ControlNet

Links

https://learnopencv.com/controlnet/#How-ControlNet-Works? - Кратко
https://arxiv.org/abs/2302.05543 - Статья
https://lllyasviel.github.io/misc/202309/cnet_supp.pdf - supplementary materials.

Проблема

The largest datasets for various specific problems (e.g., object shape/normal, human pose extraction, etc.) are usually about 100K in size, which is 50,000 times smaller than the LAION-5B [79] dataset

The direct finetuning or continued training of a large pretrained model with limited data may cause overfitting and catastrophic forgetting

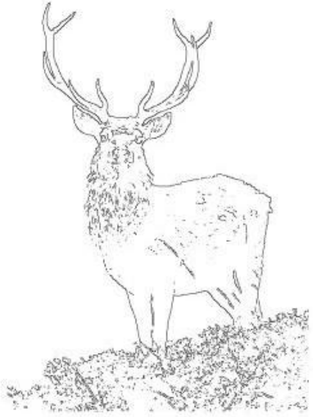
Общая идея

- разработать стратегию fine-tune для больших моделей не требующую огромных выч. мощ.
- как можно более четко управлять доменом

Решение:

- end-to-end архитектура для обучения моделей специфическим задачам
- скопируем веса (не все) и будем учить не их, а их копии

Source image
(for canny edge detection)



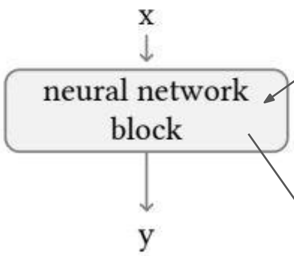
Canny edge (input)

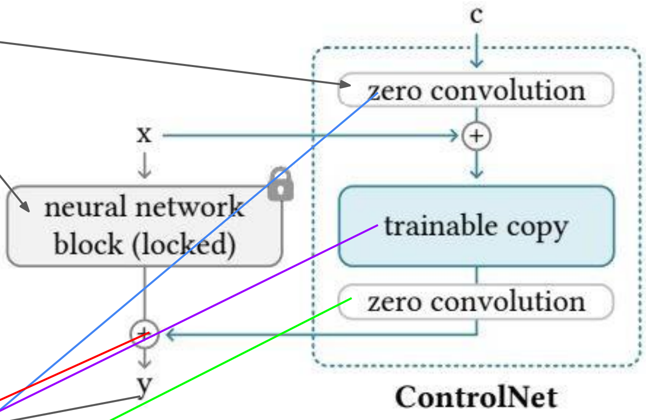
Generated images (output)

Описание метода

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \Theta)$$

$$\mathcal{Z}(\cdot;\cdot)$$





$$\boldsymbol{y}_{\mathrm{c}} = \mathcal{F}(\boldsymbol{x}; \Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{\mathrm{z1}}); \Theta_{\mathrm{c}}); \Theta_{\mathrm{z2}})$$

where non-zero gradients are obtained and the neural network begins to learn. In this way, the zero convolutions become a unique type of connection layer that progressively grows parameters from zero to optimized values in a learned way.

Gradient Calculation of Zero Convolution Layers

$$\mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i} = \boldsymbol{B}_i + \sum_{j}^{c} \boldsymbol{I}_{p,j}\boldsymbol{W}_{i,j}.$$

$$\begin{cases} \mathcal{Z}(\boldsymbol{c};\Theta_{\mathrm{z1}}) = \mathbf{0} \\ \mathcal{F}(\boldsymbol{x}+\mathcal{Z}(\boldsymbol{c};\Theta_{\mathrm{z1}});\Theta_{\mathrm{c}}) = \mathcal{F}(\boldsymbol{x};\Theta_{\mathrm{c}}) = \mathcal{F}(\boldsymbol{x};\Theta) \\ \mathcal{Z}(\mathcal{F}(\boldsymbol{x}+\mathcal{Z}(\boldsymbol{c};\Theta_{\mathrm{z1}});\Theta_{\mathrm{c}});\Theta_{\mathrm{z2}}) = \mathcal{Z}(\mathcal{F}(\boldsymbol{x};\Theta_{\mathrm{c}});\Theta_{\mathrm{z2}}) = \mathbf{0} \end{cases}$$

$$\boldsymbol{y}_{\mathrm{c}} = \boldsymbol{y}$$

$$\mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i} = \boldsymbol{B}_i + \sum_{j}^{c} \boldsymbol{I}_{p,i}\boldsymbol{W}_{i,j}$$

$$\begin{cases} \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{B}_i} = 1 \\ \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{I}_{p,i}} = \sum_{j}^{c} \boldsymbol{W}_{i,j} = 0 \\ \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{W}_{i,j}} = \boldsymbol{I}_{p,i} \neq \mathbf{0} \end{cases}$$

loss

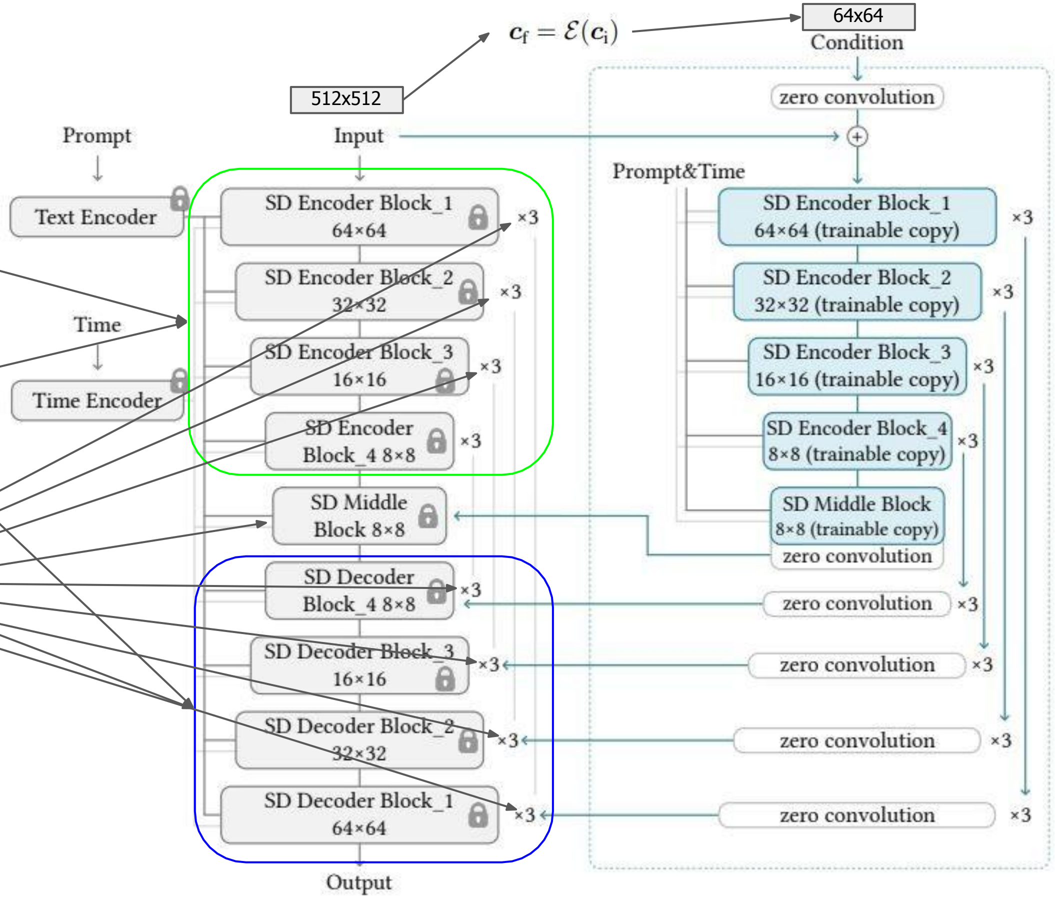
$$\boldsymbol{W}^* = \boldsymbol{W} - \beta_{\mathrm{lr}} \cdot \frac{\partial \mathcal{L}}{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})} \odot \frac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})}{\partial \boldsymbol{W}} \neq \mathbf{0}$$

$$\frac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W}^*,\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{I}_{p,i}} = \sum_{i}^{c} \boldsymbol{W}^*_{i,j} \neq \mathbf{0}$$

Hadamard product

Control Net + SD

- Encoder: 12 blocks
- Decoder: 12 blocks
- One middle block
- 4 down-sampling
- 4 upsampling
- 17 blocks res-net + 2 ViT
- Textencoder: Clip



(a) Stable Diffusion
(b) ControlNet

Обучение

z_0 начальное изображение

z_t зашумленное изображение на последнем шаге

t кол-во шагов зашумления

c_t текстовый промт

c_f наше условие которому учим

ϵ_θ семейство зашумляющих сетей

лосс функция

$$\mathcal{L} = \mathbb{E}_{z_0, t, c_t, c_f, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(z_t, t, c_t, c_f)) \|_2^2 \right]$$

Детали обучения:

- рандомно меняем промты на пустые
- Small-Scale Training
  - “SD Middle Block” and “SD Decoder Block 1,2,3,4”
  - убрать skip-connect из D увеличит скорость в 1.6 (потом можно вернуть и дотюнить)
- Large-Scale Training
  - если много GPU
  - если много данных (пару лимонов)
  - можно учить ControlNet 50k итераций, а потом разлочить и учить всю диффузию

Ablation Study Zero Convolution

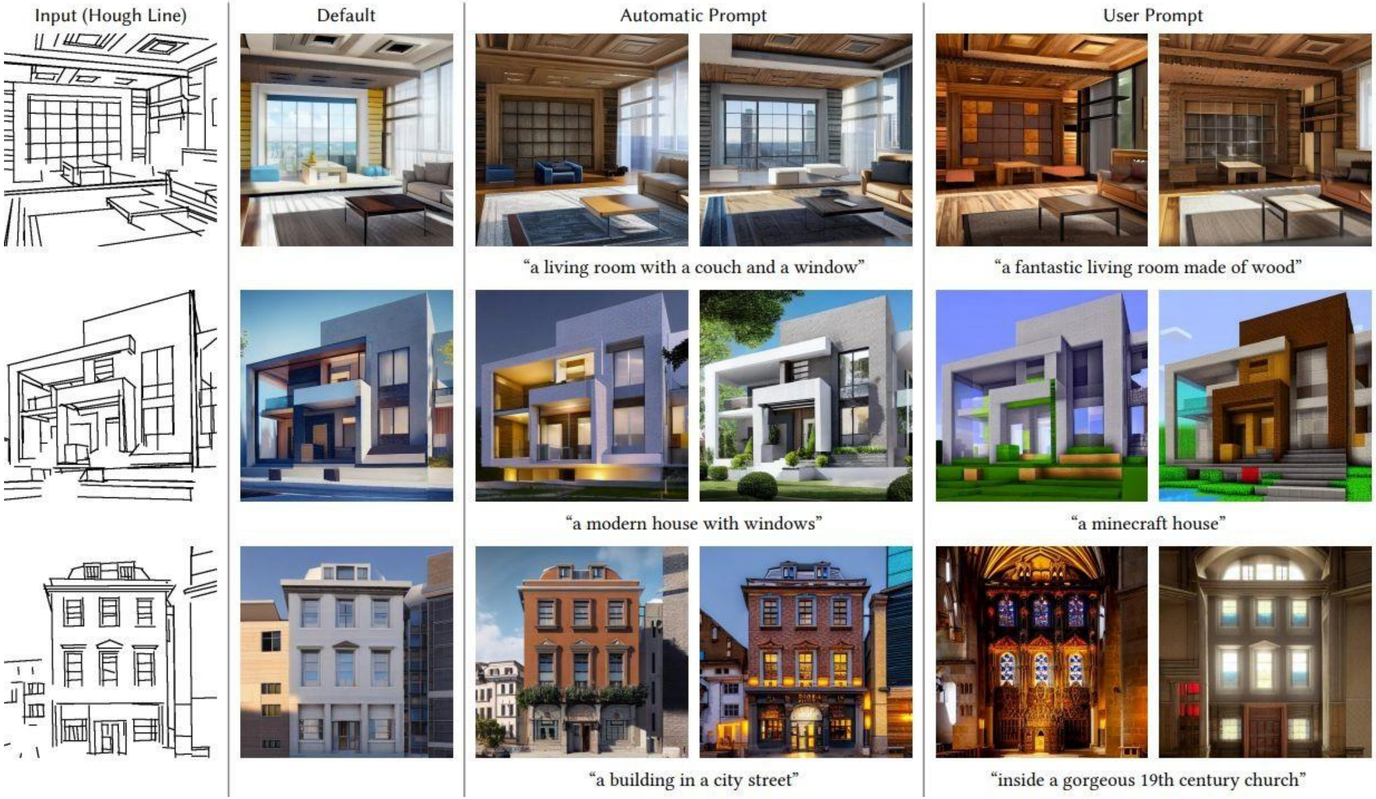
Результаты

- имеем подход позволяющий сократить ресурсы при дообучении
- можно брать небольшой датасет и не оверфитнуться!
- бери доп. условие которое надо и тюнь диффузию
- на 23% нужно больше GPU и на 34% больше времени на 1 итерацию по сравнению со SD

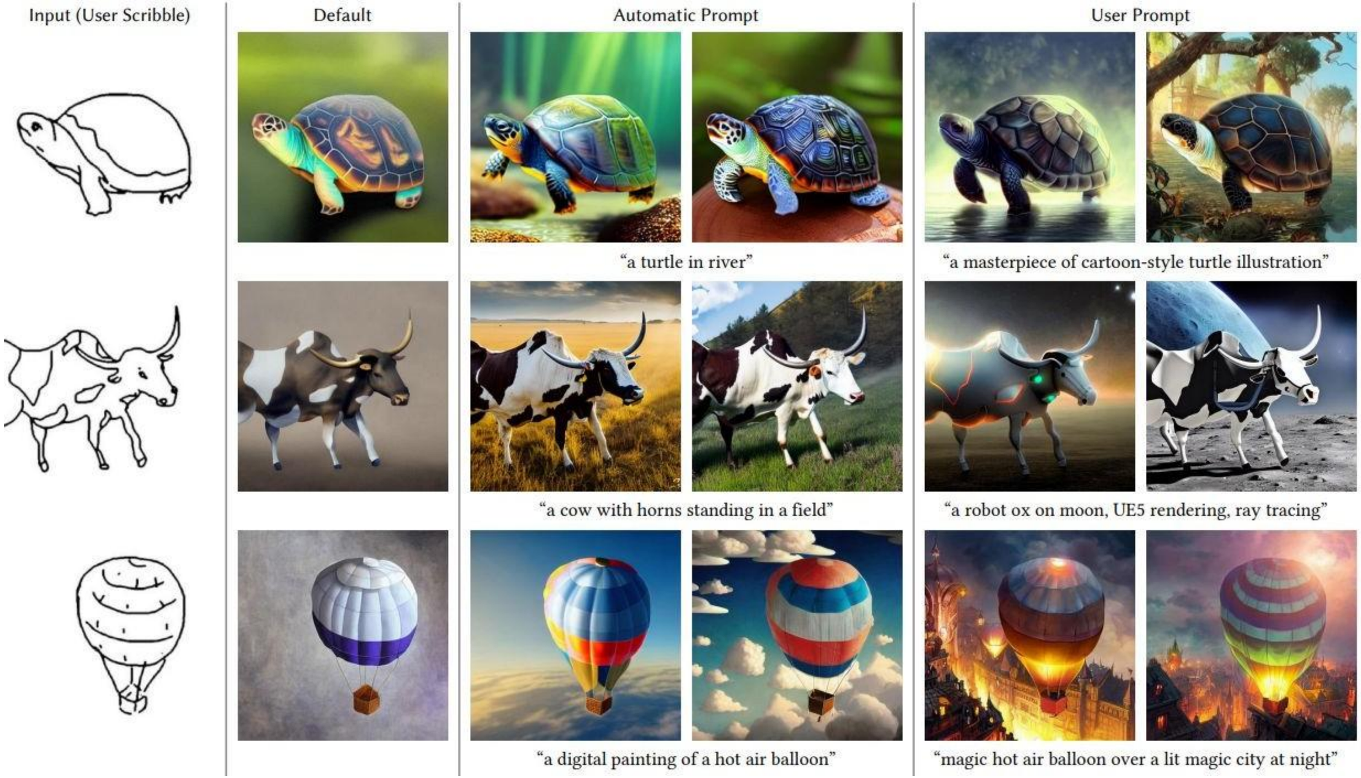
Canny Edge

Input (Canny Edge)
Default
Automatic Prompt
User Prompt
"a man with beard sitting with two children"
"mother and two boys in a room, masterpiece, artwork"
"a man in a suit and tie"
"a man in a white suit and tie"

Hough lines


Input (Hough Line)
Default
Automatic Prompt
User Prompt
"a living room with a couch and a window"
"a fantastic living room made of wood"
"a modern house with windows"
"a minecraft house"
"a building in a city street"
"inside a gorgeous 19th century church"

Human scribbles



Controlling Stable Diffusion with Openpose