Symbolic Complexity of the Colorization Model

(Akshith / 2022102048)

Assumptions and Notation

- Input is grayscale of spatial size $H \times W$, with channels NIC (typically NIC = 1).
- Convolution / transpose-convolution kernels are $k \times k$ with $k^2 = k \cdot k$, padding $p = k//2$.
- Strides:
  - Encoder: three stride-2 convs: $32 \to 16 \to 8 \to 4$ (or $64 \to 32 \to 16 \to 8$ for doubled input).
  - Decoder: three stride-2 transpose-convs: $4 \to 8 \to 16 \to 32$ (or $8 \to 16 \to 32 \to 64$).
- Channels per layer:

  L1: NIC $\to$ NF, L2: NF $\to$ 2NF, L3: 2NF $\to$ 4NF,
  D1: 4NF $\to$ 2NF, D2: 2NF $\to$ NF, D3: NF $\to$ NF, CLS (1×1): NF $\to$ NC.
- Ignore biases and BatchNorm parameters.

Per-layer formulae. For any (transpose-)convolution with $C_{\text{in}} \to C_{\text{out}}$, kernel $k \times k$, output map $H_\ell \times W_\ell$:

$$\begin{aligned} \textbf{Weights:} &\quad C_{\text{out}}\, C_{\text{in}}\, k^2, \\ \textbf{Outputs (activation elements):} &\quad C_{\text{out}}\, H_\ell\, W_\ell, \\ \textbf{Connections:} &\quad (C_{\text{out}}\, C_{\text{in}}\, k^2) \cdot (H_\ell\, W_\ell). \end{aligned}$$

For the 1×1 classifier: weights = NC · NF; outputs = NC · $H_{\text{out}} W_{\text{out}}$; connections = (NC · NF) · ($H_{\text{out}} W_{\text{out}}$).

Case A: Input 32×32

Output spatial sizes per layer:

L1: 16×16 (= 256), L2: 8×8 (= 64), L3: 4×4 (= 16),
D1: 8×8 (= 64), D2: 16×16 (= 256), D3: 32×32 (= 1024),
CLS: 32×32 (= 1024).

1) Total Number of Weights

Layer-wise weights:

L1: NF · NIC · k^2, L2: 2NF$^2k^2$, L3: 8NF$^2k^2$,
D1: 8NF$^2k^2$, D2: 2NF$^2k^2$, D3: NF$^2k^2$, CLS: NF · NC.

Sum:

$$\boxed{W_{\text{total},\,32} = k^2\big(\text{NF} \cdot \text{NIC} + 21\,\text{NF}^2\big) + \text{NF} \cdot \text{NC}.}$$

2) Total Number of Outputs (Activation Elements)

Layer-wise outputs:

$$\begin{aligned}&\text{L1: } \mathrm{NF}\cdot 256, \quad \text{L2: } 2\mathrm{NF}\cdot 64 = \mathrm{NF}\cdot 128, \quad \text{L3: } 4\mathrm{NF}\cdot 16 = \mathrm{NF}\cdot 64,\\ &\text{D1: } 2\mathrm{NF}\cdot 64 = \mathrm{NF}\cdot 128, \quad \text{D2: } \mathrm{NF}\cdot 256, \quad \text{D3: } \mathrm{NF}\cdot 1024,\\ &\text{CLS: } \mathrm{NC}\cdot 1024.\end{aligned}$$

Sum:

$$\boxed{O_{\text{total},\,32} = \mathrm{NF}\cdot 1856 + \mathrm{NC}\cdot 1024.}$$

3) Total Number of Connections

Layer-wise connections (weights × output positions):

$$\begin{aligned}&\text{L1: } (\mathrm{NF}\cdot\mathrm{NIC}\cdot k^2)\cdot 256,\\ &\text{L2: } (2\mathrm{NF}^2k^2)\cdot 64 = 128\,\mathrm{NF}^2k^2,\\ &\text{L3: } (8\mathrm{NF}^2k^2)\cdot 16 = 128\,\mathrm{NF}^2k^2,\\ &\text{D1: } (8\mathrm{NF}^2k^2)\cdot 64 = 512\,\mathrm{NF}^2k^2,\\ &\text{D2: } (2\mathrm{NF}^2k^2)\cdot 256 = 512\,\mathrm{NF}^2k^2,\\ &\text{D3: } (\mathrm{NF}^2k^2)\cdot 1024 = 1024\,\mathrm{NF}^2k^2,\\ &\text{CLS: } (\mathrm{NF}\cdot\mathrm{NC})\cdot 1024.\end{aligned}$$

Sum:

$$\boxed{C_{\text{total},\,32} = k^2\big(256\,\mathrm{NF}\cdot\mathrm{NIC} + 2304\,\mathrm{NF}^2\big) + 1024\,\mathrm{NF}\cdot\mathrm{NC}.}$$

Case B: Input 64×64 (spatial doubled)

With the same stride pattern (three $\times 2$ downs/ups), every feature-map area $H_\ell W_\ell$ is $4\times$ larger than in the 32×32 case; *weights do not depend on spatial size.*

Therefore:

$$\boxed{W_{\text{total},\,64} = W_{\text{total},\,32} = k^2\big(\mathrm{NF}\cdot\mathrm{NIC} + 21\,\mathrm{NF}^2\big) + \mathrm{NF}\cdot\mathrm{NC}.}$$

$$\boxed{O_{\text{total},\,64} = 4\,O_{\text{total},\,32} = \mathrm{NF}\cdot 7424 + \mathrm{NC}\cdot 4096.}$$

$$\boxed{C_{\text{total},\,64} = 4\,C_{\text{total},\,32} = k^2\big(1024\,\mathrm{NF}\cdot\mathrm{NIC} + 9216\,\mathrm{NF}^2\big) + 4096\,\mathrm{NF}\cdot\mathrm{NC}.}$$

Notes

- If the actual channel schedule or number of stages differs, re-apply the per-layer formulas above with your $C_{\text{in}}, C_{\text{out}}, H_\ell, W_\ell$ and sum.
- Transpose-conv layers use the same weight formula $C_{\text{out}}C_{\text{in}}k^2$; their output sizes follow the stride-2 upsampling with padding $p = k//2$.