AI Applications Lecture 15
Image Generation AI 5: Convolutional Neural Networks for Image Generation

SUZUKI, Atsushi
Jing WANG

Contents

1 Introduction

1.1 Roadmap Recap

We will review the content learned so far. Three lectures ago, we learned about the **Variational Autoencoder (VAE)** as a natural image decoder [3]. Two lectures ago, we learned about the **reverse diffusion process** that generates low-resolution latent images, namely the **denoising scheduler** [1]. In the previous lecture, we mathematically understood the sense in which the reverse diffusion process performs **distribution learning**, using the continuity equation, score, and KL divergence (see [2, 6]).

In this lecture, we will focus our discussion on **practical image generation AI** and look in detail at the **neural network architectures** used in denoising schedulers and VAEs. In particular, we will focus on the differences, as the **architecture in the original paper** and the **architecture used in actual implementations** often differ (e.g., Latent Diffusion/Stable Diffusion [4]).

1.2 Learning Outcomes

By the end of this lecture, students should be able to:

- **Mathematically describe** the **neural network architectures used in practical image generation AI**.
- **Explain** how practical image generation AI achieves support for **variable input/output sizes** by using specific **layers**.
- **Explain** how the neural network architectures used in practical image generation AI have **changed from their original proposals**.

2 Preparation: Mathematical Notations

- **Definition:**
  - $(\mathrm{LHS}) \coloneqq (\mathrm{RHS})$: Indicates that the left-hand side is defined by the right-hand side. For example, $a \coloneqq b$ indicates that a is defined as b.
- **Set:**
  - Sets are often denoted by uppercase calligraphic letters. E.g., $\mathcal{A}$.
  - $x \in \mathcal{A}$: Indicates that element x belongs to set $\mathcal{A}$.
  - $\{\}$: The empty set.
  - $\{a, b, c\}$: The set consisting of elements a, b, c (set-builder notation by extension).

- $\{x \in \mathcal{A} \mid P(x)\}$: The set of elements in set $\mathcal{A}$ for which the proposition $P(x)$ is true (set-builder notation by intension).
- $|\mathcal{A}|$: The number of elements in set $\mathcal{A}$ (in this lecture, used only for finite sets).
- $\mathbb{R}$: The set of all real numbers. Similarly for $\mathbb{R}_{>0}$, $\mathbb{R}_{\geq 0}$, etc.
- $\mathbb{Z}$: The set of all integers. Similarly for $\mathbb{Z}_{>0}$, $\mathbb{Z}_{\geq 0}$, etc.
- $[1, k]_{\mathbb{Z}}$: For $k \in \mathbb{Z}_{>0} \cup \{+\infty\}$, if $k < +\infty$, then $\{1, \ldots, k\}$; if $k = +\infty$, then $\mathbb{Z}_{>0}$.

- **Function:**
  - $f : \mathcal{X} \to \mathcal{Y}$ denotes a mapping.
  - $y = f(x)$ denotes the output $y \in \mathcal{Y}$ for the input $x \in \mathcal{X}$.
- **Vector:**
  - Vectors are denoted by bold italic lowercase letters. E.g., $\boldsymbol{v}$. $\boldsymbol{v} \in \mathbb{R}^n$.
  - The i-th component is written as v_i:

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}. \tag{1}$$

  - Standard inner product:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle \coloneqq \sum_{i=1}^{d_{\mathrm{emb}}} u_i v_i. \tag{2}$$

- **Sequence:**
  - We call $\boldsymbol{a} : [1, n]_{\mathbb{Z}} \to \mathcal{A}$ a sequence of length n. If $n < +\infty$, $\boldsymbol{a} = (a_1, \ldots, a_n)$; if $n = +\infty$, $\boldsymbol{a} = (a_1, a_2, \ldots)$.
  - The length is written as $|\boldsymbol{a}|$.
- **Matrix:**
  - Matrices are denoted by bold italic uppercase letters. E.g., $\boldsymbol{A} \in \mathbb{R}^{m,n}$.
  - The elements are $a_{i,j}$, and

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}. \tag{3}$$

- Transpose $\boldsymbol{A}^\top \in \mathbb{R}^{n,m}$:

$$\boldsymbol{A}^\top = \begin{bmatrix} a_{1,1} & \cdots & a_{m,1} \\ \vdots & \ddots & \vdots \\ a_{1,n} & \cdots & a_{m,n} \end{bmatrix}. \tag{4}$$

- The transpose of a vector is

$$\boldsymbol{v}^\top = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}. \tag{5}$$

- **Tensor:**
  - A tensor as a multi-dimensional array is denoted by an underlined bold italic uppercase letter $\underline{\boldsymbol{A}}$.

3 General Theory: Reconfirming Architectural Freedom

3.1 Training and Inference of Noise Estimators are Architecture-Independent

The training of the **noise estimator** used in the denoising scheduler was given by the optimization problem of minimizing the squared error of the noise estimation. The objective function is identical to the previous lecture:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} \left\| \boldsymbol{\epsilon}^{(i)} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}\left(\boldsymbol{\zeta}^{(i)}, \boldsymbol{c}^{(i)}, t^{(i)}\right) \right\|_2^2. \tag{6}$$

(6) is defined **independently of the neural network's architecture**. Inference also just involves inserting the trained $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}$ into a sequential algorithm, which is also **architecture-independent** (e.g., DDPM/DDIM steps [1]).

3.2 VAE Training and Inference are Likewise Architecture-Independent

VAE training is regularized reconstruction error minimization [3]. For each input image $\underline{\boldsymbol{X}} \in \mathcal{I}$, we define an encoder that outputs a mean vector

$$\mathsf{MeanEnc}_{\boldsymbol{\eta}_{\mathrm{mean}}} : \mathcal{I} \to \mathbb{R}^d, \qquad \boldsymbol{\mu}(\underline{\boldsymbol{X}}) \coloneqq \mathsf{MeanEnc}_{\boldsymbol{\eta}_{\mathrm{mean}}}(\underline{\boldsymbol{X}}), \tag{7}$$

and an encoder that outputs the standard deviation for each component

$$\mathsf{SDEnc}_{\boldsymbol{\eta}_{\mathrm{SD}}} : \mathcal{I} \to \mathbb{R}^d_{>0}, \qquad \boldsymbol{\sigma}(\underline{\boldsymbol{X}}) \coloneqq \mathsf{SDEnc}_{\boldsymbol{\eta}_{\mathrm{SD}}}(\underline{\boldsymbol{X}}) \tag{8}$$

(σ is a positive real number for each element). The entire set of parameters for the VAE encoder is then

$$\boldsymbol{\eta} \coloneqq (\boldsymbol{\eta}_{\mathrm{mean}}, \boldsymbol{\eta}_{\mathrm{SD}}) \tag{9}$$

. We **sample** a standard normal random vector $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and generate the latent vector by

$$z_{\boldsymbol{\epsilon}} \coloneqq \boldsymbol{\mu}(\underline{\boldsymbol{X}}) + \boldsymbol{\sigma}(\underline{\boldsymbol{X}}) \odot \boldsymbol{\epsilon} \tag{10}$$

($\odot$ is the element-wise product). The decoder provides

$$\hat{\boldsymbol{X}}_{\boldsymbol{\epsilon}} \; \coloneqq \; \mathrm{Dec}_{\gamma}\left(z_{\boldsymbol{\epsilon}}\right) \tag{11}$$

. Using a reconstruction loss function $\ell : \mathcal{I} \times \mathcal{I} \to \mathbb{R}_{\geq 0}$, the objective function is

$$\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\gamma}) \coloneqq \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}\Big[\quad \underbrace{\ell\big(\underline{\boldsymbol{X}}, \underline{\hat{\boldsymbol{X}}_{\boldsymbol{\epsilon}}}\big)}_{\text{Reconstruction term}} \quad \Big] + \beta \underbrace{\sum_{i=1}^{d} \Big(\mu_i(\underline{\boldsymbol{X}})^2 + \sigma_i(\underline{\boldsymbol{X}})^2 - \log \sigma_i(\underline{\boldsymbol{X}})^2 - 1\Big)}_{\text{Regularization (concentration to origin)}}. \tag{12}$$

Each term in (12) is also **architecture-independent**. Once the decoder is trained, inference is **only the application of the decoder**.

3.3 Consequence of the General Theory

From the above, it is clear that, outside the context of image generation, both the noise estimator and the VAE decoder can adopt **any architecture**.

4 Necessity of Variable I/O Sizes and Convolutional Layers

4.1 Necessity of Variable Input/Output Sizes

In practical image generation, the **output resolution (dimensions)** depends on user requirements and cannot be fixed at training time. Therefore, an architecture with **variable input/output sizes**, which allows selecting the output size by choosing the input size (latent resolution), is necessary.

4.2 Layer Achieving Variable I/O Sizes: Convolutional Layer

To construct variable input/output sizes, each layer only needs to be a **parametric function compatible with variable sizes**. A typical example is the **convolution layer (implemented as cross-correlation)**. The noise estimator in Stable Diffusion 1.5 is a **U-Net** [5] family (conditional, with attention), and the VAE is also constructed with convolutional systems [4].

5 Overview of Stable Diffusion 1.5 U-Net and VAE with Diagrams

5.1 Influence of Different Motivations for U-Net and VAE on Architectural Differences

U-Net and VAE have very different expected functionalities.

- U-Net's purpose is to generate a **globally coherent** low-resolution latent image from noise (which has no information), using information from a text encoder.
- VAE's purpose is to convert a low-resolution latent image, which already has some global coherence, into a high-resolution natural image by defining the details.

This is evident even when observing the intermediate states of image generation.

Corresponding to these differences in motivation, the architectures actually used for U-Net and VAE encoders also differ.

- U-Net, to achieve global coherence, uses downsampling, giving it a structure that efficiently allows pixels at one edge to influence pixels at the opposite edge with a relatively small number of layers.
- The VAE encoder is composed of pure convolutional layers and upsampling layers, adopting a structure that restricts the use of parameters to local value transformations.

Let's confirm these differences by looking at the actual architectures.

5.2 Understanding the Implemented Architecture

The actual architecture may differ in details from the paper's description. Using tools like **torchview**, one can visualize the **computation graph** from the implemented model, making it easier to grasp implementation differences[1].

5.3 Block Diagram of U-Net (Conditional)

Figure 1 shows a schematic of the conditional U-Net in Stable Diffusion 1.5 (including downsample, bottleneck, upsample, skip connections, and cross-attention). This figure is obtained by applying torchview with `depth=1` to the U-Net part of Stable Diffusion 1.5 uploaded to Hugging Face, reflecting the actual implementation.

Remark 5.1. Difference from the original U-Net: Ronneberger et al.'s U-Net [5] is an **image-to-image** mapping for **medical image segmentation**, where both input and output

[1] `torchview`: `https://github.com/mert-kurttutan/torchview`

are images. In image generation AI, it needs to accept **time (noise level)** and **text conditions** as input, and perform **noise estimation (or v-prediction)**. Therefore, the significant differences are the addition of a **time encoder** and **cross-attention layers (for text)** [4].

5.4 Block Diagram of VAE (Decoder)

Figure 2 shows the architecture of the VAE decoder in Stable Diffusion 1.5 (Conv/ResBlock-/Upsample). This figure is obtained by applying torchview with `depth=1` to the VAE part of Stable Diffusion 1.5 uploaded to Hugging Face, reflecting the actual implementation.

5.5 Confirmation of Variable Input/Output Sizes

The fact that U-Net and VAE have variable input/output sizes follows from each component layer (convolution, normalization, attention, up/down-sample) being defined **convolutionally (translationally equivariant under isomorphism)** with respect to the **spatial resolution** $H \times W$.

Remark 5.2. The **variable input/output sizes** mentioned here refer to the **variability in the image width W and height H**; the **number of channels C is fixed** (although the channel width may change in steps inside the U-Net, the number of channels at the input/output interface is specified).

6 Formal Definitions of Layers (Using only cross-correlation, matrix operations, and elementary functions)

In the following, all layers (functions) used in the U-Net (conditional) and VAE decoder will be **formally** defined using only **cross-correlation (convolution in practice), matrix operations, elementary functions, and well-known special functions**. In each definition, the **function name (matching the library class/function name)** is given by $\cdot$, **non-learnable hyperparameters** (e.g., stride, padding, groups) are indicated in superscript parentheses $^{(\cdot)}$, and the **entire set of learnable parameters** is specified in the subscript. The URL of the corresponding implementation is provided in a footnote.

6.1 Cross-Correlation (2D "Convolution")

Definition 6.1 (Conv2d (Convolution in practice is cross-correlation))**.** For input $\underline{\boldsymbol{X}} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$ and output $\underline{\boldsymbol{Y}} \in \mathbb{R}^{C_{\text{out}} \times H' \times W'}$, we fix **hyperparameters** kernel size (k_h, k_w), stride (s_h, s_w), and padding (p_h, p_w). The set of **learnable parameters** (weights and biases) is

$$\Theta_{\text{Conv2d}} = \left\{ \boldsymbol{W}^{(o)} \in \mathbb{R}^{C_{\text{in}} \times k_h \times k_w},\ b_o \in \mathbb{R} \right\}_{o=1}^{C_{\text{out}}} \tag{13}$$

At this time,

$$\left(\mathrm{Conv2d}^{(k_h,k_w;\ s_h,s_w;\ p_h,p_w)}_{\Theta_{\mathrm{Conv2d}}}(\underline{\boldsymbol{X}})\right)_{o,i,j} = b_o + \sum_{c=1}^{C_{\mathrm{in}}}\sum_{u=1}^{k_h}\sum_{v=1}^{k_w} W^{(o)}_{c,u,v}\, X_{c,\ i\cdot s_h+u-p_h,\ j\cdot s_w+v-p_w}. \tag{14}$$

The output spatial size is $H' = \left\lfloor \frac{H - k_h + 2p_h}{s_h} \right\rfloor + 1,\ W' = \left\lfloor \frac{W - k_w + 2p_w}{s_w} \right\rfloor + 1.$[a]

[a] torch.nn.Conv2d: `https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html`, torch.nn.functional.conv2d: `https://pytorch.org/docs/stable/generated/torch.nn.functional.conv2d.html`.

Remark 6.1. ”Convolution” in implementations is cross-correlation (**does not flip the kernel**) and matches (14).

6.2 Linear Layer (Fully-Connected)

Definition 6.2 (Linear)**.** For input $\boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{in}}}$, output $\boldsymbol{y} \in \mathbb{R}^{d_{\mathrm{out}}}$, and learnable parameters $\Theta_{\mathrm{Linear}} = \{\boldsymbol{W} \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{in}}},\ \boldsymbol{b} \in \mathbb{R}^{d_{\mathrm{out}}}\}$,

$$\mathrm{Linear}_{\Theta_{\mathrm{Linear}}}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}. \tag{15}$$

[a]

[a] torch.nn.Linear: `https://pytorch.org/docs/stable/generated/torch.nn.Linear.html`, torch.nn.functional.linear: `https://pytorch.org/docs/stable/generated/torch.nn.functional.linear.html`.

6.3 Activation Functions and Softmax

Definition 6.3 (SiLU (Swish))**.** For a component u of an arbitrary-dimensional tensor,

$$\mathrm{SiLU}(u) = u\,\sigma(u), \quad \sigma(u) = \frac{1}{1+e^{-u}}. \tag{16}$$

[a]

[a] torch.nn.SiLU: `https://pytorch.org/docs/stable/generated/torch.nn.SiLU.html`, torch.nn.functional.silu: `https://pytorch.org/docs/stable/generated/torch.nn.functional.silu.html`.

Definition 6.4 (Softmax)**.** For $\boldsymbol{x} \in \mathbb{R}^n$, fixing the temperature $\tau > 0$,

$$\left(\mathrm{Softmax}^{(\tau)}(\boldsymbol{x})\right)_i = \frac{\exp(x_i/\tau)}{\sum_{j=1}^n \exp\left(x_j/\tau\right)}. \tag{17}$$

6.4 Normalization

Definition 6.5 (GroupNorm)**.** For input $\underline{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$, **hyperparameter** number of groups $G \mid C$, and **learnable parameters** $\Theta_{\mathrm{GroupNorm}} = \{\gamma \in \mathbb{R}^C,\ \beta \in \mathbb{R}^C\}$, the mean and variance for each group g are

$$\mu_g = \frac{1}{|S_g|}\sum_{(c,i,j)\in S_g} X_{c,i,j}, \qquad \sigma_g^2 = \frac{1}{|S_g|}\sum_{(c,i,j)\in S_g} (X_{c,i,j}-\mu_g)^2, \tag{18}$$

The output is

$$\left(\text{GroupNorm}^{(G)}_{\Theta_{\mathrm{GroupNorm}}}(\underline{\boldsymbol{X}})\right)_{c,i,j} = \gamma_c \frac{X_{c,i,j}-\mu_{g(c)}}{\sqrt{\sigma^2_{g(c)}+\varepsilon}} + \beta_c. \tag{19}$$

[a]

[a] torch.nn.GroupNorm: `https://pytorch.org/docs/stable/generated/torch.nn.GroupNorm.html`.

6.5 Up/Downsample

6.5.1 Definition of Downsample by Strided Cross-Correlation

Using Conv2d with stride 2:

Definition 6.6 (Downsample2D (by Conv2d))**.**

$$\text{Downsample2D}^{(2)}_{\Theta_{\mathrm{Down}}}(\underline{\boldsymbol{X}}) := \text{Conv2d}^{(k_h,k_w;\ 2,2;\ p_h,p_w)}_{\Theta_{\mathrm{Down}}}(\underline{\boldsymbol{X}}). \tag{20}$$

(Θ_{Down} is the set of learnable parameters for Conv2d).

6.5.2 Explicit Definition based on Element-wise Formulas (Nearest/Average Pooling)

Definition 6.7 (Downsample2DNearest (nearest pooling; element-wise))**.** For $\underline{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$, $\underline{\boldsymbol{Y}} \in \mathbb{R}^{C\times\lfloor H/2\rfloor\times\lfloor W/2\rfloor}$ is defined as

$$Y_{c,i,j} = X_{c,\,2i,\,2j} \quad (0 \le i < \lfloor H/2\rfloor,\ 0 \le j < \lfloor W/2\rfloor) \tag{21}$$

(No learnable parameters)

Definition 6.8 (Downsample2DAverage (average pooling; element-wise))**.** For $\underline{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$,

$$Y_{c,i,j} = \frac{1}{4}\sum_{u=0}^{1}\sum_{v=0}^{1} X_{c,\,2i+u,\,2j+v}. \tag{22}$$

(No learnable parameters)

6.5.3 Upsample by Nearest-Neighbor Interpolation and Cross-Correlation

Definition 6.9 (Interpolate (nearest; element-wise))**.** For $\underline{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$, $\underline{\boldsymbol{Z}} = \text{Interpolate}^{(\times 2,\text{nearest})}(\underline{\boldsymbol{X}}) \in \mathbb{R}^{C\times 2H\times 2W}$ is defined as

$$Z_{c,\,2i+u,\,2j+v} = X_{c,i,j} \quad (u, v \in \{0, 1\}) \tag{23}$$

[a]

[a] torch.nn.functional.interpolate: `https://pytorch.org/docs/stable/generated/torch.nn.functional.interpolate.html`.

Definition 6.10 (Upsample2D (nearest+conv))**.**

$$\text{Upsample2D}^{(2)}_{\Theta_{\text{Up}}}(\underline{\boldsymbol{X}}) := \text{Conv2d}^{(k_h,k_w;\ 1,1;\ p_h,p_w)}_{\Theta_{\text{Up}}}\big(\text{Interpolate}^{(\times 2,\text{nearest})}(\underline{\boldsymbol{X}})\big). \tag{24}$$

6.6 Reshape, Concatenate

Definition 6.11 (flatten, unflatten, Concat)**.** For $\underline{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$,

$$\text{flatten}_{(H,W)}(\underline{\boldsymbol{X}}) \in \mathbb{R}^{(HW)\times C}, \quad \big(\text{flatten}_{(H,W)}(\underline{\boldsymbol{X}})\big)_{(i-1)W+j,\ c} = X_{c,i,j}, \tag{25}$$

$$\text{unflatten}_{(H,W)}(\boldsymbol{Y}) \in \mathbb{R}^{C\times H\times W}, \quad \big(\text{unflatten}_{(H,W)}(\boldsymbol{Y})\big)_{c,i,j} = Y_{(i-1)W+j,\ c}, \tag{26}$$

$$\text{Concat}(\underline{\boldsymbol{A}}, \underline{\boldsymbol{B}}) = \underline{\boldsymbol{A}} \oplus \underline{\boldsymbol{B}} \text{ (concatenation along the channel dimension).} \tag{27}$$

6.7 Timestep Embedding (Noise Level)

Definition 6.12 (TimestepEmbedding (Sinusoidal + MLP))**.** For scalar $t \in \mathbb{R}$ and frequency sequence $\omega_r = \omega_0 \beta^{r-1}$ $(r = 1, \ldots, R)$,

$$\boldsymbol{e}(t) = \big[\cos(\omega_1 t), \sin(\omega_1 t), \ldots, \cos(\omega_R t), \sin(\omega_R t)\big]^\top \in \mathbb{R}^{2R}. \tag{28}$$

For learnable parameters $\Theta_{\text{TE}} = \{\boldsymbol{U}_1 \in \mathbb{R}^{d_h\times 2R},\ \boldsymbol{b}_1 \in \mathbb{R}^{d_h},\ \boldsymbol{U}_2 \in \mathbb{R}^{d_t\times d_h},\ \boldsymbol{b}_2 \in \mathbb{R}^{d_t}\}$,

$$\text{TimestepEmbedding}_{\Theta_{\text{TE}}}(t) = \boldsymbol{U}_2\, \text{SiLU}(\boldsymbol{U}_1\, \boldsymbol{e}(t) + \boldsymbol{b}_1) + \boldsymbol{b}_2 \in \mathbb{R}^{d_t}. \tag{29}$$

[a]

[a] Diffusers implementation (`embeddings.py` / `unet_2d_condition.py`): `https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/embeddings.py`, `https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/unet_2d_condition.py`.

Remark 6.2. Components with small ω represent the **coarse position (low frequency, long period)** of t, while components with large ω represent the **fine position (high frequency, short period)**. This is analogous to the **positional numeral system**, where **upper digits**

are useful for approximate estimation, and **lower digits** independently provide information about divisors or remainders.

6.8 Scaled Dot-Product Attention (Self/Cross Attention)

Definition 6.13 (ScaledDotProductAttention)**.** For query $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$, key $\boldsymbol{K} \in \mathbb{R}^{M\times d}$, and value $\boldsymbol{V} \in \mathbb{R}^{M\times d_v}$,

$$\text{ScaledDotProductAttention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}^{(\sqrt{d})^{-1}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{V}. \tag{30}$$

[a]

[a]torch.nn.functional.scaled_dot_product_attention: https://pytorch.org/docs/stable/generated/torch.nn.functional.scaled_dot_product_attention.html.

Definition 6.14 (MultiheadAttention (SDPA with Projections))**.** For input sequence $\boldsymbol{X} \in \mathbb{R}^{N\times d_{\text{in}}}$ and context sequence $\boldsymbol{C} \in \mathbb{R}^{M\times d_{\text{ctx}}}$, using learnable parameters

$$\Theta_{\text{MHA}} = \{\boldsymbol{W}_Q \in \mathbb{R}^{d_{\text{in}}\times d},\ \boldsymbol{W}_K \in \mathbb{R}^{d_{\text{ctx}}\times d},\ \boldsymbol{W}_V \in \mathbb{R}^{d_{\text{ctx}}\times d_v},\ \boldsymbol{W}_O \in \mathbb{R}^{d_v\times d_{\text{out}}}\} \tag{31}$$

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q, \quad \boldsymbol{K} = \boldsymbol{C}\boldsymbol{W}_K, \quad \boldsymbol{V} = \boldsymbol{C}\boldsymbol{W}_V, \tag{32}$$

$$\text{MultiheadAttention}_{\Theta_{\text{MHA}}}(\boldsymbol{X}, \boldsymbol{C}) = \text{ScaledDotProductAttention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})\,\boldsymbol{W}_O. \tag{33}$$

[a]

[a]torch.nn.MultiheadAttention: https://pytorch.org/docs/stable/generated/torch.nn.MultiheadAttention.html. Diffusers attention mechanism (attn_processor.py): https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/attn_processor.py.

6.8.1 Programming Intuition of a Dictionary (Key-Value Map)

A **dictionary (map)** in programming is a correspondence of {key : value}, a structure that retrieves the corresponding value when a key is given.

Example 6.1 (Python Dictionary)**.** For `D = {"cat" : 1, "dog" : 2}`, `D["dog"] = 2`. The ScaledDotProductAttention in attention implements a **soft dictionary** that ”retrieves a weighted sum of values closest to the key” in a continuous vector space.

Proposition 6.1 (Attention as a Soft Dictionary (Symbol correspondence for ScaledDotProductAttention))**.** Let $\boldsymbol{K} = [\boldsymbol{k}_1^\top; \ldots; \boldsymbol{k}_M^\top] \in \mathbb{R}^{M\times d}$, $\boldsymbol{V} = [\boldsymbol{v}_1^\top; \ldots; \boldsymbol{v}_M^\top] \in \mathbb{R}^{M\times d_v}$, and consider a single query $\boldsymbol{q} \in \mathbb{R}^d$ with $\boldsymbol{Q} = [\boldsymbol{q}^\top]$. Then

$$\text{ScaledDotProductAttention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \Big[\sum_{m=1}^{M} \pi_m(\boldsymbol{q})\, \boldsymbol{v}_m \Big], \tag{34}$$

$$\pi_m(\boldsymbol{q}) = \frac{\exp\Big(\langle \boldsymbol{q}, \boldsymbol{k}_m \rangle / \sqrt{d}\Big)}{\sum_{j=1}^{M} \exp\Big(\langle \boldsymbol{q}, \boldsymbol{k}_j \rangle / \sqrt{d}\Big)}. \tag{35}$$

Here, $\pi(\boldsymbol{q})$ is the first row of Softmax$\big((\boldsymbol{Q}\boldsymbol{K}^\top)/\sqrt{d}\big)$ and corresponds perfectly to the Softmax in (30).

Proof. The m-th component of $\boldsymbol{Q}\boldsymbol{K}^\top/\sqrt{d} \in \mathbb{R}^{1\times M}$ is $\langle \boldsymbol{q}, \boldsymbol{k}_m \rangle/\sqrt{d}$, so applying Softmax yields the weight $\pi_m(\boldsymbol{q})$. Multiplying by $\boldsymbol{V}$ from the right yields $\sum_m \pi_m(\boldsymbol{q})\, \boldsymbol{v}_m$. This can be shown by expanding equation (30) component-wise. □

Remark 6.3. If $\pi(\boldsymbol{q})$ becomes a one-hot vector (1 for some $m^\star$, 0 otherwise), then $\sum_m \pi_m(\boldsymbol{q})\, \boldsymbol{v}_m = \boldsymbol{v}_{m^\star}$, which matches the exact retrieval from a dictionary.

Example 6.2 (Numerical Calculation of ScaledDotProductAttention)**.** Let $d = 2$, $\boldsymbol{q} = (1,0)^\top$, $\boldsymbol{k}_1 = (1,0)^\top$, $\boldsymbol{k}_2 = (0,1)^\top$, $\boldsymbol{v}_1 = (2,0)^\top$, $\boldsymbol{v}_2 = (0,3)^\top$. At this time, the inner products are

$$\langle \boldsymbol{q}, \boldsymbol{k}_1 \rangle = 1\cdot 1 + 0 \cdot 0 = 1, \quad \langle \boldsymbol{q}, \boldsymbol{k}_2 \rangle = 1 \cdot 0 + 0 \cdot 1 = 0 \tag{36}$$

As the exponentials of the scores divided by the square root of d,

$$\sqrt{d} = \sqrt{2}, \tag{37}$$

$$s_1 = \exp\big(\langle \boldsymbol{q}, \boldsymbol{k}_1 \rangle/\sqrt{2}\big) = \exp\Big(1/\sqrt{2}\Big), \quad s_2 = \exp\big(\langle \boldsymbol{q}, \boldsymbol{k}_2 \rangle/\sqrt{2}\big) = \exp(0) = 1 \tag{38}$$

we obtain. Numerically, from $1/\sqrt{2} \approx 0.70710678$,

$$s_1 \approx e^{0.70710678} \approx 2.02811498, \qquad s_2 = 1. \tag{39}$$

The softmax weights are

$$\pi_1 = \frac{s_1}{s_1 + s_2} = \frac{e^{1/\sqrt{2}}}{e^{1/\sqrt{2}} + 1}, \qquad \pi_2 = \frac{s_2}{s_1 + s_2} = \frac{1}{e^{1/\sqrt{2}} + 1} \tag{40}$$

and numerically,

$$\pi_1 \approx \frac{2.02811498}{2.02811498 + 1} \approx 0.66976155, \qquad \pi_2 \approx \frac{1}{2.02811498 + 1} \approx 0.33023845. \tag{41}$$

The output is

$$
\begin{aligned}
\boldsymbol{o} &= \pi_1 \boldsymbol{v}_1 + \pi_2 \boldsymbol{v}_2 = \pi_1 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + \pi_2 \begin{bmatrix} 0 \\ 3 \end{bmatrix} = \begin{bmatrix} 2\pi_1 \\ 3\pi_2 \end{bmatrix} \\
&= \left(\frac{2s_1}{s_1 + 1}, \; \frac{3}{s_1 + 1} \right)^{\top} \quad \text{(Substituting Eq. (38))},
\end{aligned} \tag{42}
$$

and numerically,

$$
\boldsymbol{o} \approx \begin{bmatrix} 2 \times 0.66976155 \\ 3 \times 0.33023845 \end{bmatrix} = \begin{bmatrix} 1.33952310 \\ 0.99071535 \end{bmatrix}. \tag{43}
$$

As shown above, we can follow step-by-step in the order of exponential calculation → normalization → weighted sum of values.

Exercise 6.1 (Numerical Example with 2D Vectors)**.** Let $d = 2$, $\boldsymbol{q} = (2, 1)^{\top}$, $\boldsymbol{k}_1 = (1, 0)^{\top}$, $\boldsymbol{k}_2 = (0, 1)^{\top}$, $\boldsymbol{v}_1 = (1, 2)^{\top}$, $\boldsymbol{v}_2 = (4, -1)^{\top}$. Calculate the output vector $\boldsymbol{o}$ of ScaledDotProductAttention both as an exact expression and numerically ($\sqrt{d} = \sqrt{2}$).

Answer. First, calculate the inner products:

$$
\langle \boldsymbol{q}, \boldsymbol{k}_1 \rangle = 2 \cdot 1 + 1 \cdot 0 = 2, \qquad \langle \boldsymbol{q}, \boldsymbol{k}_2 \rangle = 2 \cdot 0 + 1 \cdot 1 = 1. \tag{44}
$$

Thus, the scores are

$$
\begin{aligned}
s_1 &= \exp(\langle \boldsymbol{q}, \boldsymbol{k}_1 \rangle / \sqrt{2}) = \exp\left(2/\sqrt{2}\right) = \exp\left(\sqrt{2}\right), \\
s_2 &= \exp(\langle \boldsymbol{q}, \boldsymbol{k}_2 \rangle / \sqrt{2}) = \exp\left(1/\sqrt{2}\right)
\end{aligned} \tag{45}
$$

Numerically, from $\sqrt{2} \approx 1.41421356$ and $1/\sqrt{2} \approx 0.70710678$,

$$
s_1 \approx e^{1.41421356} \approx 4.11325038, \qquad s_2 \approx e^{0.70710678} \approx 2.02811498. \tag{46}
$$

Therefore, the softmax weights are

$$
\pi_1 = \frac{s_1}{s_1 + s_2} = \frac{e^{\sqrt{2}}}{e^{\sqrt{2}} + e^{1/\sqrt{2}}}, \qquad \pi_2 = \frac{s_2}{s_1 + s_2} = \frac{e^{1/\sqrt{2}}}{e^{\sqrt{2}} + e^{1/\sqrt{2}}}, \tag{47}
$$

Numerically,

$$
\pi_1 \approx \frac{4.11325038}{4.11325038 + 2.02811498} \approx 0.66976155, \qquad \pi_2 \approx \frac{2.02811498}{4.11325038 + 2.02811498} \approx 0.33023845. \tag{48}
$$

The output is

$$\boldsymbol{o} = \pi_1 \boldsymbol{v}_1 + \pi_2 \boldsymbol{v}_2 = \pi_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \pi_2 \begin{bmatrix} 4 \\ -1 \end{bmatrix} = \begin{bmatrix} \pi_1 + 4\pi_2 \\ 2\pi_1 - \pi_2 \end{bmatrix} \tag{49}$$

$$= \begin{bmatrix} 4 - 3\pi_1 \\ 3\pi_1 - 1 \end{bmatrix} \quad \text{(Simplified using } \pi_2 = 1 - \pi_1\text{)}.$$

Numerically,

$$\boldsymbol{o} \approx \begin{bmatrix} 4 - 3 \times 0.66976155 \\ 3 \times 0.66976155 - 1 \end{bmatrix} = \begin{bmatrix} 1.99071535 \\ 1.00928465 \end{bmatrix}. \tag{50}$$

This completes the derivation of the exact expression and the numerical calculation.

Proposition 6.2 (Limit to a Hard Dictionary (ScaledDotProductAttention))**.** Suppose for some $m^\star$, $\boldsymbol{k}_{m^\star} \parallel \boldsymbol{q}$ and $\boldsymbol{k}_m \perp \boldsymbol{q}$ $(m \neq m^\star)$. Then, for any $\alpha > 0$, let $\boldsymbol{q}_\alpha = \alpha \boldsymbol{q}$,

$$\lim_{\alpha \to +\infty} \pi_m(\boldsymbol{q}_\alpha) = \begin{cases} 1, & m = m^\star, \\ 0, & m \neq m^\star. \end{cases} \tag{51}$$

Proof. $\langle \boldsymbol{q}_\alpha, \boldsymbol{k}_{m^\star} \rangle = \alpha \|\boldsymbol{q}\| \, \|\boldsymbol{k}_{m^\star}\|$, and others are 0. Therefore, only the $m^\star$-th component of $\boldsymbol{Q}\boldsymbol{K}^\top / \sqrt{d}$ increases linearly with α. From the definition of Softmax (17), the exponential of the largest component dominates the others, and (51) follows. □

Remark 6.4. $\boldsymbol{k}_{m^\star} \parallel \boldsymbol{q}$ means "**query and key are in the same direction (identical content)**", which causes the corresponding $\boldsymbol{v}_{m^\star}$ to be selected. That is, it matches the "**retrieval of the value corresponding to the key**".

6.9 Time-Conditioned Residual Block

Definition 6.15 (ResnetBlock2D (Affine modulation by time embedding; FiLM))**.** For input $\underline{\boldsymbol{X}} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$ and time embedding $\boldsymbol{h} \in \mathbb{R}^{d_t}$, using learnable parameters

$$\Theta_{\text{ResnetBlock2D}} = \left(\Theta^{(1)}_{\text{Conv2d}}, \Theta^{(2)}_{\text{Conv2d}}, \Theta^{(s)}_{\text{Conv2d}}, \Theta^{(1)}_{\text{GN}}, \Theta^{(2)}_{\text{GN}}, \Theta^{(\gamma)}_{\text{Linear}}, \Theta^{(\beta)}_{\text{Linear}} \right),$$

$$\underline{\boldsymbol{U}}_1 = \text{GroupNorm}^{(G)}_{\Theta^{(1)}_{\text{GN}}}(\underline{\boldsymbol{X}}), \quad \underline{\boldsymbol{V}}_1 = \text{SiLU}(\underline{\boldsymbol{U}}_1), \quad \underline{\boldsymbol{W}}_1 = \text{Conv2d}^{(k,k;\ 1,1;\ p,p)}_{\Theta^{(1)}_{\text{Conv2d}}}(\underline{\boldsymbol{V}} \tag{52}$$

$$\boldsymbol{\gamma}(\boldsymbol{h}) = \text{Linear}_{\Theta^{(\gamma)}_{\text{Linear}}}(\boldsymbol{h}), \quad \boldsymbol{\beta}(\boldsymbol{h}) = \text{Linear}_{\Theta^{(\beta)}_{\text{Linear}}}(\boldsymbol{h}), \tag{53}$$

$$\underline{\boldsymbol{U}}_2 = \text{GroupNorm}^{(G)}_{\Theta^{(2)}_{\text{GN}}}(\underline{\boldsymbol{W}}_1), \quad \widehat{\underline{\boldsymbol{U}}}_2 = \boldsymbol{\gamma}(\boldsymbol{h}) \odot \underline{\boldsymbol{U}}_2 + \boldsymbol{\beta}(\boldsymbol{h}), \tag{54}$$

$$\underline{\boldsymbol{V}}_2 = \text{SiLU}(\widehat{\underline{\boldsymbol{U}}}_2), \quad \underline{\boldsymbol{W}}_2 = \text{Conv2d}^{(k,k;\ 1,1;\ p,p)}_{\Theta^{(2)}_{\text{Conv2d}}}(\underline{\boldsymbol{V}}_2), \tag{55}$$

$$\underline{\boldsymbol{S}} = \text{Conv2d}^{(1,1;\ 1,1;\ 0,0)}_{\Theta^{(s)}_{\text{Conv2d}}}(\underline{\boldsymbol{X}}) \quad \text{(channel matching)}, \tag{56}$$

$$\text{ResnetBlock2D}_{\Theta_{\text{ResnetBlock2D}}}(\underline{\boldsymbol{X}}, \boldsymbol{h}) = \underline{\boldsymbol{S}} + \underline{\boldsymbol{W}}_2. \tag{57}$$

[a]

[a] Diffusers ResnetBlock2D implementation: `https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/resnet.py`.

6.10 U-Net Construction Blocks (Down/Up/Mid)

Definition 6.16 (DownBlock2D)**.** Taking the number of residual layers within the level $n \in \mathbb{Z}_{>0}$ as a hyperparameter, and for learnable parameters $\Theta_{\text{DownBlock2D}} = \big(\{\Theta^{(r)}_{\text{Res}}\}_{r=1}^{n},\ \Theta_{\text{Down}}\big)$,

$$\underline{\boldsymbol{H}}_0 = \underline{\boldsymbol{X}}, \quad \underline{\boldsymbol{H}}_r = \text{ResnetBlock2D}_{\Theta^{(r)}_{\text{Res}}}(\underline{\boldsymbol{H}}_{r-1}, \boldsymbol{h}) \ (r = 1, \ldots, n), \tag{58}$$

$$\text{DownBlock2D}_{\Theta_{\text{DownBlock2D}}}(\underline{\boldsymbol{X}}, \boldsymbol{h}) = \text{Downsample2D}^{(2)}_{\Theta_{\text{Down}}}(\underline{\boldsymbol{H}}_n). \tag{59}$$

[a]

[a] Diffusers block implementation: `https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/unet_2d_blocks.py`.

Definition 6.17 (UpBlock2D)**.** Combining the skip connection $\underline{\boldsymbol{S}}$ and the input from the bottom $\underline{\boldsymbol{X}}$ with Concat, and for learnable parameters $\Theta_{\text{UpBlock2D}} = \big(\{\Theta^{(r)}_{\text{Res}}\}_{r=1}^{n},\ \Theta_{\text{Up}}\big)$,

$$\underline{\boldsymbol{Y}}_0 = \text{Concat}\Big(\text{Upsample2D}^{(2)}_{\Theta_{\text{Up}}}(\underline{\boldsymbol{X}}),\ \underline{\boldsymbol{S}}\Big), \tag{60}$$

$$\underline{\boldsymbol{Y}}_r = \text{ResnetBlock2D}_{\Theta^{(r)}_{\text{Res}}}(\underline{\boldsymbol{Y}}_{r-1}, \boldsymbol{h}) \ (r = 1, \ldots, n), \tag{61}$$

$$\text{UpBlock2D}_{\Theta_{\text{UpBlock2D}}}(\underline{\boldsymbol{X}}, \underline{\boldsymbol{S}}, \boldsymbol{h}) = \underline{\boldsymbol{Y}}_n. \tag{62}$$

[a]

[a] Implementation reference: `https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/unet_2d_blocks.py`.

Definition 6.18 (CrossAttentionMidBlock2D (Using Self/Cross Attention))**.** For learnable parameters $\Theta_{\mathrm{MidBlock2D}} = (\Theta^{(1)}_{\mathrm{Res}},\ \Theta^{\mathrm{self}}_{\mathrm{MHA}},\ \Theta^{\mathrm{cross}}_{\mathrm{MHA}},\ \Theta^{(2)}_{\mathrm{Res}})$ and text context $\boldsymbol{C} \in \mathbb{R}^{M\times d_{\mathrm{ctx}}}$,

$$\underline{\boldsymbol{A}}_0 = \text{ResnetBlock2D}_{\Theta^{(1)}_{\mathrm{Res}}}(\underline{\boldsymbol{X}}, \boldsymbol{h}), \tag{63}$$

$$\boldsymbol{X}_{\mathrm{flat}} = \text{flatten}_{(H,W)}(\underline{\boldsymbol{A}}_0) \in \mathbb{R}^{(HW)\times d_{\mathrm{in}}}, \tag{64}$$

$$\boldsymbol{B}_1 = \text{MultiheadAttention}_{\Theta^{\mathrm{self}}_{\mathrm{MHA}}}(\boldsymbol{X}_{\mathrm{flat}}, \boldsymbol{X}_{\mathrm{flat}}), \tag{65}$$

$$\boldsymbol{B}_2 = \text{MultiheadAttention}_{\Theta^{\mathrm{cross}}_{\mathrm{MHA}}}(\boldsymbol{B}_1, \boldsymbol{C}), \tag{66}$$

$$\underline{\boldsymbol{A}}_1 = \text{unflatten}_{(H,W)}(\boldsymbol{B}_2), \qquad \text{CrossAttentionMidBlock2D}_{\Theta_{\mathrm{MidBlock2D}}}(\underline{\boldsymbol{X}}, \boldsymbol{h}, \boldsymbol{C}) = \text{ResnetBlock2D}_{\Theta^{(2)}_{\mathrm{Res}}}(\underline{\boldsymbol{A}}_1, \tag{67}$$

a

[a]Processor for self/cross attention: `https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/attn_processor.py`.

Definition 6.19 (CrossAttentionDownBlock2D)**.** Let $n \in \mathbb{Z}_{>0}$ be the number of residual layers in the level. Self-attention and cross-attention use the mapping defined in §6.8 for the 'sequence of tokens per spatial position'. Let the learnable parameters be

$$\Theta_{\mathrm{XDown}} = \left(\{\Theta^{(r)}_{\mathrm{Res}}\}_{r=1}^{n},\ \{\Theta^{(r)}_{\mathrm{MHA,self}}\}_{r=1}^{n},\ \{\Theta^{(r)}_{\mathrm{MHA,cross}}\}_{r=1}^{n},\ \Theta_{\mathrm{Down}}\right)$$

For an input $\underline{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$, time embedding $\boldsymbol{h} \in \mathbb{R}^{d_t}$, and text context $\boldsymbol{C} \in \mathbb{R}^{M\times d_{\mathrm{ctx}}}$,

$$\underline{\boldsymbol{H}}_0 = \underline{\boldsymbol{X}},$$

$$\underline{\boldsymbol{H}}_r = \text{ResnetBlock2D}_{\Theta^{(r)}_{\mathrm{Res}}}(\underline{\boldsymbol{H}}_{r-1}, \boldsymbol{h}), \tag{68}$$

$$\boldsymbol{T}_r = \text{flatten}_{(H,W)}(\underline{\boldsymbol{H}}_r), \tag{69}$$

$$\boldsymbol{U}_r = \text{MultiheadAttention}_{\Theta^{(r)}_{\mathrm{MHA,self}}}(\boldsymbol{T}_r, \boldsymbol{T}_r), \tag{70}$$

$$\boldsymbol{V}_r = \text{MultiheadAttention}_{\Theta^{(r)}_{\mathrm{MHA,cross}}}(\boldsymbol{U}_r, \boldsymbol{C}), \tag{71}$$

$$\underline{\boldsymbol{H}}^{\star}_r = \text{unflatten}_{(H,W)}(\boldsymbol{V}_r), \qquad r = 1, \ldots, n, \tag{72}$$

$$\text{CrossAttentionDownBlock2D}_{\Theta_{\mathrm{XDown}}}(\underline{\boldsymbol{X}}, \boldsymbol{h}, \boldsymbol{C}) = \text{Downsample2D}^{(2)}_{\Theta_{\mathrm{Down}}}\left(\underline{\boldsymbol{H}}^{\star}_n\right). \tag{73}$$

Definition 6.20 (CrossAttentionUpBlock2D)**.** For learnable parameters

$$\Theta_{\mathrm{XUp}} = \left(\{\Theta^{(r)}_{\mathrm{Res}}\}_{r=1}^{n},\ \{\Theta^{(r)}_{\mathrm{MHA,self}}\}_{r=1}^{n},\ \{\Theta^{(r)}_{\mathrm{MHA,cross}}\}_{r=1}^{n},\ \Theta_{\mathrm{Up}}\right)$$

and for an input $\underline{\boldsymbol{X}}$ from the bottom, a skip connection $\underline{\boldsymbol{S}}$ with the same resolution, a time embedding $\boldsymbol{h}$, and context $\boldsymbol{C}$:

$$
\begin{aligned}
\underline{\boldsymbol{Y}}_0 &= \mathsf{Concat}\Big(\mathsf{Upsample2D}^{(2)}_{\Theta_{\mathrm{Up}}}(\underline{\boldsymbol{X}}),\ \underline{\boldsymbol{S}}\Big), && (74)\\
\underline{\boldsymbol{Y}}_r &= \mathsf{ResnetBlock2D}_{\Theta^{(r)}_{\mathrm{Res}}}(\underline{\boldsymbol{Y}}_{r-1}, \boldsymbol{h}), \qquad r = 1, \dots, n, && (75)\\
\boldsymbol{T}_r &= \mathsf{flatten}_{(H,W)}(\underline{\boldsymbol{Y}}_r), \quad \boldsymbol{U}_r = \mathsf{MultiheadAttention}_{\Theta^{(r)}_{\mathrm{MHA,self}}}(\boldsymbol{T}_r, \boldsymbol{T}_r), && (76)\\
\boldsymbol{V}_r &= \mathsf{MultiheadAttention}_{\Theta^{(r)}_{\mathrm{MHA,cross}}}(\boldsymbol{U}_r, \boldsymbol{C}), \quad \underline{\boldsymbol{Y}}^{\star}_r = \mathsf{unflatten}_{(H,W)}(&& (77)
\end{aligned}
$$

$$\mathsf{CrossAttentionUpBlock2D}_{\Theta_{\mathrm{XUp}}}(\underline{\boldsymbol{X}}, \underline{\boldsymbol{S}}, \boldsymbol{h}, \boldsymbol{C}) = \underline{\boldsymbol{Y}}^{\star}_n. \tag{78}$$

6.11 VAE Decoder's Terminal Mapping (RGB Output; 1×1 conv)

Definition 6.21 (Conv1x1 (Final Projection))**.** We define $\mathsf{Conv1x1}_{\Theta_{\mathrm{out}}} \coloneqq \mathsf{Conv2d}^{(1,1;\ 1,1;\ 0,0)}_{\Theta_{\mathrm{out}}}$ and use it for the mapping to RGB output $\mathbb{R}^{3\times H\times W}$.

7 Why U-Net Reaches the Entire Area "Shallowly": Quantitative Comparison of Receptive Fields

7.1 Receptive Field of a Pure CNN (Conv2d only)

When L layers of Conv2d with kernel size $k = 3$, stride 1, and padding 1 are stacked, the **receptive field** in one dimension is

$$R_{\mathrm{pure}}(L) = 1 + (k-1)L = 1 + 2L. \tag{79}$$

The condition to reach the entire width W is $R_{\mathrm{pure}}(L) \ge W$, i.e.,

$$L \ \ge\ \frac{W-1}{2}. \tag{80}$$

7.2 Receptive Field of U-Net (with staged Downsample2D)

Performing Downsample2D$^{(2)}$ with stride 2 L times at each level, and performing n_ℓ 3×3 Conv2d (stride 1) at each resolution, one step at the final (coarsest) level corresponds to 2^L pixels in the original resolution. Therefore, the receptive field converted to the original resolution is

$$R_{\mathrm{unet}} \ =\ 1\ +\ \sum_{\ell=0}^{L} (2^\ell) \cdot (k-1)\, n_\ell \ =\ 1\ +\ 2\sum_{\ell=0}^{L} 2^\ell n_\ell. \tag{81}$$

If we uniformly set $n_\ell = n$,

$$R_{\text{unet}} = 1 + 2n\,(2^{L+1} - 1). \tag{82}$$

Theorem 7.1 (U-Net reaches the entire area with $O(\log W)$ depth)**.** Assuming $k = 3$ and $n \geq 1$ layers at each level, the sufficient condition $R_{\text{unet}} \geq W$ to reach the entire width W is

$$L \;\geq\; \log_2\!\left(\frac{W-1}{2n} + 1\right) - 1. \tag{83}$$

Therefore, the required number of levels L is $O(\log W)$, which is **significantly fewer layers** to express dependencies from end to end compared to the $O(W)$ of a pure CNN (Conv2d only) in (80).

Proof. Substituting (82) into $R_{\text{unet}} \geq W$, we get $1 + 2n(2^{L+1} - 1) \geq W \iff 2^{L+1} \geq \frac{W-1}{2n} + 1$. Taking $\log_2$, we get $L + 1 \geq \log_2(\frac{W-1}{2n} + 1)$, which yields (83). □

8 Full Definition of U-Net and VAE Decoder "as Functions"

8.1 The U-Net (Conditional) Overall Function

Definition 8.1 (Parametric Function of UNet2DConditionModel)**.** The inputs are latent $\underline{\boldsymbol{Z}} \in \mathbb{R}^{C\times H\times W}$, time $t \in \mathbb{R}$, and text embedding sequence $\boldsymbol{C} \in \mathbb{R}^{M\times d_{\text{ctx}}}$. The learnable parameter vector is

$$\boldsymbol{\Theta}_{\text{U}} = \Big(\Theta_{\text{TE}},\ \{\Theta_\ell^{\downarrow}\}_{\ell=1}^{L},\ \Theta^{\text{mid}},\ \{\Theta_\ell^{\uparrow}\}_{\ell=1}^{L},\ \Theta^{\text{out}}\Big) \tag{84}$$

(where each Θ is the concatenation of all coefficients from §6 for Conv2d, GroupNorm, SiLU, Linear, ScaledDotProductAttention, MultiheadAttention, TimestepEmbedding, ResnetBlock2D, Downsample2D, Upsample2D, DownBlock2D, UpBlock2D, CrossAttentionMidBlock2D, Conv1x1). Injecting the time embedding $\boldsymbol{h} = \text{TimestepEmbedding}_{\Theta_{\text{TE}}}(t)$ into each residual block,

$$\underline{\boldsymbol{D}}_0 = \underline{\boldsymbol{Z}}, \tag{85}$$

$$\underline{\boldsymbol{D}}_\ell = \text{(CrossAttention)DownBlock2D}_{\Theta_\ell^{\downarrow}}\big(\underline{\boldsymbol{D}}_{\ell-1},\ \boldsymbol{h}, \boldsymbol{C}\big), \quad \ell = 1, \ldots, L, \tag{86}$$

$$\underline{\boldsymbol{B}} = \text{CrossAttentionMidBlock2D}_{\Theta^{\text{mid}}}\big(\underline{\boldsymbol{D}}_L,\ \boldsymbol{h},\ \boldsymbol{C}\big), \tag{87}$$

$$\underline{\boldsymbol{U}}_L = \text{UpBlock2D}_{\Theta_L^{\uparrow}}\big(\underline{\boldsymbol{B}},\ \underline{\boldsymbol{D}}_L,\ \boldsymbol{h}\big), \tag{88}$$

$$\underline{\boldsymbol{U}}_{\ell-1} = \text{(CrossAttention)UpBlock2D}_{\Theta_{\ell-1}^{\uparrow}}\big(\underline{\boldsymbol{U}}_\ell,\ \underline{\boldsymbol{D}}_{\ell-1},\ \boldsymbol{h}, \boldsymbol{C}\big), \quad \ell = L, \ldots, 1, \tag{89}$$

$$\underline{\hat{\boldsymbol{E}}} = \text{Conv1x1}_{\Theta^{\text{out}}}(\underline{\boldsymbol{U}}_0) \in \mathbb{R}^{C\times H\times W}, \tag{90}$$

$$\underline{\hat{\boldsymbol{E}}} = \text{UNet2DConditionModel}_{\boldsymbol{\Theta}_{\text{U}}}\big(\underline{\boldsymbol{Z}},\ t,\ \boldsymbol{C}\big). \tag{91}$$

[a]

[a] Diffusers UNet2DConditionModel (U-Net Conditional): `https://huggingface.co/docs/diffusers/api/`

`models/unet2d`.

Remark 8.1. Each block is a composition of the basic operations from §6, and the output spatial size continuously follows $H \times W$ (**variable input/output sizes**).

8.2 The VAE Decoder Overall Function

Stable Diffusion series VAEs have a decoder structure (Decoder) included in AutoencoderKL. Here, we formulate the **deterministic architecture** of the decoder as a composition of ResnetBlock2D and Upsample2D.

Definition 8.2 (Decoder (VAE Decoder))**.** For input latent $\underline{\boldsymbol{Z}} \in \mathbb{R}^{C_z \times H_z \times W_z}$, using learnable parameters

$$\boldsymbol{\Theta}_{\mathrm{Dec}} = \left(\Theta^{\mathrm{in}},\ \{\Theta_\ell^{\uparrow}\}_{\ell=1}^{L_d},\ \Theta^{\mathrm{out}}\right) \tag{92}$$

$$\underline{\boldsymbol{H}}_0 = \text{ResnetBlock2D}_{\Theta^{\mathrm{in}}}(\underline{\boldsymbol{Z}}, \mathbf{0}) \quad \text{(no time dependence, so } \boldsymbol{h} = \mathbf{0}\text{)}, \tag{93}$$

$$\underline{\boldsymbol{H}}_\ell = \text{ResnetBlock2D}_{\Theta_\ell^{\uparrow}}\big(\text{Upsample2D}^{(2)}_{\Theta^{(\ell)}_{\mathrm{Up}}}(\underline{\boldsymbol{H}}_{\ell-1}),\ \mathbf{0}\big), \quad \ell = 1, \ldots, L_d, \tag{94}$$

$$\hat{\underline{\boldsymbol{X}}} = \text{Conv1x1}_{\Theta^{\mathrm{out}}}(\underline{\boldsymbol{H}}_{L_d}) \in \mathbb{R}^{3 \times H \times W}, \quad (H = 2^{L_d} H_z,\ W = 2^{L_d} W_z). \tag{95}$$

Definition 8.3 (AutoencoderKL (Decoder part))**.** We define the decoder mapping $\mathcal{D}$ of AutoencoderKL as

$$\mathcal{D}_{\boldsymbol{\Theta}_{\mathrm{Dec}}} : \ \mathbb{R}^{C_z \times H_z \times W_z} \longrightarrow \mathbb{R}^{3 \times (2^{L_d} H_z) \times (2^{L_d} W_z)}, \quad \mathcal{D}_{\boldsymbol{\Theta}_{\mathrm{Dec}}}(\underline{\boldsymbol{Z}}) = \text{Decoder}_{\boldsymbol{\Theta}_{\mathrm{Dec}}}(\underline{\boldsymbol{Z}}) \tag{96}$$

[a]

[a] Diffusers AutoencoderKL implementation (includes Decoder): `https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/autoencoder_kl.py`, API: `https://huggingface.co/docs/diffusers/api/models/autoencoderkl`.

9 Summary (Correspondence to Learning Outcomes)

Correspondence with Learning Outcomes

- **Mathematical description of architectures**: We **formally defined as functions** the U-Net and VAE decoder.
- **Explanation of variable I/O sizes**: We confirmed that U-Net/VAE satisfy variable input/output sizes because each layer is defined in a **form independent of spatial size**.
- **Explanation of differences from the proposal**: We clarified the configuration of the U-Net in image generation AIs is different from the originally proposed form.

10 Next Lecture Preview

Next time, we will explain the **Text Encoder**.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

[3] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In International Conference on Learning Representations (ICLR), 2014.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[6] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations (ICLR), 2021.

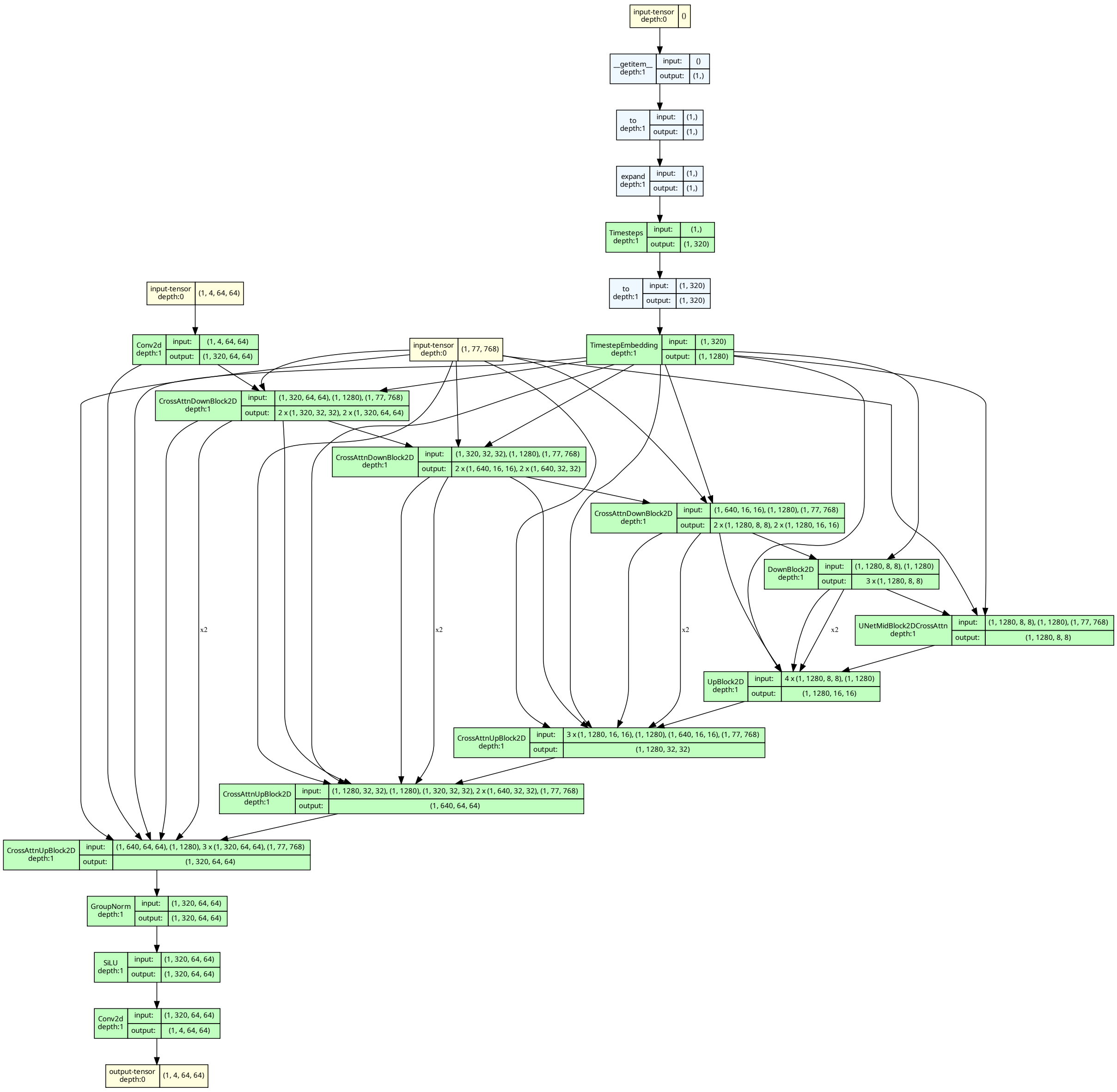

Figure 1: The neural network architecture of the U-Net in Stable Diffusion 1.5. The tensor size corresponds to the input with: batch size = 1, the number of channels = 4, width = 64, height = 64.

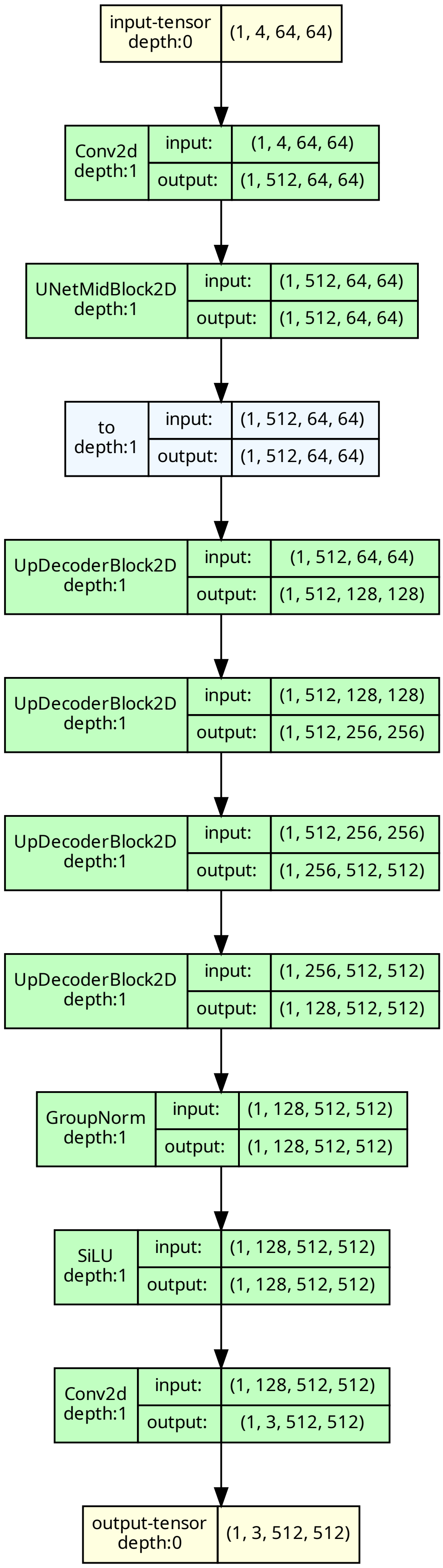

Figure 2: The neural network architecture of the VAE Decoder in Stable Diffusion 1.5. The tensor size corresponds to the input with: batch size = 1, the number of channels = 4, width = 64, height = 64.

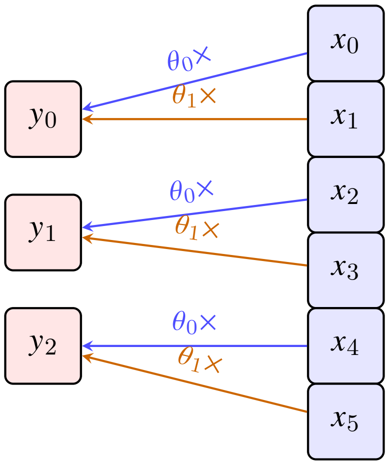

Figure 3: 1D **Downsample1D** (Example: input length 6, output length 3, filter width 2, stride 2). Identical weights θ_0, θ_1 are shown by edges of the same color.

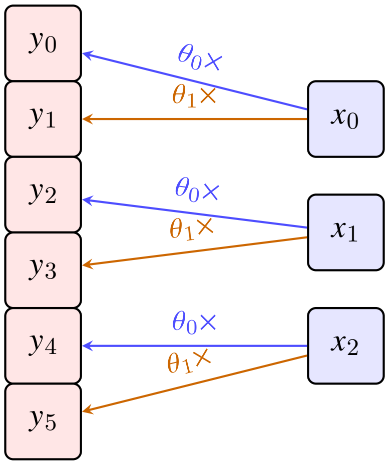

Figure 4: 1D **Upsample1D** (Example: input length 3, output length 6, filter width 2, stride 2). Identical weights θ_0, θ_1 are shown by edges of the same color.