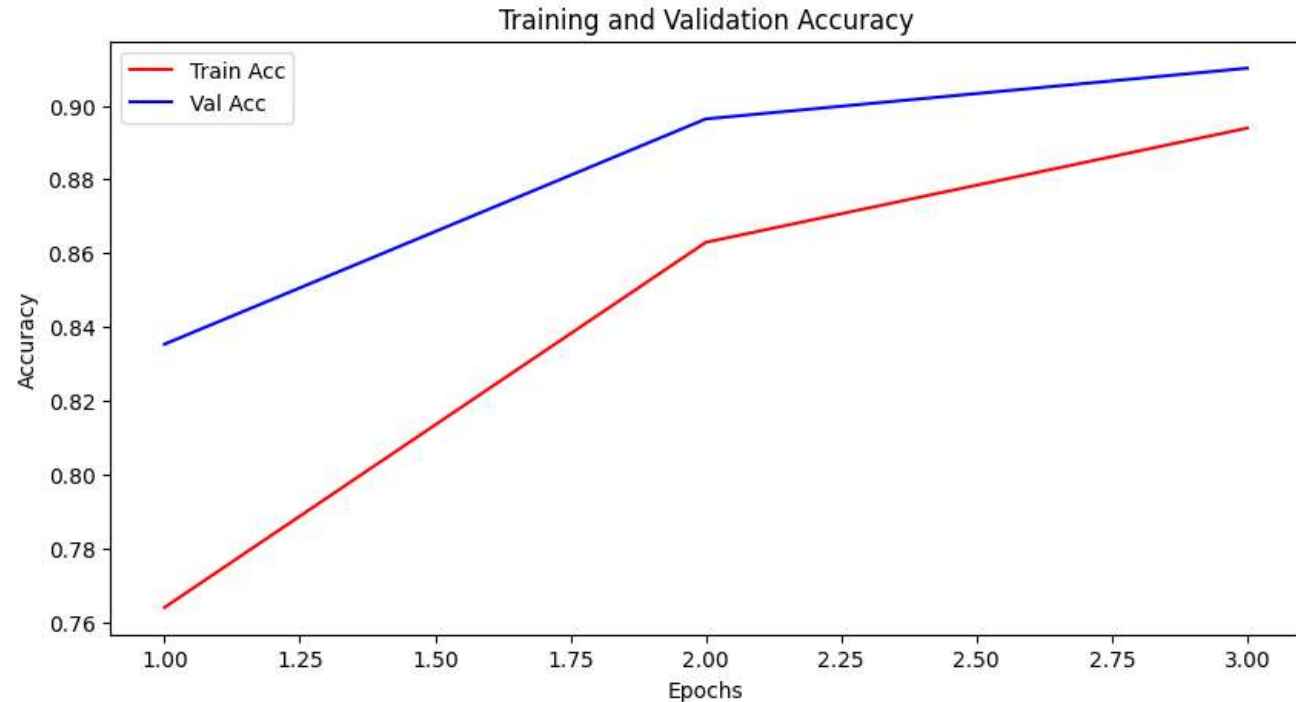


# Results From LeNet!

So, after training and feeding through the validation set, we get a final validation accuracy of 91.02%

Now, we expected this one to be the lowest, but this accuracy is actually not that bad for being trained on handwriting 40 years ago.

Not sure we'd consider this the 'high accuracy' we are looking for. I'd say that belongs to >95%



That's  
you!

A-B

C

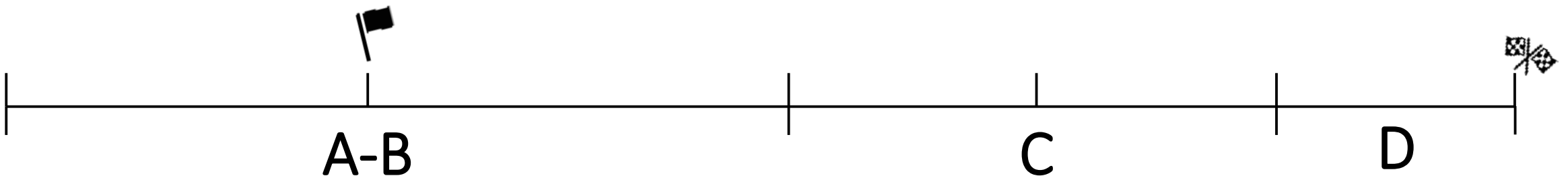
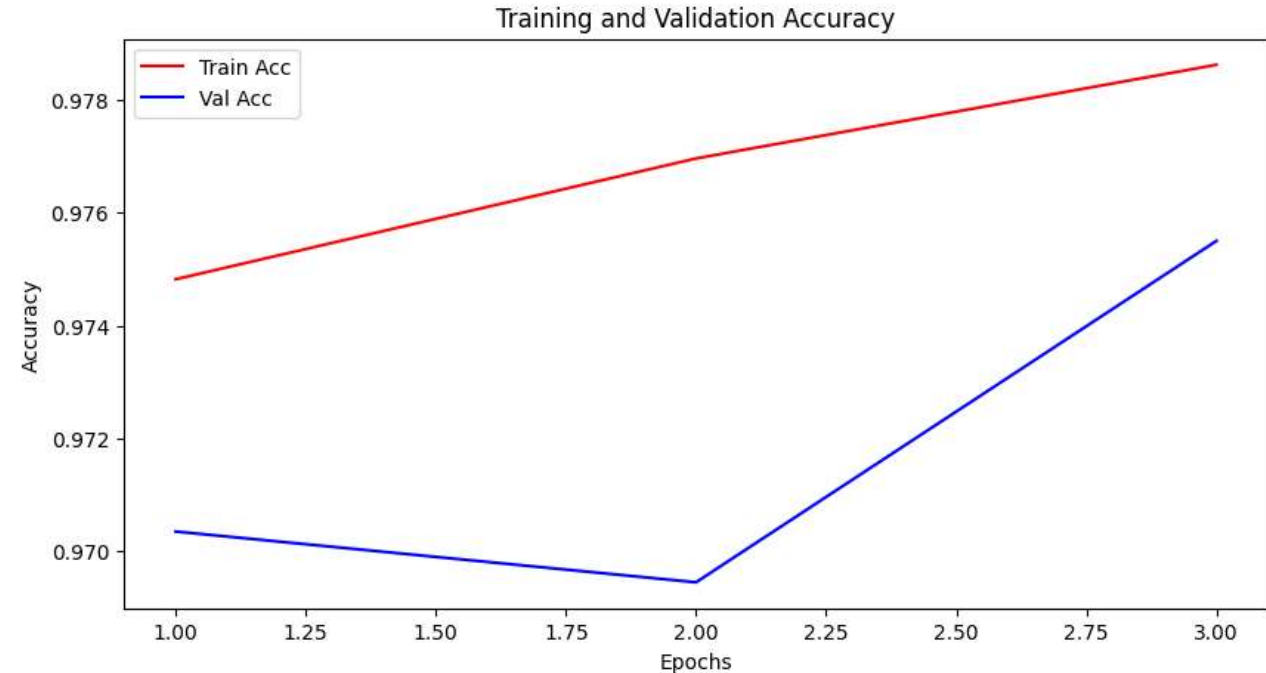
D

# Results From AlexNet

For AlexNet, we get a final validation accuracy of 97.55%

This is great and also doesn't show tell-tale signs of over/under fitting"

Let's see how ResNet-50 fends

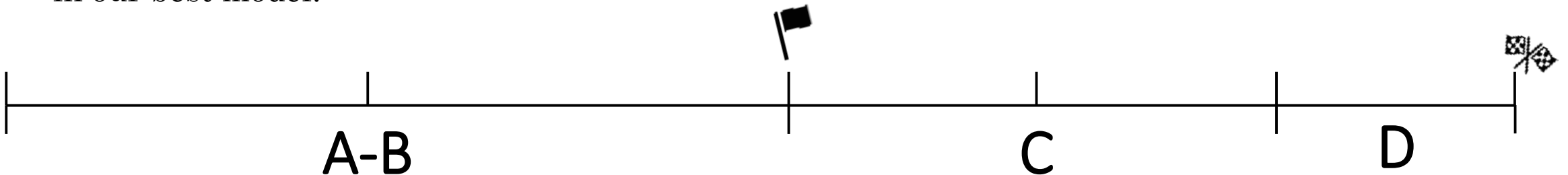
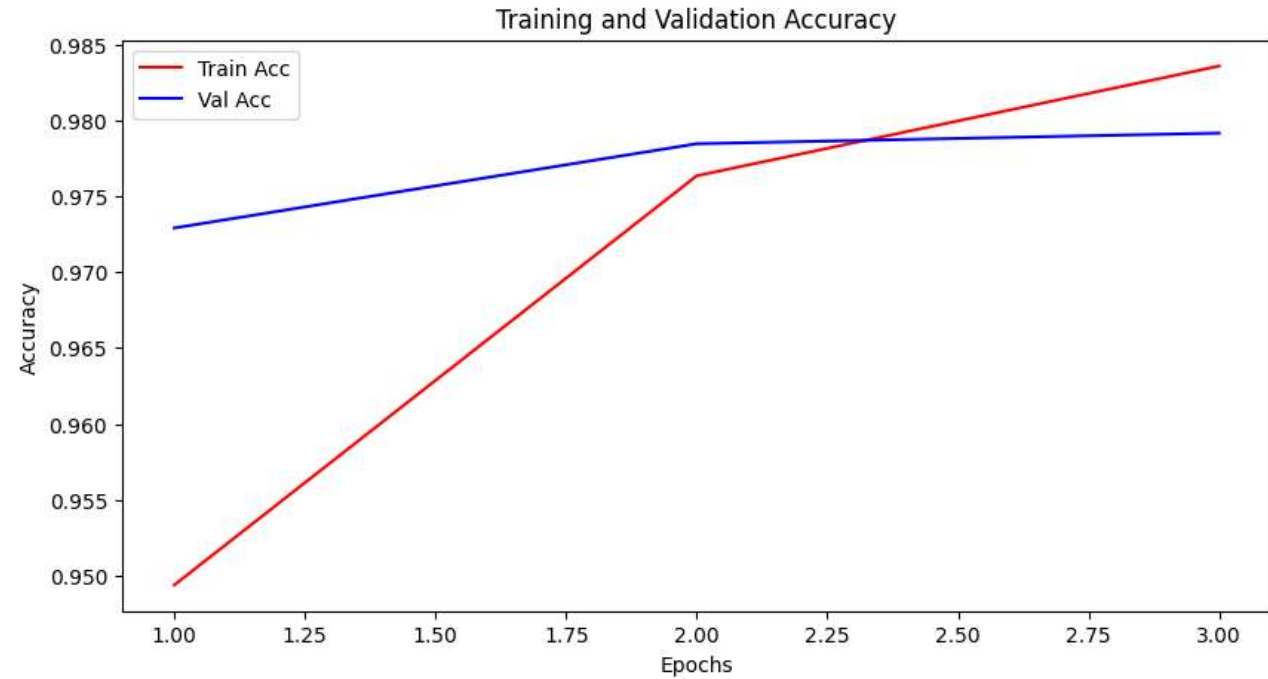


# Results from ResNet-50

Finally, our accuracy from the most complicated CNN is 97.9%

So ResNet-50 has the highest accuracy and hence we have our winner!

But we have some signs of underfitting as we end up having training accuracy overtaking validation accuracy. Well, that's why our next part is about fixing problems in our best model!

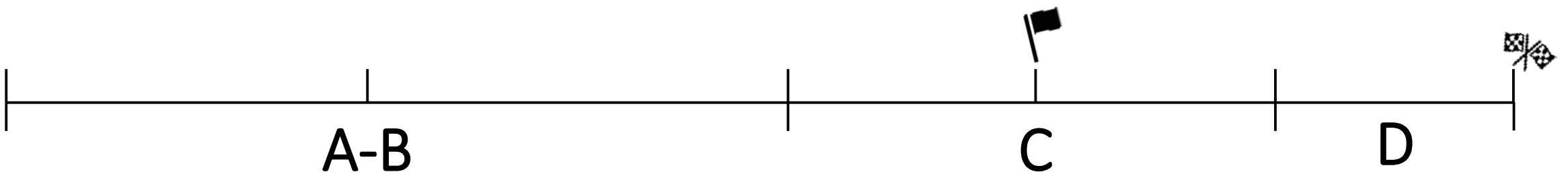
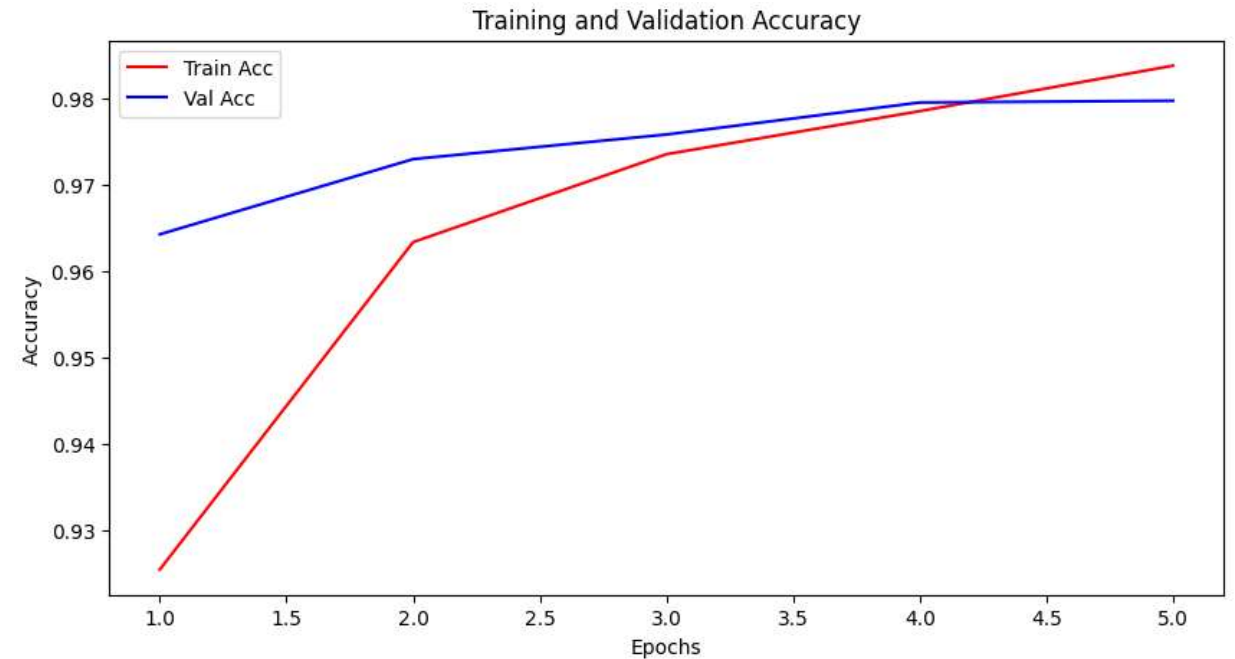




# Finetuning ResNet-50, Experiment 1

For experiment 1, we'll do the classic approach of getting a better fit by increasing the number of epochs from 3 to 5 (such a *big* change, we know. But each extra epoch takes hours to train). Also, let's decrease the learning rate to try and get a better local minima.

We end up getting a slightly lower accuracy at 97.65% (still better than AlexNet). Trading a small amount of accuracy for a *slightly* better-fitted model is something we'll take!

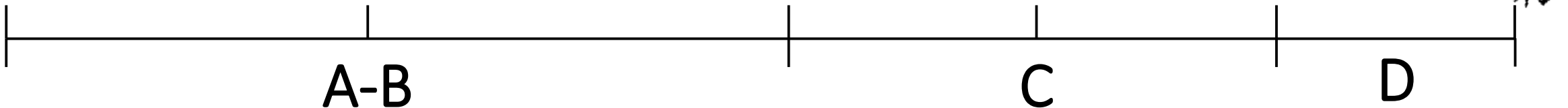
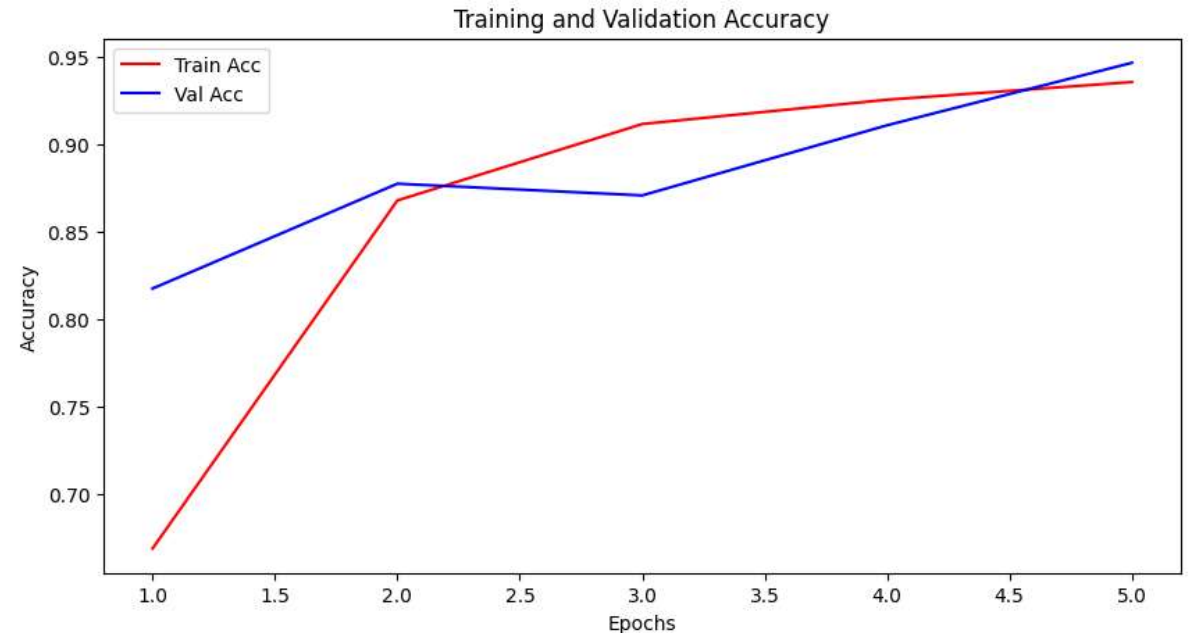


# Finetuning ResNet-50, Experiment 2

This time, we'll keep the epoch count at 5 but see what happens if we decrease the learning rate because fine-tuning neural networks as complex as this can lead to results by doing the things that you think wouldn't work.

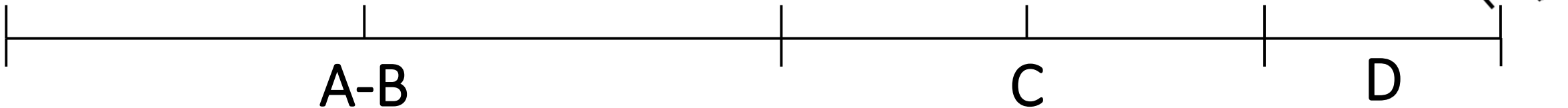
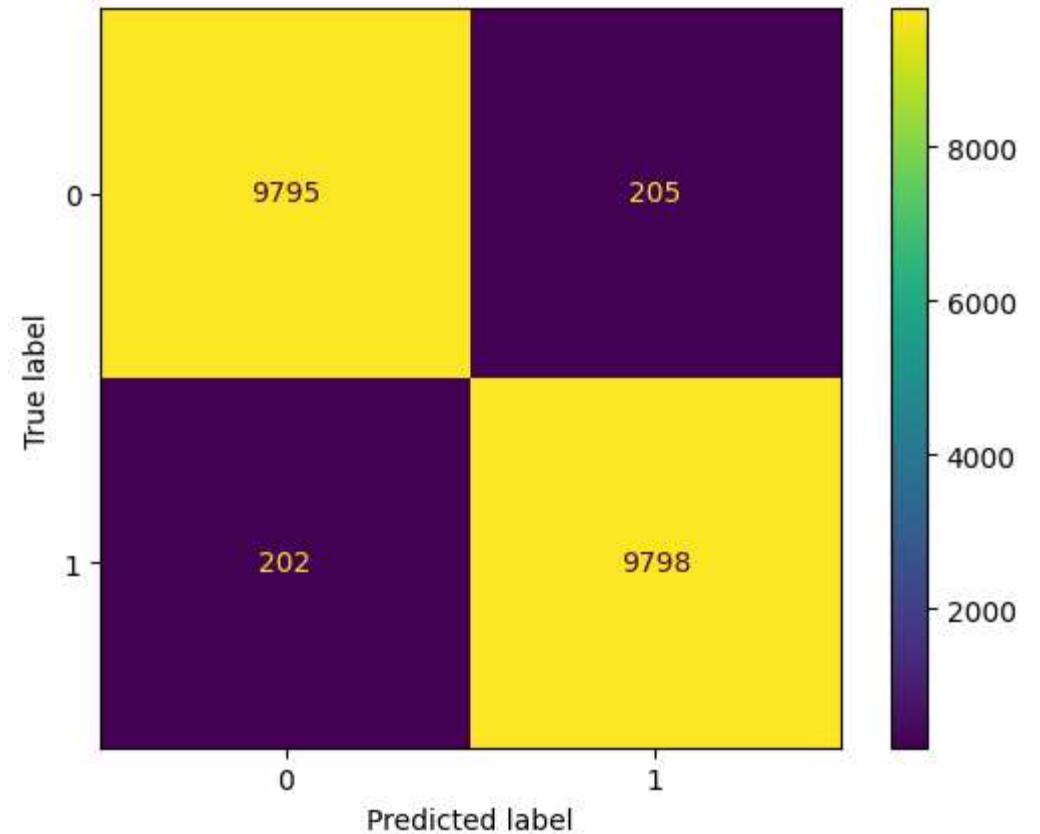
In fact, computer scientists often spend months at a time trying to understand why changing parameters leads to a certain change!

But we end up getting the worst validation accuracy of them all at 94.68%. This trade-off is too much for us to take. We'll take Experiment 1 as the chosen one!



# Test-Drive Time!

- Take a look at the confusion matrix on the right. We are concerned with the yellow squares.
- Out of 10,000 images, that's a pretty good outcome!
- And just like that, our objective is complete, an accurate model to distinguish between real and AI images!







# Final Thoughts

---

- So, our best model is a modified ResNet-50 with a validation accuracy of 97.65%. Note that this model still has signs of minor underfitting. Given time and resources, you can just run more epochs to slowly chip away and eventually get an even better model. But still, an accuracy of 97.65% is amazing and the model does fulfil our objective]
- In terms of future-proofing, generative models are only getting better. Our dataset is from last year and today, generative art through a process called stable diffusion is better than ever. Hence, this model will need to be done again with another more modern dataset to keep its validation accuracy high!