Architecture patterns

Deep learning - BMDC 2025-2026

Gheorghe Cosmin Silaghi

Universitatea Babeș-Bolyai

October 15, 2025

The model architecture

- architecture: the sum of choices that went into creating the DL model: which layers to use, how to configure and connect them etc.
- these choices define the hypothesis space of the model
- good hypothesis space encode prior knowledge about the problem

What is a good architecture?

- an architecture that reduces the size of the search space or
- makes it easier to converge to a good point in the search space

Modularity, hierarchy and reuse (MHR

how to make a complex system simpler?

- structure your amorphous soup of complexity into *modules*
- organize modules into a *hierarchy*
- start reusing the same modules in multiple places as appropriate

MHR

Its at the heart of the organization of any system of meaningful complexity (a cathedral, the human body, the US Navy etc).

DL - an expression of MHR

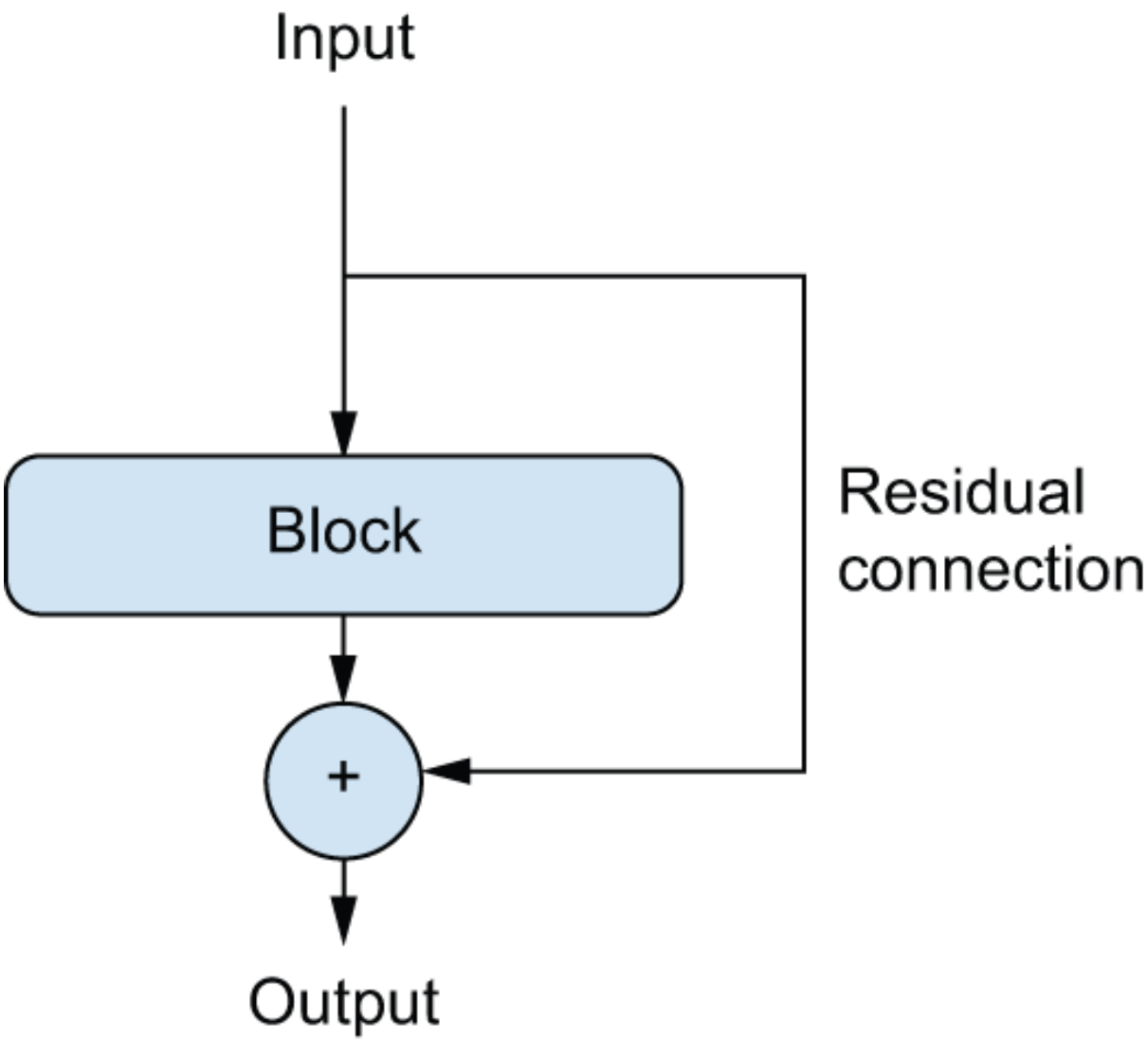
- take a classic optimization technique, structure the search space into modules (layers) organized into deep hierarchies, and reuse whatever you can
- popular ConvNet architectures are structured in repeated groups of layers (called *blocks*)
- ConvNet features a pyramid-like structure (feature hierarchies) - number of filters grows with layer depth while the size of feature maps shrinks
- deeper hierarchies are intrinsically good because they encourage feature reuse and therefore, abstraction
- a deep stack of narrow layers performs better than a shallow stack of large layers

Ablation studies

- DL architecture are more evolved than designed. If you take any complicated DL setup, chances are that you can remove few modules (or replace some trained features with random ones) with no loss in performance
- incentive for DL researchers: to make a system more complex than necesary in order to appear more interesting
  - if you read a lot of DL papers, you notice that they are optimized for peer review in both style and content in ways that actively hurt the clarity of explanations and reliability of results.
- the goal of a research is *to generate reliable knowledge*
- *understanding causality* is the most straightforward way to generate reliable knowledge
- **ablation studies**: systematically trying to remove parts of a system to identify where its performance actually comes from
  - if X + Y + Z givea good results, try also X, Y, X, X + Y, X + Z and Y +Z to see what happens

Residual connection

- *telefonul fara fir*: if a message is transmitted in oral form between multiple peers, there is a high chance that the last receiver to get an altered message
- cumulative errors that occur in a sequential transmission over a noisy channel could be high
- given a chain of functions $y = f4(f3(f2(f1(x))))$, backpropagation adjusts the parameters of each function based on the error recorded at the output of $f4$. to adjust $f1$, you need to percolate information through $f2$, $f3$ and $f4$.
- each successive function in the chain introduces some amount of noise. if the function is deep in the chain, the amount of noise overwhelmes the gradient information, and backpropagation stops working: **vanishing gradient problem**
- to fix this problem: force each function in the chain to be non-destructive: retain a noiseless version of its input - **residual connection**


Input
Block
+
Output
Residual
connection

Technical details of residual connection

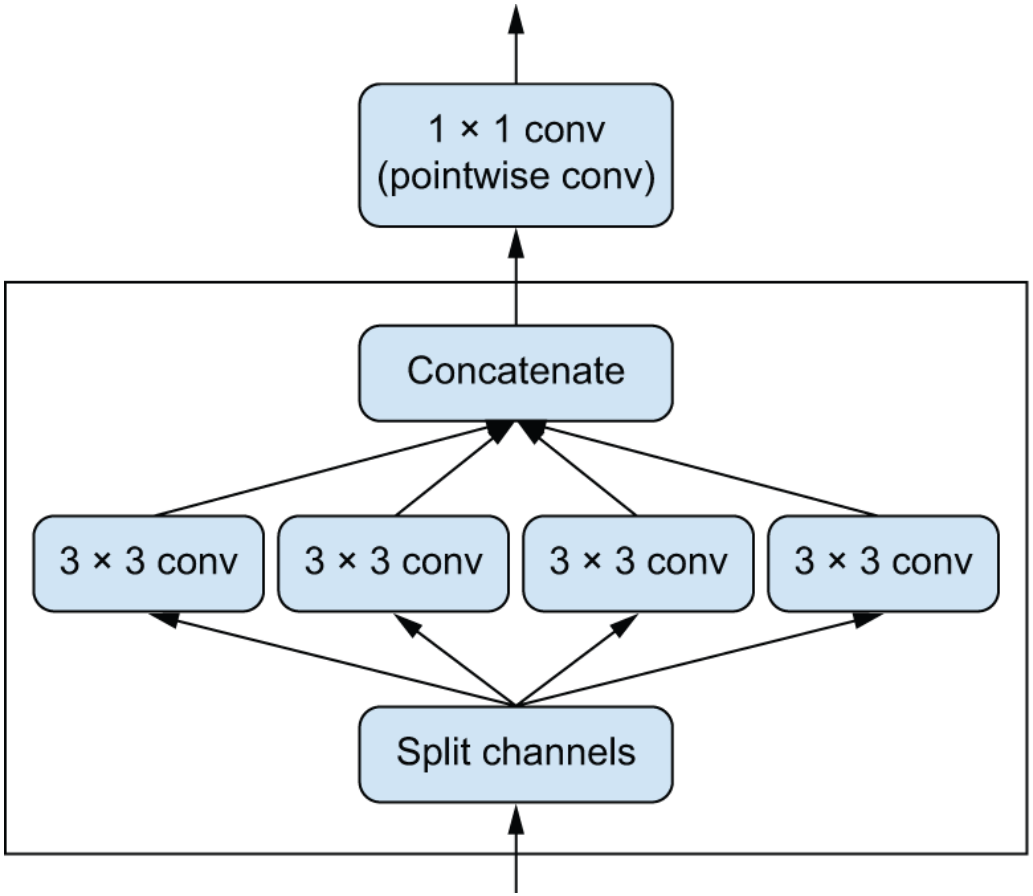
- the output of a block should be the same as its input in order to be able to apply the residual connection (this is not the case in the CNN layers)
- apply a 1x1 Conv2D layer with no activation on the input
- use padding="same" on the convolution layers to preserve the input size
- use strides in the residual projection to match the downsampling of a MaxPooling

Batch normalization

- normalization: make different samples seen by the ML model more similar to each other. Helps the model to learn and generalize well to new data
- the most common way of normalization: subtract the mean and give the data a unit standard deviation: consider the assumption that the data follows a normal distribution
- BatchNormalization layer: during training it uses the mean and the variance of the current batch of data to normalize samples. during inference uses exponential moving average of the batchwise mean and variance of data seen during training
- the main effect of batch normalization: helps gradient propagation - thus allows for deeper networks
- if a Conv2D is followed by a BatchNormalization, the bias in the kernel is not anymore needed.
- eventually, the Conv2D operation could be split in 2 parts: the affine transformation and the layer activation, and use BatchNormalization in between

Depthwise separable convolutions

- makes the model smaller (fewer weights) and leaner (fewer floating point operations)
- this layer performs a spatial convolution on each channel of its input independently, before mixing the output channels via a pointwise convolution
- separates the learning of spatial features and the learning of channel-wise features
- is based on the assumption that spatial locations in intermediate activations are highly correlated, but different channels are highly independent
- Depthwise separable convolution requires significantly fewer parameters and involves fewer computations compared with the regular convolution, converges faster are are less prone to overfitting



Depthwise convolution: independent spatial convs per channel

Architectural principles (up-to-now)

- the model should be organized in repeated blocks of layers, usually made of multiple convolution layers and a max pooling layer
- the number of filters in the layers should increase as the size of feature maps decreases
- deep and narrow is better than broad and shallow
- residual connections around blocks of layers helps to train deeper networks
- it can be beneficial to introduce batch normalization layers after your convolution layers
- it can be beneficial to replace Conv2D with SeparableConv2D layers, which are more parameter efficient

The Xception block

- successive blocks, each block being surounded by a residual connection
- each block is composed of two depthwise separable convolutions, followed by a max pooling operation
- data entering the separable convolution is normalized with BatchNormalization

Vision Transformers (ViT)

- Transformer: is a sequence-processing architecture, developed to process text
- Vision transformer: split the image into 1D sequence of patches, turns each patch into a flat vector, process the vector sequence
- the Transformer architecture allows ViT to capture long-range relationships between different parts of the image - which is a problem where Conv2D struggles
- Transformers are a great choice when working with massive datasets.
- a ViT is really large.
- they end up being slow for anything smaller than ImageNet