ControlNet

状态：ICCV2023

单位：斯坦福大学

文章链接：https://arxiv.org/abs/2302.05543

Github 链接：未开源

目录

Abstract

本文推出了 ControlNet，一种向大型预训练文生图扩散模型中加入控制条件的神经网络模块。ControlNet 冻结了预训练的扩散模型，并复用了其深且强效的编码层，经过十亿级别的训练来使其学习如何加入控制条件。这些结构通过“零卷积层”链接，这是一个从零开始逐步学习的卷积层，避免了有害噪声影响结果。本文尝试了多种控制条件，如边界图，深度图，分割图，姿态图等。使用的预训练模型是 Stable Diffusion，能够使用单一或多种控制，以及尝试了是否加入提示词。我们展示了 ControlNet 在小数据集（<50k）和大数据集（>1m）上的训练结果，展现了 ControlNet 在给扩散模型加入控制的广泛运用前景。

Introduction

文生图模型对我们绘制想要的图像有很大的帮助，但使用文本控制一些空间信息，具体细节还是十分困难的。如果能够使用一张图像作为参考，再进行相对应的生成，用户体验将会好很多。

这种使用额外的图像作为参考，一般被认为是以其作为一种控制条件，而控制条件的种类也有许多，如深度图，描边图，手绘稿。这就需要一种可以在训练中学会利用不同种类控制条件的方案了。

但控制条件毕竟是个小问题，其数据集大小远小于预训练模型使用的数据集，如果直接进行微调或者继续训练，效果都不会很好。

本文提出了 ControlNet，一种神经网络结构，可以对预训练的扩散模型进行高效地微调。其将预训练模型冻结，而复制出一个可训练的副本，作为旁路。其中使用“零卷积层”链接，保证了开始训练时不存在有害噪声的干扰。

实验则尝试了多种控制方案，包括深度，描边，等等，且尝试了是否使用提示词的不同方案，以及单一或多种的控制方案。证明了这一方案的有效性。

并于当前流行的几种加入控制方案进行了横向对比，验证了这一方案的优势。

相关工作

微调神经网络

微调神经网络的最简单方案就是用额外的数据进行进一步训练，但这 有可能会造成过拟合，模式坍塌，灾难性遗忘等问题，因此研究者提出了许多微调方案来避免这些问题。

HyperNetwork

起源于 NLP 领域的微调方案。尝试训练一个小的神经网络来影响大网络的结果，曾被用于调节 SD 的生成风格。

Adapter

通过在神经网络中增加额外层的方式来调整网络，同样起源于 NLP，常用在 Transformer 模型中。

Additive Learning

通过冻结原有的预训练模型，而增加额外参数（旁路）进行微调的方案。

LoRA

通过引入低秩矩阵的方式进行微调。

零初始化层

这一方案被应用在本文中连接各个模块。对于初始化参数的讨论一直都存在，比如有研究者发现高斯初始化比全零初始化更加稳定。而对于 diffusion 的初始化，不少工作都提及了使用零初始化的卷积层效果更佳。

图扩散模型

文生图扩散模型

从 DDPM 被提出，LDM 引入了隐空间机制，加快了运算速率。且有诸多相关的工作。

额外控制的扩散模型

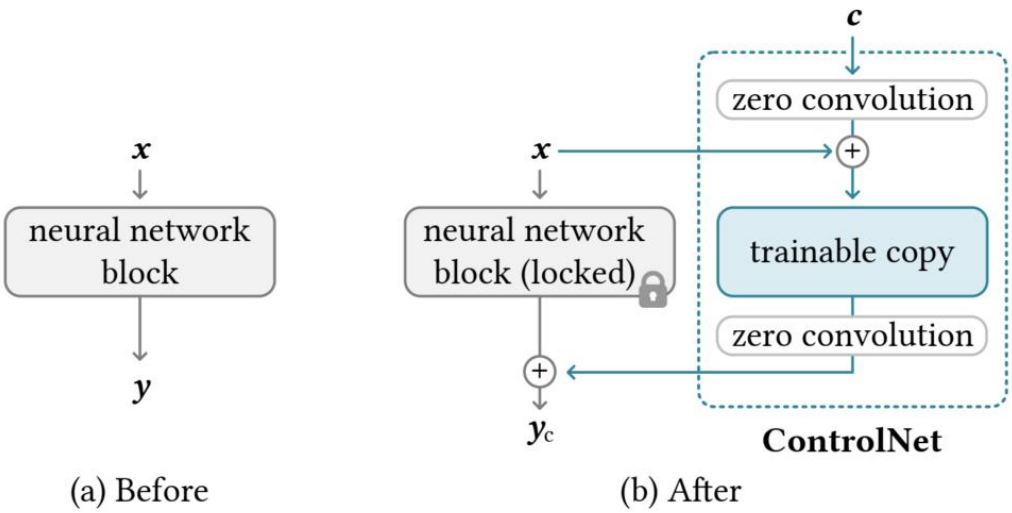
具有额外控制的扩散模型在定制化运用中十分有效，许多工作也提出了多种不同的加入控制条件的方案。

图像翻译

有条件的 GAN 和 transfomer 都可以做到构建两个不同领域照片间的联系。Diffusion 也行，但需要微调。

实现方案

ControlNet

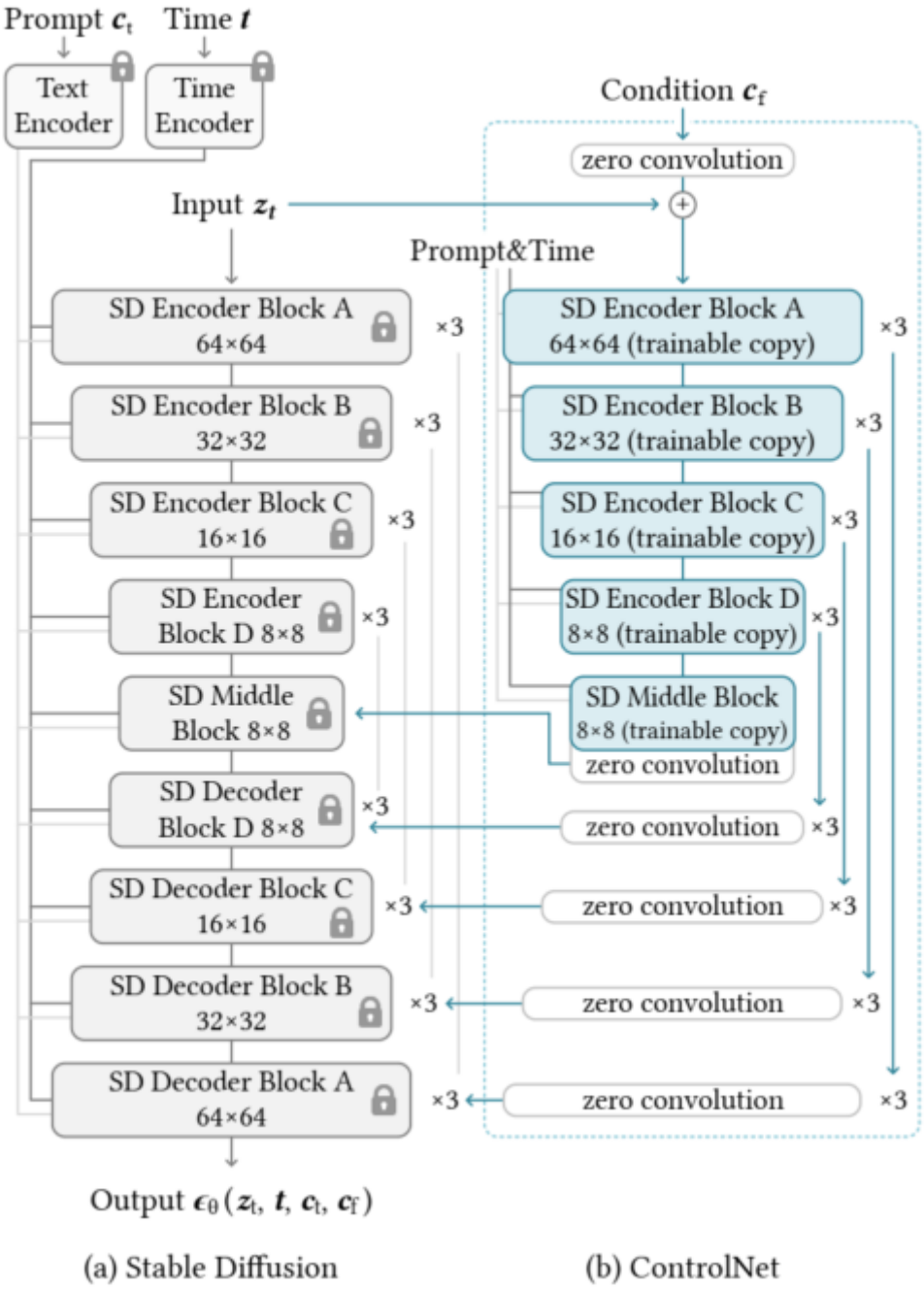


对于一个完整的神经网络，其输入 x，输出为 y。加入一个 ContrNet 即将原本的模型冻结，而复制一个可以训练的副本。副本接受额外的输入，且连接处在最初都为零卷积层（保证训练开始时加入的部分完全不影响输出）。

在使用时则将旁路的输出结合到原本的输出 y 上，形成最后的输出 yc。

以 SD 为例

本部分以 SD 为例，详细介绍了其结构以及加入一个 ControlNet 的方法。外加一个 ControlNet 只会额外增加 23%的 GPU 显存，以及 34%的训练时间。



训练

训练的 loss 与常规的 SD 训练相同。

训练时，50%的情况保持 prompts 输入为空，这是为了进一步强化 condition 的用途。

且训练过程中注意到，ControlNet 对于额外控制信息的学习不是平滑的，而是会在某一阶段骤然发生的。

推理

CFG

CFG 策略用于帮助模型取得高质量的输出。这原本是用于控制语义条件强度的，公式为：

$$\epsilon_{prd} = \epsilon_{uc} + \beta_{cfg}(\epsilon_c - \epsilon_{uc})$$

其中ϵ_{prd}为最终输出，ϵ_{uc}为无控制输出（此处的控制指 prompt），ϵ_c为有控制输出，β_{cfg}为 CFG 参数。

而本文中，需要考虑的问题是 controlnet 的输入应该如何加入ϵ_{uc}和ϵ_c中。通过实验展示，如果同时加入，则无 prompt 时，CFG 的指导毫无效果，如果只输入到ϵ_c中，又会使得控制条件过强。

本文采取的方案为，为了减少控制强度，而在 block 上进行不断衰减的控制，比如在第一个 block 加入四倍的控制，第二块 2 倍，第三块 1 倍，依次类推。

多条件控制

同时加入多个控制条件时，直接将对于的 ControlNet 的输出相加即可，不需要额外模块。

实验

定性分析

用结果图定性展示了多种不同的控制条件，在 ControlNet 上都十分有效。

消融实验

进行了四组实验的对比，得出了使用 copy 结构和零卷积层的有效性。

定量分析

用户实验

首先，通过用户评分将多种草图生成完整图片模型进行对比，本文的方案效果最佳。

然后，对于深度图生成图像，与工业化的 SDv2-D2I 模型对比，用户实验的结果表明二者的输出结果几乎难以区分，证明两种方案的输出大致相同，且本方案的数据需求量，训练计算量和训练时长都远远更小。

对比实验

和相关的工作对比了效果。

对比之前方案

通过展示图像和之前的方案进行的定性的对比。

讨论

训练数据集只有 1K 的时候就以及有较为优秀的输出，而继续征打数据集，效果会更好。

当输入的 condition 并不明确时，会尽量模仿其形状。

且由于 Controlnet 并没有改变 SD 的结构，其可以与用到多种不同的预训练扩散模型中去。

结论

简单总结本文。