CMM536 Topic 9 - GPT

Content

1. (Vanilla) CNN is just the start!

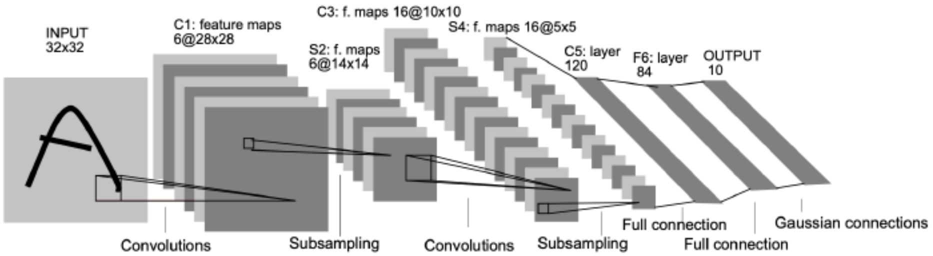
2. P: Pre-Trained Models → Transfer Learning

3. G: Generative Models

4. T: Transformers!

(Vanilla) CNN is just the start!

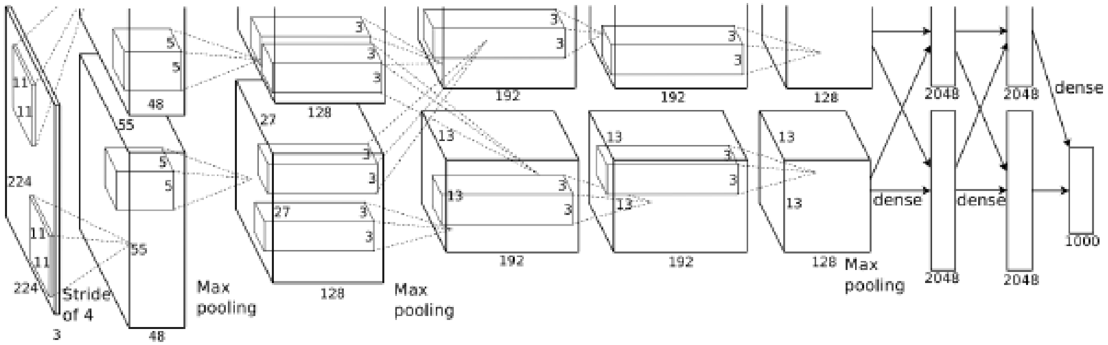
- The network we saw last week is often called “LeNet” or “Vanilla-CNN”
- It was “tailor-made” to solve the MNIST problem, although it could solve many more!
- The biggest drawback is that the filters used are too simple to tackle other challenges!



Y. LeCun et al., “Gradient-based learning applied to document recognition”. Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998, doi: https://doi.org/10.1109/5.726791

AlexNet

- Won the 2012 ImageNet ILSVRC challenge (by a large margin).
  - Achieved top error rate of 17% (second best achieved only 26%)
- It is much larger and deeper than LeNet, and the authors used dropout and data augmentation to reduce overfitting
  - It was the first network trained in a GPU, thanks to Krizhevsky's gaming expertise!



A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks". Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS) 2012, doi: https://dl.acm.org/doi/10.5555/2999134.2999257

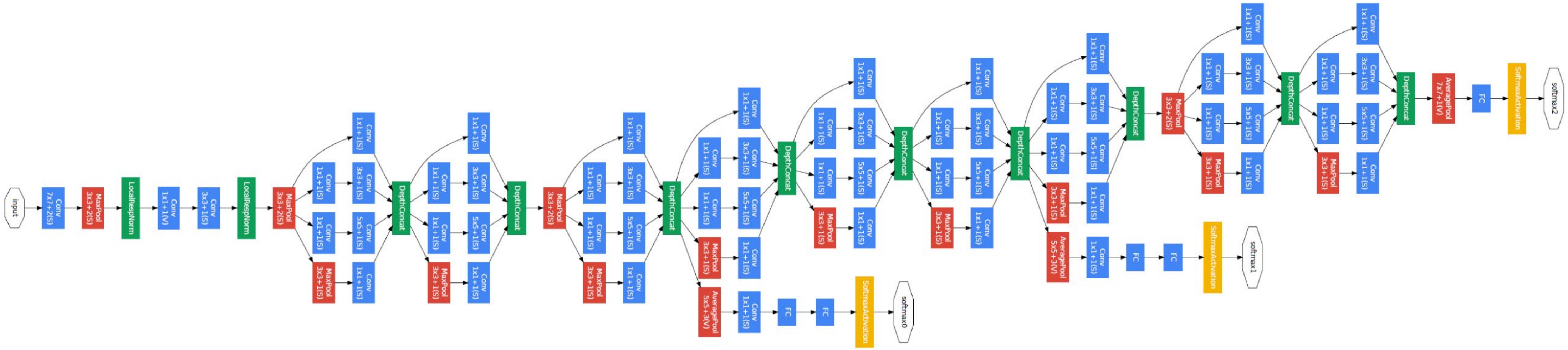
GoogLeNet

- Won the ILSVRC challenge in 2014, developed by Szegedy et al.
  - Much deeper network than previous CNNs (one early version is made of 22 conv layers!)
- Several extensions of GoogLeNet were developed later by Google researchers, most notably the Inception architectures
  - Allow the network to choose between multiple convolutional filter sizes in each block.
  - An Inception network stacks these modules on top of each other, with occasional max-pooling layers with stride 2 to halve the resolution of the grid

---

C. Szegedy et al., “Going deeper with convolutions“. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, doi: https://doi.org/10.1109/CVPR.2015.7298594

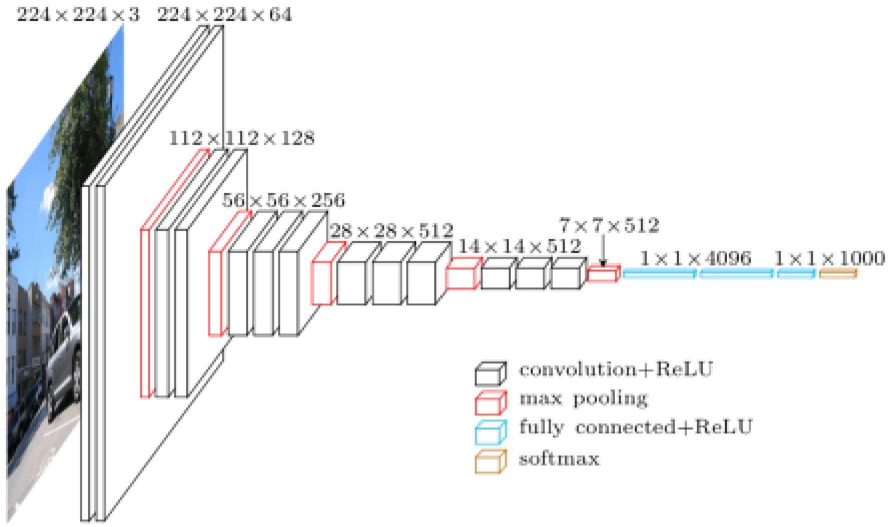
GoogLeNet



https://paperswithcode.com/method/googlenet

Visual Geometry Group (VGG)

- Very deep architecture of 16 or 19 layers in total
- Designed to have 2 or 3 conv layers, and a pooling layer, then again 2 or 3 conv layers followed by a pooling layers to reach 16 layers (in VGGNet16) and 19 layers (in VGGNet19)
- Used for multiple object detection.

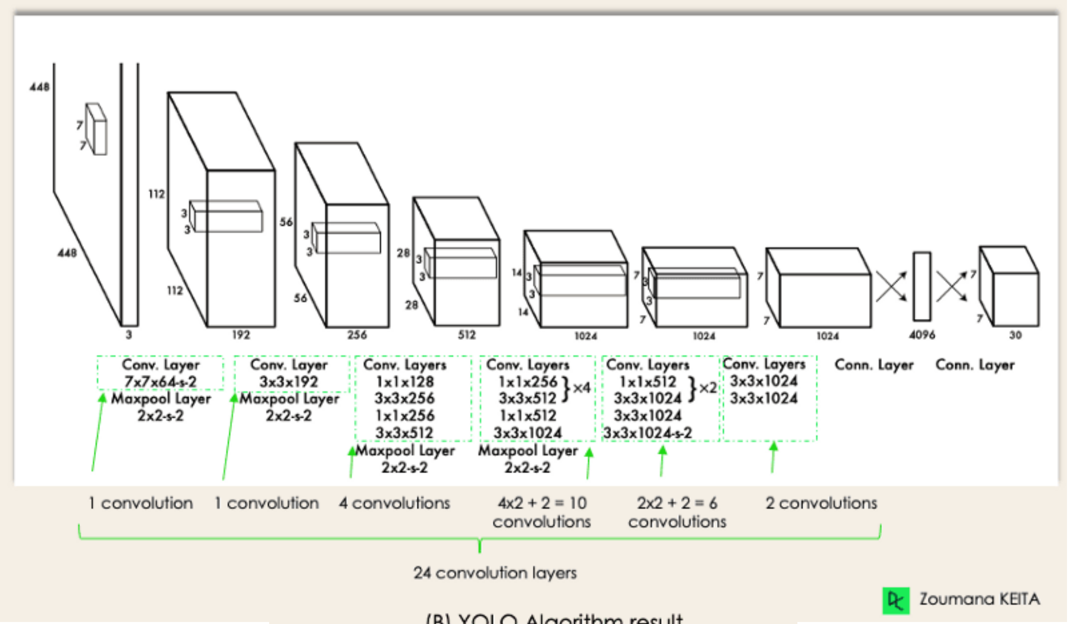


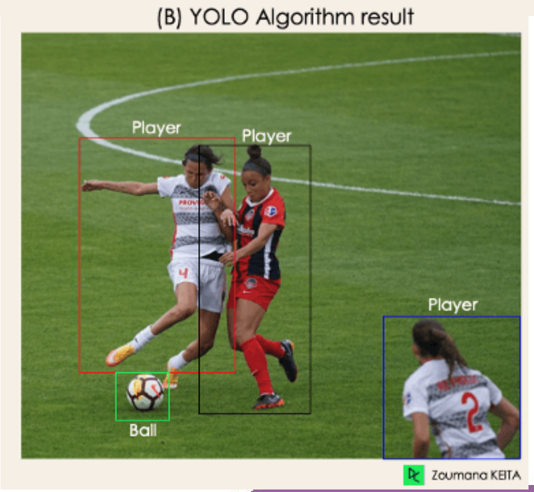
K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition". International Conference on Learning Representations (ICLR) 2015, doi: https://arxiv.org/abs/1409.1556

https://www.datacamp.com/blog/yolo-object-detection-explained

You Only Look Once (YOLO)

- One of the most popular architectures in recent times
- Authors framed object detection as a regression rather than a classification problem by spatially separating bounding boxes and associating probabilities to each of the detected images using a single CNN
- Architecture similar to GoogLeNet
  - 24 convs. 4 max pool, 2 FCN
- Advantages: Speed, accuracy, generalisation, open source





Redmond et al. , “You Only Look Once: Unified, Real-Time Object Detection”. ArXiV 2015, doi: https://arxiv.org/abs/1506.02640 https://paperswithcode.com/method/resnet

ResNet

- Training very deep networks proved to be problematic and can cause problems such as vanishing/exploding gradients
- However, ResNet made it possible to train a very deep network without harming the performance
  - Residual: Learning from a reference rather than the direct output

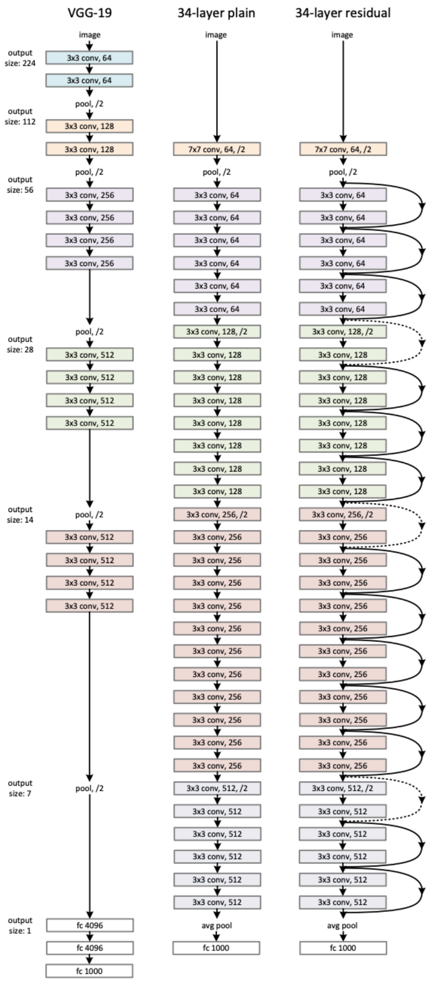
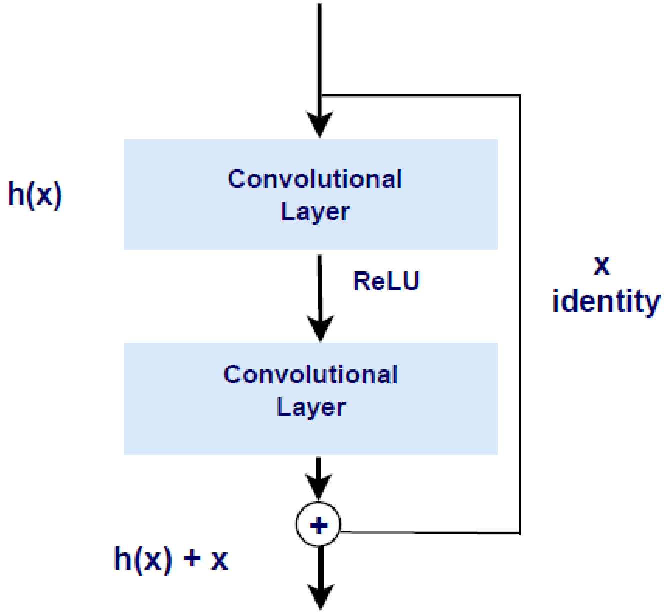


Figure 3. Example network architectures for ImageNet. **Left**: the VGG-19 model [41] (19.6 billion FLOPs) as a reference. **Middle**: a plain network with 34 parameter layers (3.6 billion FLOPs). **Right**: a residual network with 34 parameter layers (3.6 billion FLOPs). The dotted shortcuts increase dimensions. **Table 1** shows more details and other variants.

He et al. , "Deep Residual Learning for Image Recognition". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, doi: https://arxiv.org/pdf/1512.03385.pdf

ResNet

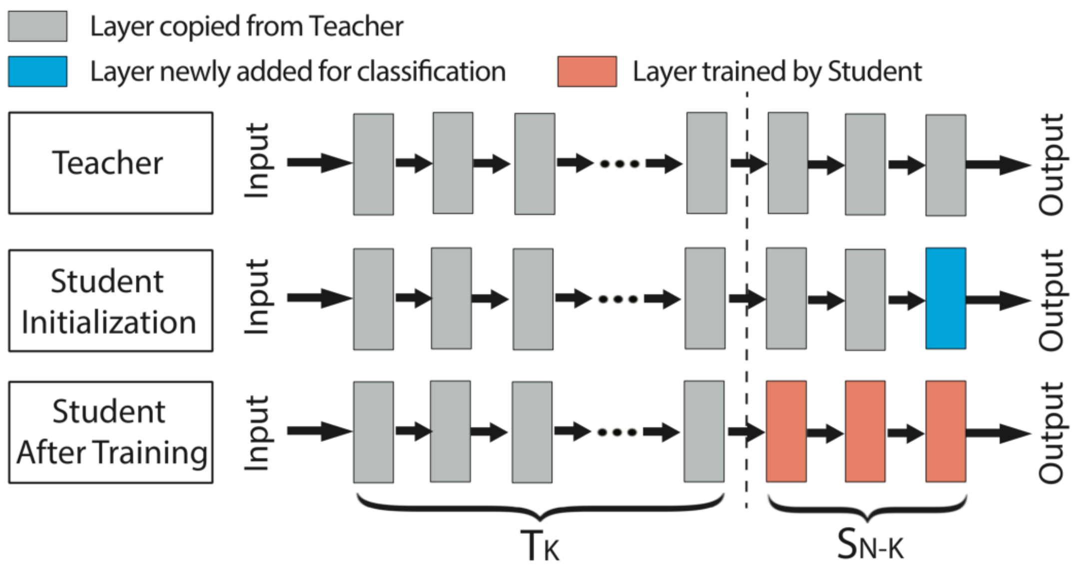
- Recall that the target of training is to model a function h(x), and so if you add input x to the output of the network (add a skip connection), then the network will end up modelling f(x) = h(x) - x instead of h(x)
  - This is called residual learning
- At the start of the training, the weights are initialized to be close to zero, so the network will simply output values close to zero.
- When adding the skip connection, the network will end up outputting a copy of its inputs.
  - This simply means if the target function is close to the identity function (often the case)
- This will speed up the training process and the network can start making progress even if some layers haven't started learning yet.



@Eyad Elyan

Pre-Trained Models → Transfer Learning

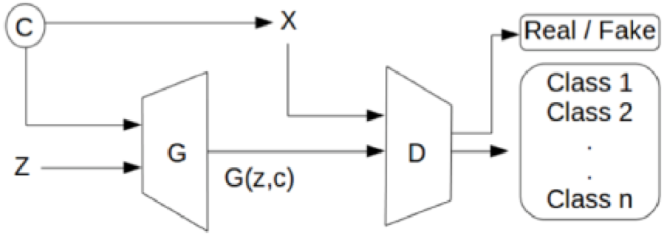
- (Almost) nobody implements these modules from scratch
  - Use pre-trained models with a single line of code!
- Most have been trained on large volumes of data (e.g. ImageNet which has 1 million images)
  - Therefore, they "know" how to recognise the most basic objects (e.g. people, vehicles, animals)
  - We "freeze" the lower layers and train the higher ones (fine-tuning).
  - Obviously, you also need to change the output layer (to predict your labels)
  - As a result, you need less training data to achieve better results

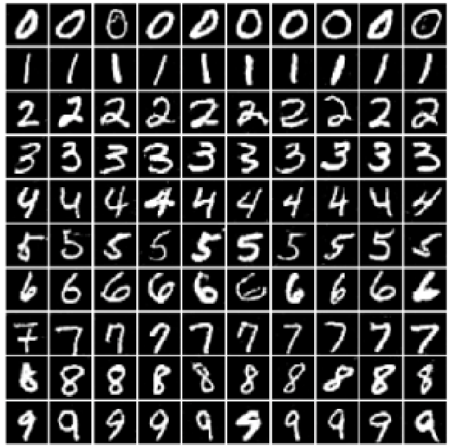


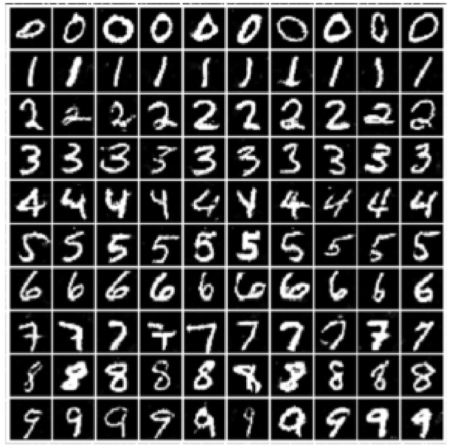
https://bdtechtalks.com/2019/06/10/what-is-transfer-learning/

Generative Models

- While the previous models classify/detect data better, attempts were made to make them generate data in parallel
- Goodfellow et al. realised that if you train two DNNs to compete against each other, not only they can classify better, but also, they can generate images better!
  - One model is called the generator, and the other one is the discriminator
  - The generator tries to create simples close to the original, and the discriminator tries to identify real from fake
  - Gradually, the generator will improve and beat the discriminator (and possibly you!)



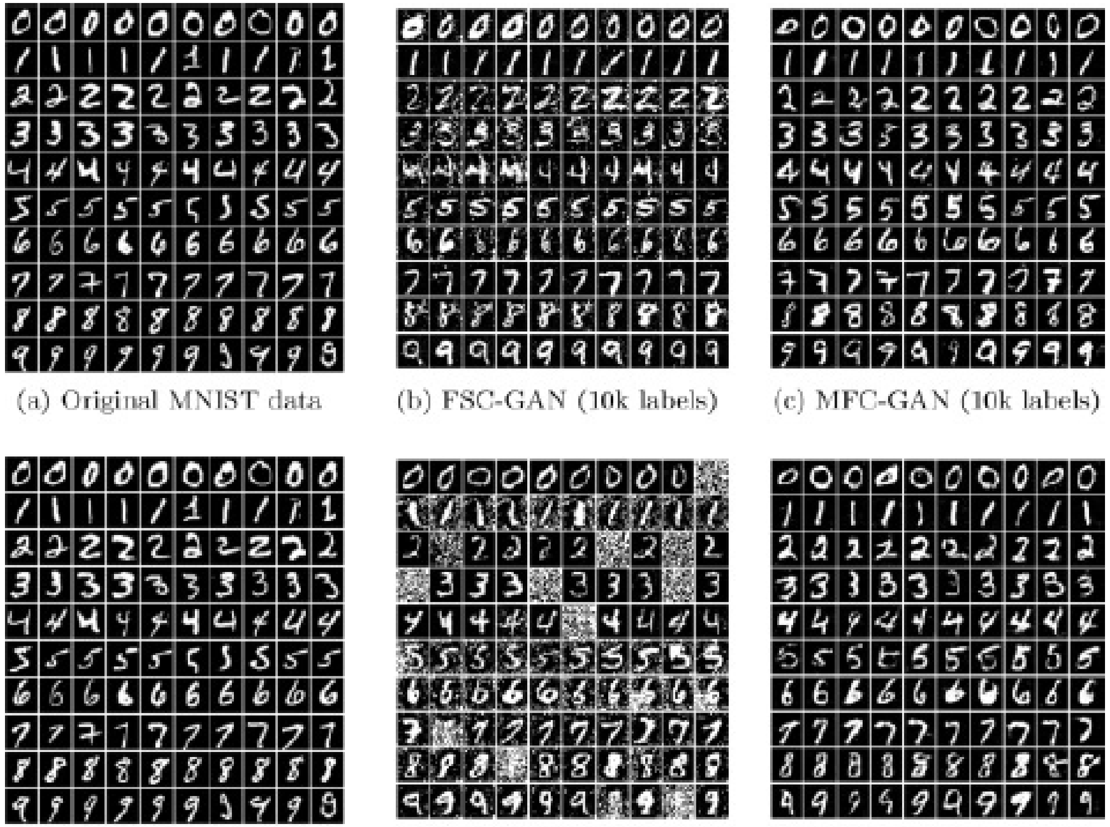




Goodfellow et al. , “Generative Adversarial Nets”. Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS) 2014, doi: https://doi.org/10.48550/arXiv.1406.2661

Generative Models

- At the moment, there's more than 1000 different GANs!
- https://github.com/hindupuravinash/the-gan-zoo

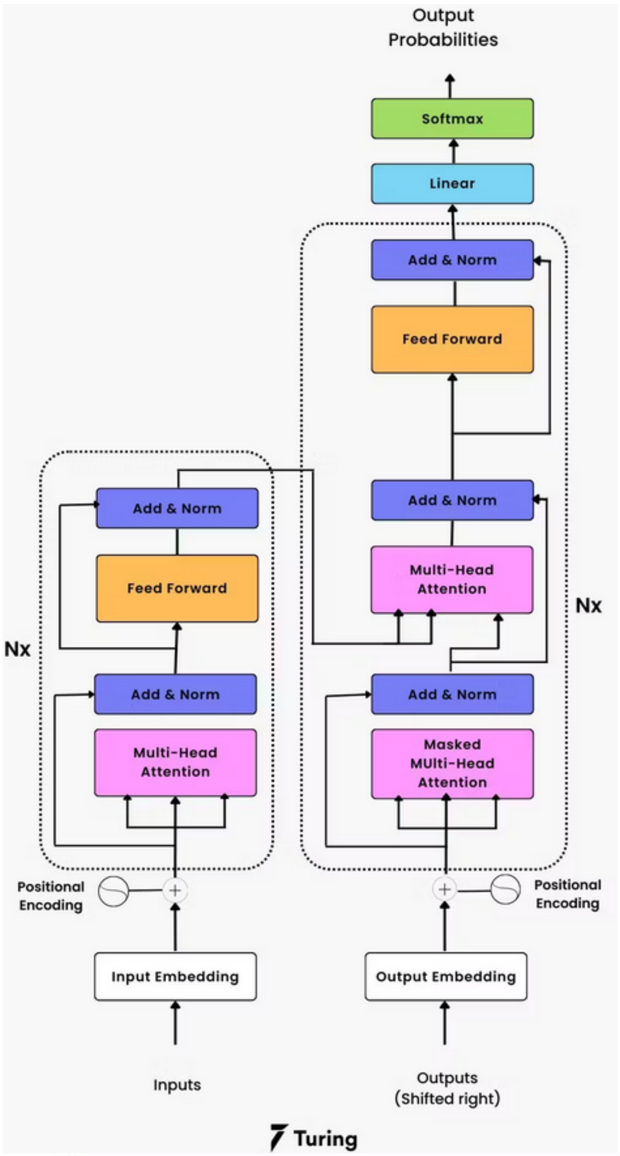


(a) Original MNIST data (b) FSC-GAN (10k labels) (c) MFC-GAN (10k labels)
(d) Original MNIST data (e) FSC-GAN (all labels) (f) MFC-GAN (all labels)

Transformers

- https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- https://www.turing.com/kb/brief-introduction-to-transformers-and-their-power

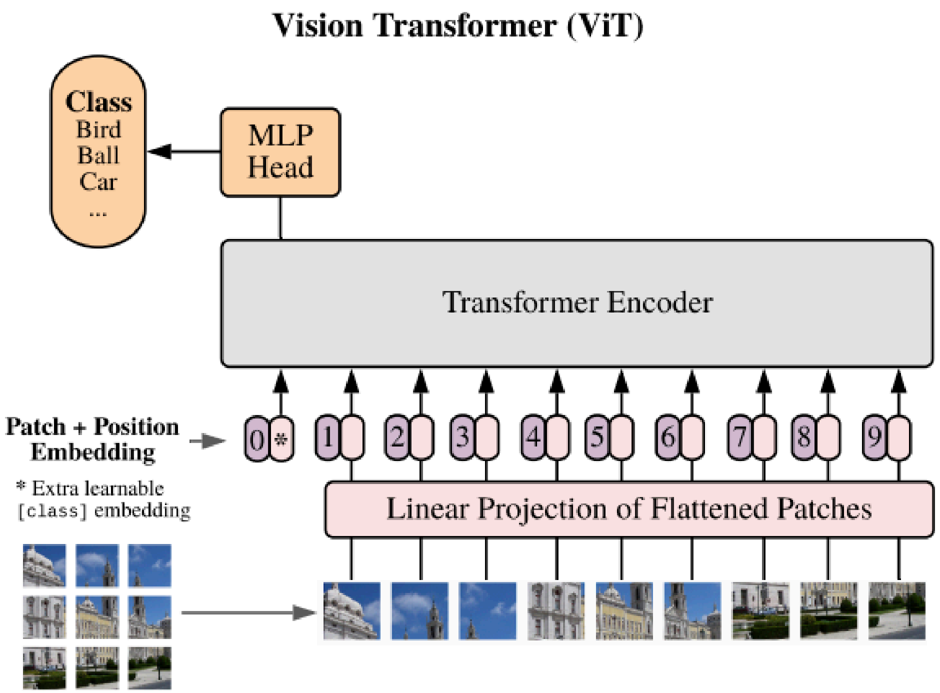
- There were previous attempts at understanding sequential data and adding “memory” to NNs
  - Examples include Recurrent Neural Networks (RNNs) and Long-Short Term Memory (LSTM)
- However, in 2017 , Vaswani et al. cracked a way to not only understand sequential data, but also to introduce “attention” mechanisms!
  - In fact, it is based on RNN but it is not sequential!
  - Encoder & Decoder using embedded text data



Vaswani et al. , “Attention is all you need”. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS) 2017, doi: https://doi.org/10.48550/arXiv.1706.03762

Vision Transformers (ViT)

- Dosovitskiy et al. “borrowed” this idea and implemented it for images!
- The method Works very well, but it has high computational demand
- Currently, they are the closest competitor vs CNN based architectures
  - Although you could “freeze” layers from a CNN and use their output to train a ViT instead of 16x16 patches!
- Most likely use: Action recognition
  - They can recognise “complex” actions



Dosovitskiy et al. , “An image is worth 16x16 words: Transformers for Image Recognition at Scale”. Proceedings of the 9th International Conference on Learning Representations (ICLR) 2021, doi: https://doi.org/10.48550/arXiv.2010.11929

Other interesting models to explore

- (Variational) Autoencoders
  - Feedforward networks to solve unsupervised learning
- U-Net
  - The best option for medical image segmentation
- Self Organising Maps (SOMs)
  - PCA for images!

https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-algorithm

Lab

Option 1: Transfer Learning

Option 2: Bayesian Classification using R