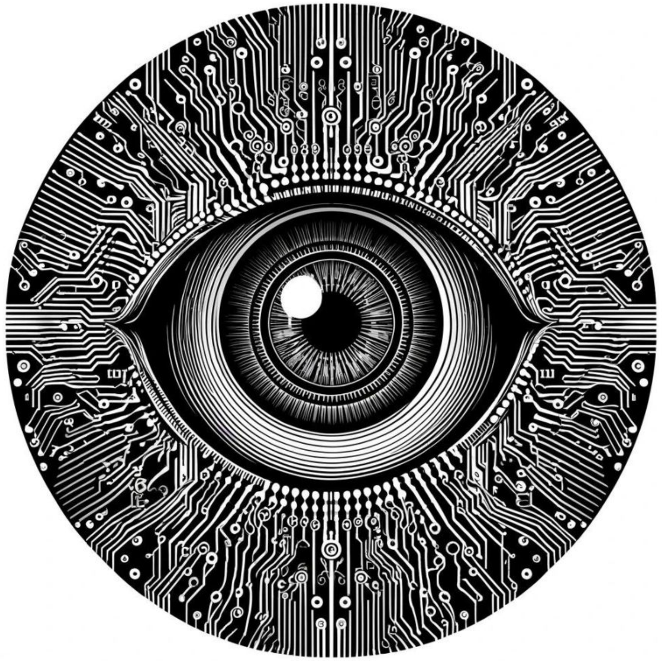
KI Service Zentrum by Hasso-Plattner-Institut HPI

Skip Connections



Antonio Rueda-Toicen

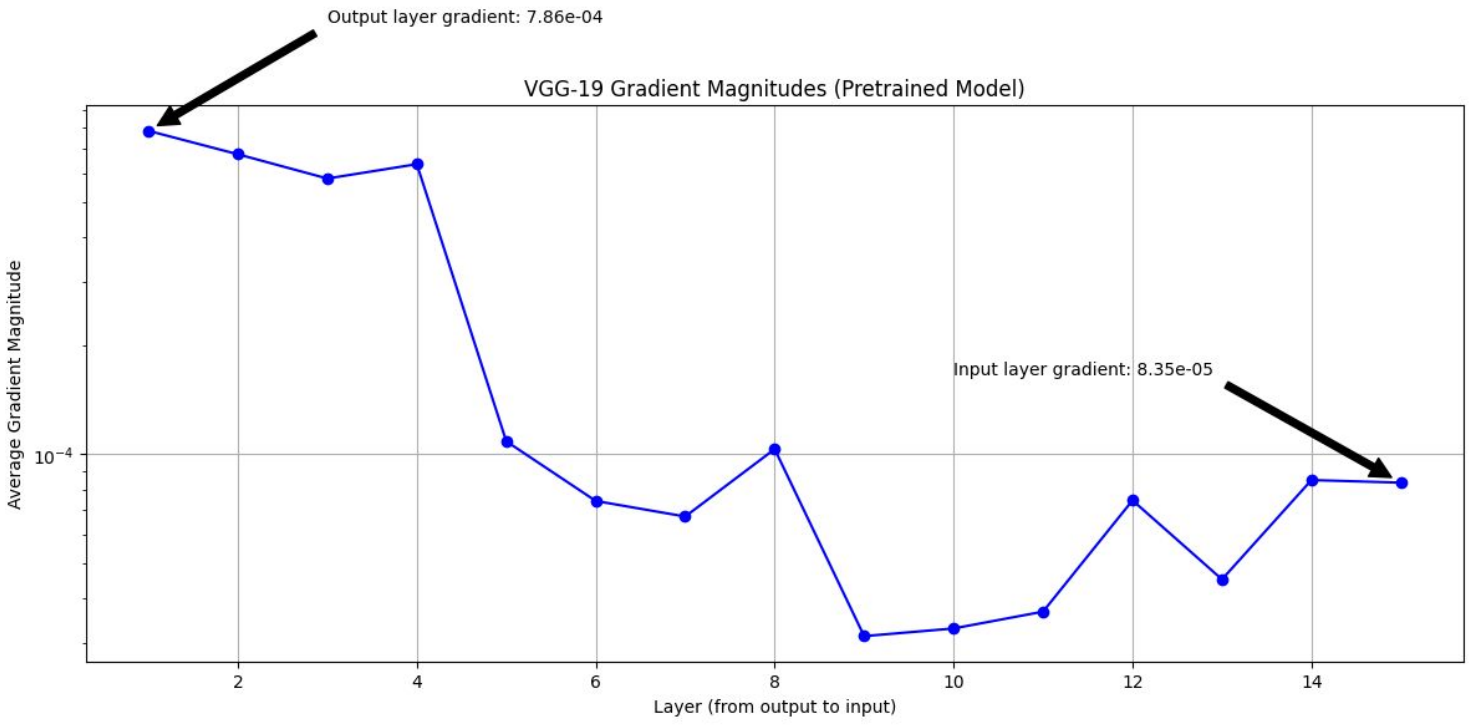
SPONSORED BY THE
Federal Ministry of Education and Research

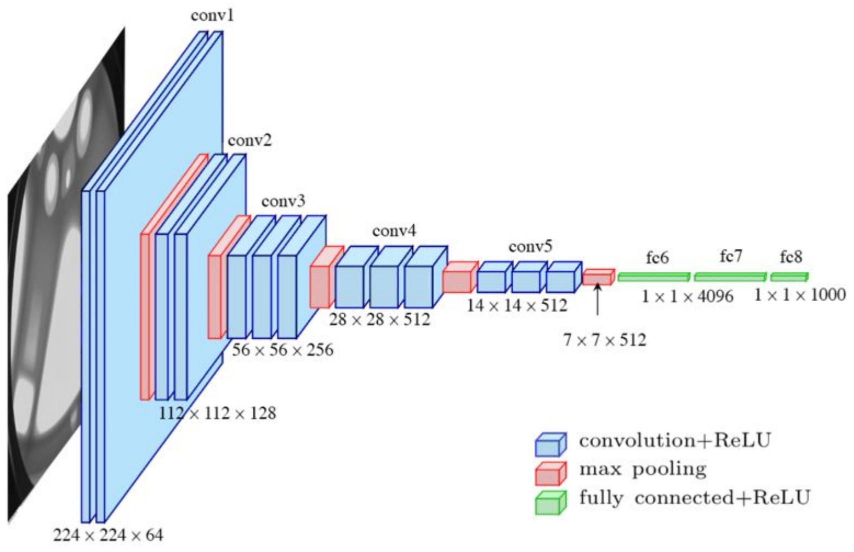
Learning goals

- Understand the vanishing gradient as a numerical problem
- Implement skip connections as element-wise addition or concatenation of activation maps

The vanishing gradient

$$w_{ij} = w_{ij} - \left(\text{learning rate} * \frac{dL}{dw_{ij}}\right)$$

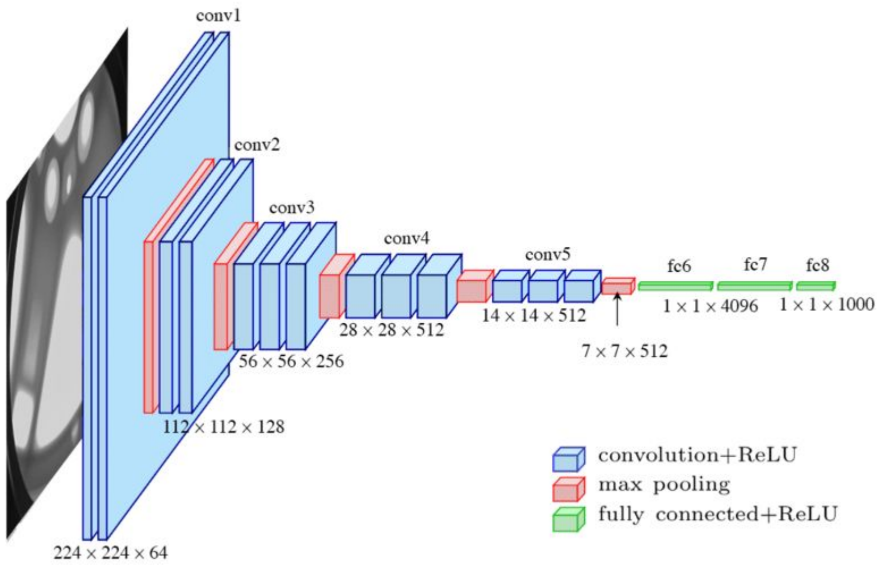




VGG-19 network (source)

Numerical underflow in neural networks

```python
import numpy as np
from scipy.signal import convolve2d

# Example data (any image or 2D array)
image = np.ones((8,8), dtype=np.float32)

# A 3×3 kernel with sum=0.8
kernel = np.array([[0.05, 0.10, 0.05],
                   [0.10, 0.20, 0.10],
                   [0.05, 0.10, 0.05]], dtype=np.float32)

for i in range(1000):
   image = convolve2d(image, kernel, mode='same', boundary='fill', fillvalue=0)
   # Underflow can show up when values drop below np.finfo(np.float32).tiny
   if (image > 0).sum() == 0:
       print("All values underflowed to 0 at iteration", i)
       break
```



VGG-19 network (source)

Numerical underflow

```python
import numpy as np

a = 1e-8 # Equal to 1 x 10 ** -8
b = 2

print(np.float32(a) ** b)  # Gives a value close to 1e-16
print(np.float16(a) ** b)  # Underflows to 0.0
```

$$(0.00000001)^2 = (1 \times 10^{-8})^2 = 10^{-16}$$

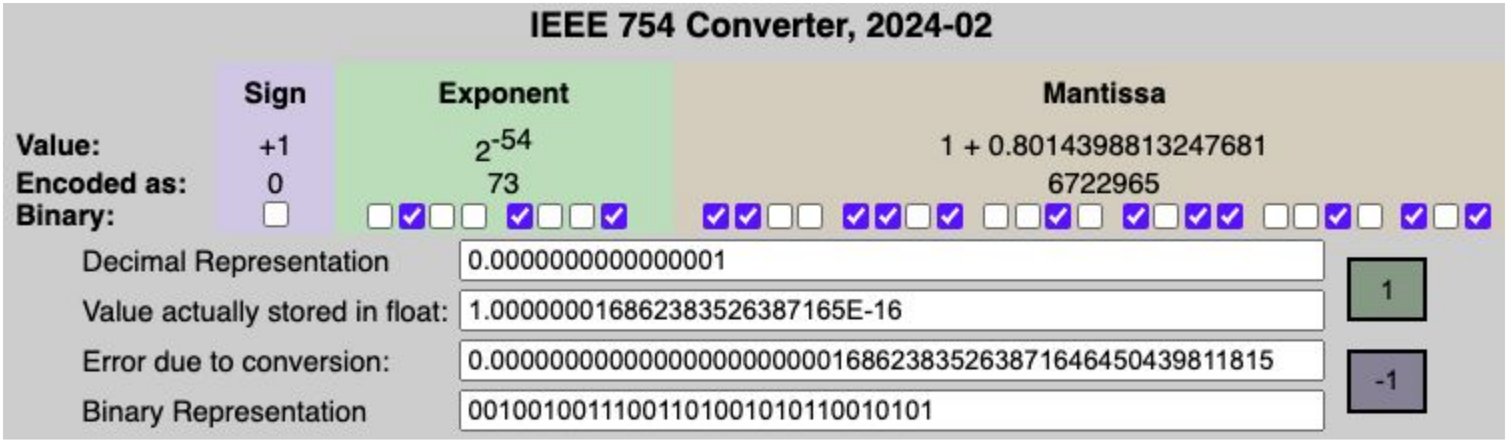
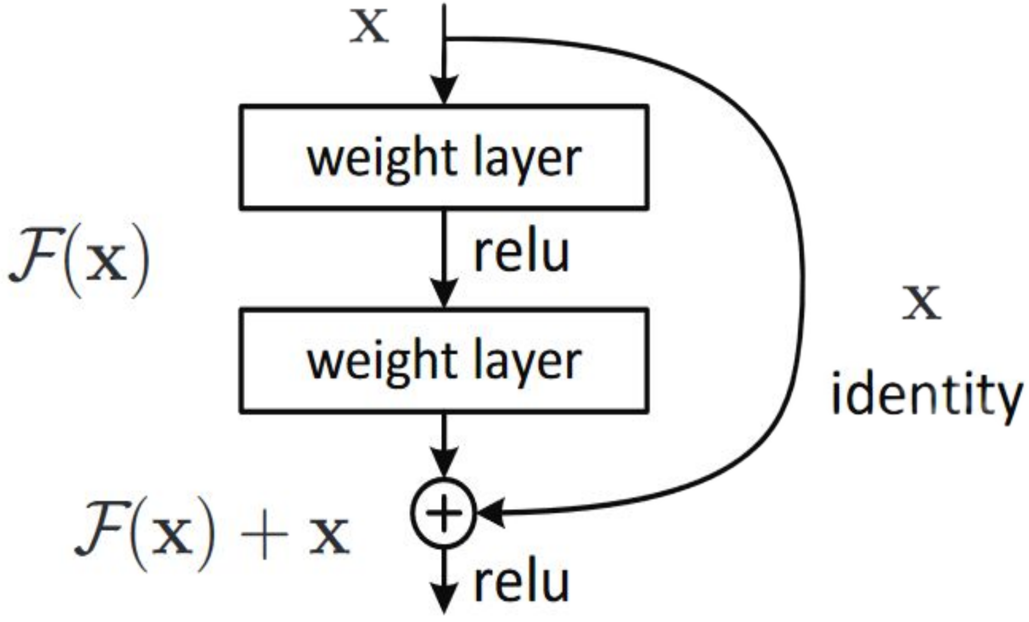


Image from [IEEE-754 Floating Point Converter]

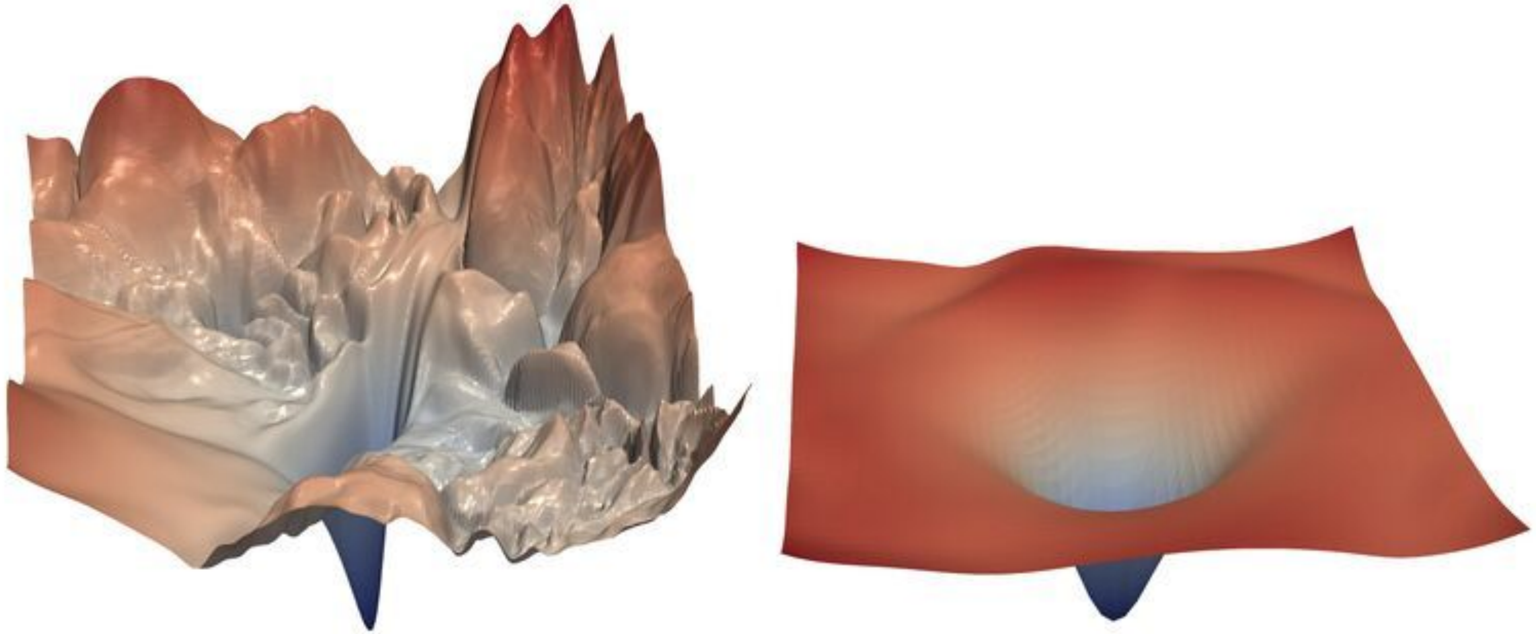
Skip connections on Resnet

```python
import torch.nn as nn

class ResidualBlock(nn.Module):
    def __init__(self, channels):
        super().__init__()
        # Main path - "city route"
        self.conv1 = nn.Conv2d(channels, channels, kernel_size=3, padding=1)
        self.bn1 = nn.BatchNorm2d(channels)
        self.conv2 = nn.Conv2d(channels, channels, kernel_size=3, padding=1)
        self.bn2 = nn.BatchNorm2d(channels)
        self.relu = nn.ReLU()

    def forward(self, x):
        # Save input for skip connection - "highway route / checkpoint"
        identity = x

        # Main path through convolutions
        out = self.conv1(x)
        out = self.bn1(out)
        out = self.relu(out)
        out = self.conv2(out)
        out = self.bn2(out)

        # Add skip connection - "merging highway with city route (adding checkpoint)"
        out += identity

        # Final activation
        out = self.relu(out)

        return out
```



$$\underbrace{\begin{bmatrix}-0.12 & 0.24\\ 0.35 & -0.25\end{bmatrix}}_{\text{Identity Path (x)}} + \underbrace{\begin{bmatrix}0.25 & -0.14\\ -0.45 & 0.35\end{bmatrix}}_{\text{Main Path (F(x))}} = \underbrace{\begin{bmatrix}0.13 & 0.10\\ -0.10 & 0.10\end{bmatrix}}_{\text{Output (F(x) + x)}}$$

Effects on the loss landscape



(a) without skip connections (b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Image from [Visualizing the Loss Landscape of Neural Nets]

Relevance on current architectures

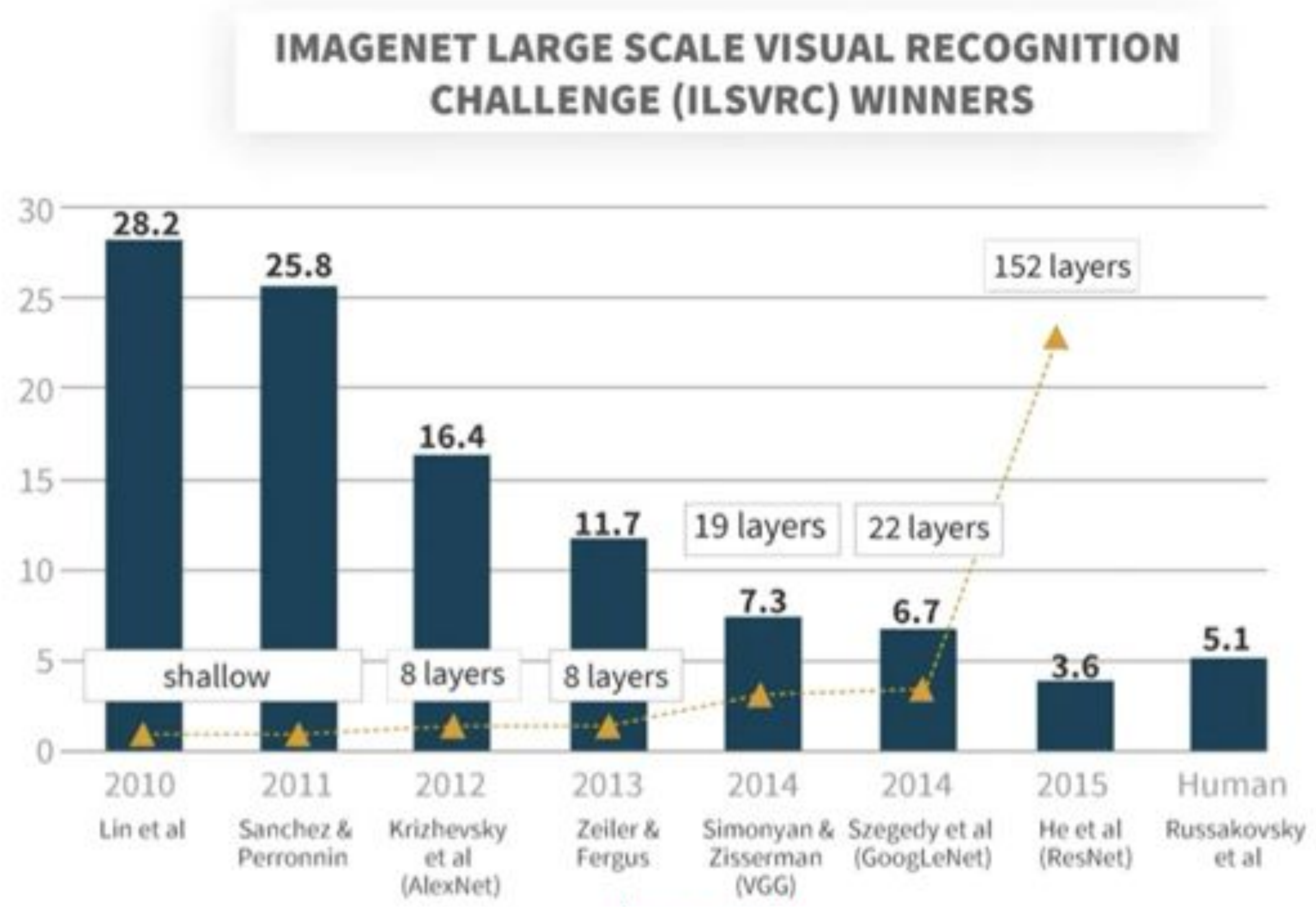
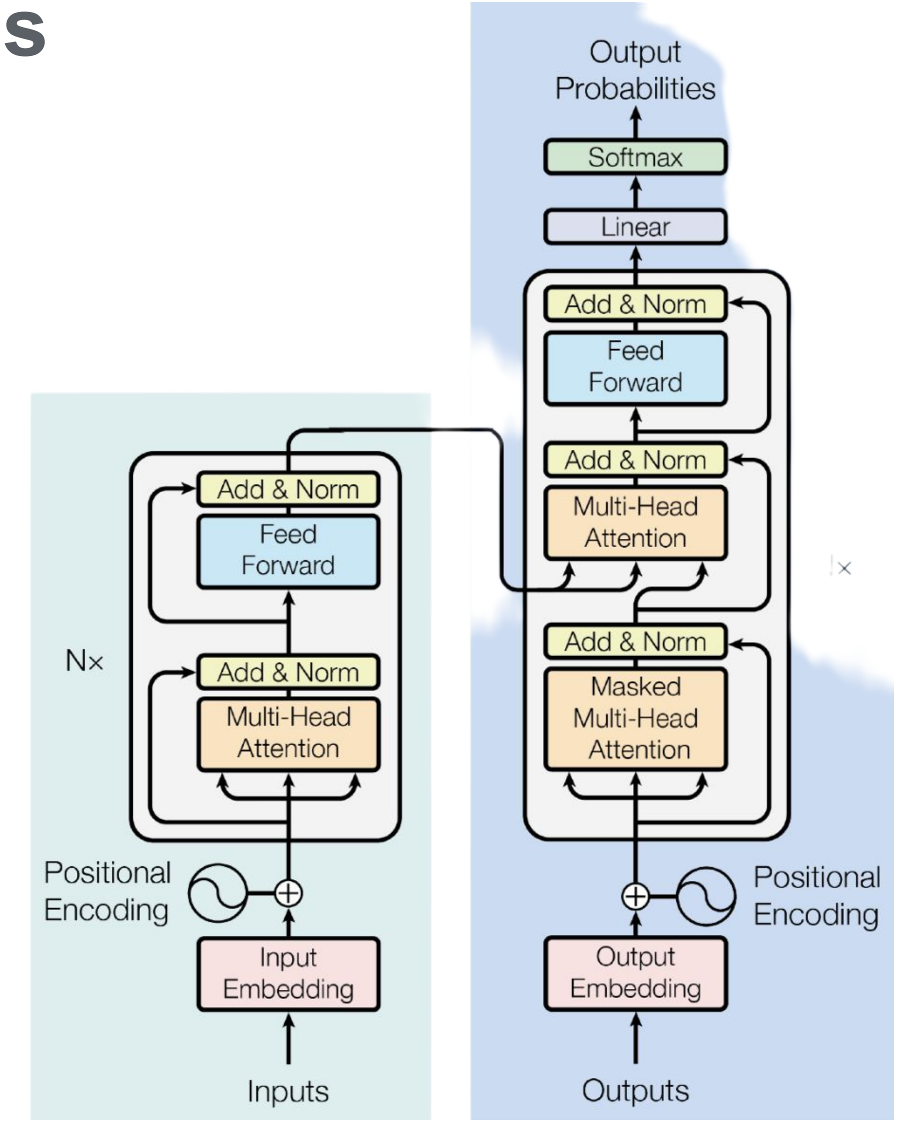


Image source



Transformer architecture from source

Skip connections on Densenet

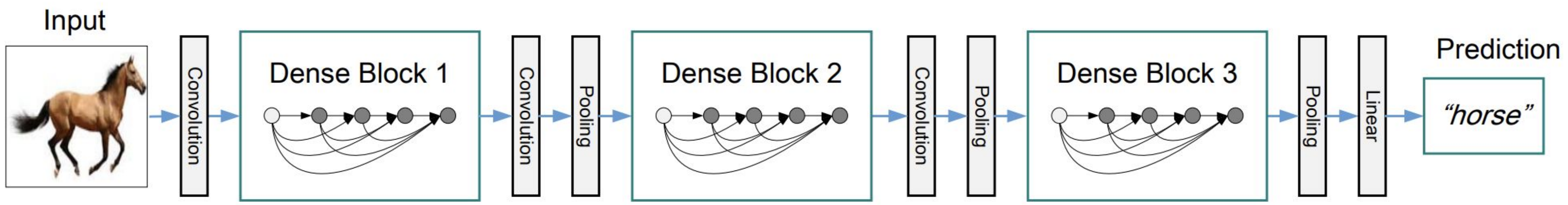
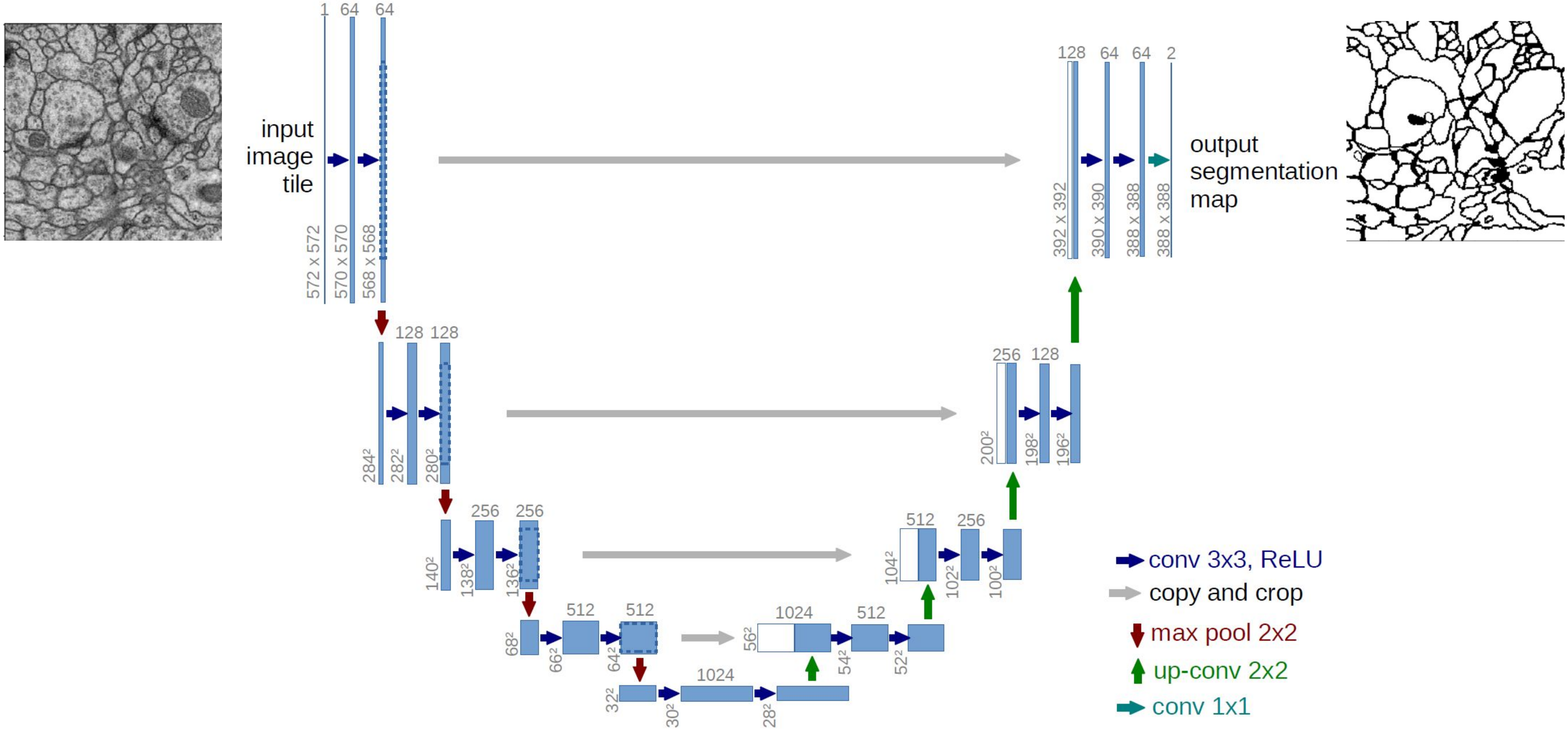


Figure 2: A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling.

```
# Feature maps are concatenated instead of added
# We can control the number of feature maps by using 1x1 convolutions
torch.cat(features, dim=1)
```

Skip connections on U-net



Summary

The vanishing gradient is a numerical problem

- Computers have limited precision to represent small numbers

Skip connections serve as “checkpoints” for what the model has learned

- A skip connection gives us the chance to preserve information that could have been destroyed due to numerical underflow
- Skip connections are what allow neural networks to be deep and increase their number of parameters while avoiding vanishing gradients

Two types of skip connections: addition and concatenation

- We use either element wise addition or concatenation of feature maps as skip connections

SPONSORED BY THE
Federal Ministry
of Education
and Research

Further reading and references

Deep Residual Learning for Image Recognition

- **https://arxiv.org/abs/1512.03385**

Densely Connected Convolutional Networks

- https://arxiv.org/abs/1608.06993

Visualizing the Loss Landscape of Neural Nets

- https://arxiv.org/abs/1712.09913

SPONSORED BY THE
Federal Ministry
of Education
and Research