Classifier Free Guidance

MIT Diffusion Models Lecture

Made with Goodnotes

Classifier-Free Guidance

- Goal: Generate stuff with prompts.

How do we make models fit to prompts.

* Vanilla guidance → Vector set

Data dist: $(z, y) \sim p_{data}$

$z \in \mathbb{R}^d$

pairs of images & prompts

N.N

$u^\theta : \mathbb{R}^d \times \mathcal{Y} \times [0,1] \to \mathbb{R}^d$

$(x, y, t) \to u_t^\theta(x|y)$

$x_0 \sim p_{init}$

$dx_t = u_t^\theta(x_t|y)dt + \sigma_t dw_t$

$x_1 \sim p_1(\cdot|y)$

How would we change our vector field?

Flow models: $u_t^\theta(x)$ → Make it prompts dependent

In our case

guided by prompts

guided vector field: $u_t^\theta(x|y) \in \mathbb{R}^d$

How to train this? As before

Made with Goodnotes

$z \sim p_{data}(\cdot|y)$ $y \to$ "original" $z \to$ image of that

Vector field

Conditional V.F.

Prompt as input

* Guided FM Loss

$$L_{CFM}(\theta) = \mathbb{E}_{t,(z,y)\sim p_{data}(z,y)} \left\| u_t^\theta(x|y) - u_t^{target}(x|z) \right\|^2$$

Compare to eqn 26 in notes

$t=0 \to$ Noise

$t=1 \to p_{data}(\cdot|y)$

Pairs we are sampling

(Show Vanilla guided sampling)

Same way as before except we are sampling a prompt & plugging it into the ODE at every step.

But we get suboptimal results

* How do we make images really fit to the prompts?

Artificially reinforce

1. Model might underfit
2. text-image pairs may have noise

Made with Goodnotes

* Technique: Classifier Free Guidance

But first we talk about classifier guidence

* Classifier guidance

Bayes Rule

$$P_t(x|y) = \frac{P_t(x)\, P_t(y|x)}{P_t(y)}$$

$P_t(x|y)$: Posterior; $P_t(x)$: Prior; $P_t(y|x)$: likelihood; $P_t(y)$: Marginal

$$\Rightarrow \log P_t(x|y) = \log P_t(y|x) + \log P_t(x) - \log P_t(y)$$

Applying gradient of log likelihood (the score)

$$\nabla_x \log P_t(x|y) = \nabla_x \log P_t(y|x) + \nabla_x \log P_t(x) \quad (1)$$

The guided score is given by R.H.S.

Made with Goodnotes

By using the conversion formula (Eqn 41 & ...) & substituting (1)

$$u_t^{target}(x|y) = b_t x + a_t \nabla_x \log p_t(x) + a_t \nabla_x \log p_t(y|x)$$

$$u_t^{target}(x|y) = u_t^{target}(x) + a_t \nabla_x \log p_t(y|x)$$

- $u_t^{target}(x|y)$ → Guided V.F
- $u_t^{target}(x)$ → Not dependent on prompt (direction)
- $\nabla_x \log p_t(y|x)$ → Classifier; Prompt dependent

$Y \rightarrow$ Label

$X \rightarrow$ Data (Image)

So we scale up the part that is prompt dependent. → Show Slides

* Idea: Reinforce Classifier

Weight : $W \geq 1$

Made with Goodnotes

$$\tilde{u}_t^w(x|y) = u_t^{target}(x) + w a_t \nabla \log p_t(y|x)$$

Train the classifier for guidence.

& add that up

scale up contribution

Issues :- 1) We have to train a classifier along with the flow/diffusion model.

2) If y is high dimensional, $p_t(y|x)$ may be very hard to learn & $\nabla \log p_t(y|x)$ is hard to obtain

So Classifier free guidence

* Classifier-free guidance (CFG)

$$\tilde{u}_t^w(x|y) = u_t^{target}(x) + w a_t \nabla_x \log p_t(y|x)$$

$$= u_t^{target}(x) + w a_t (\nabla_x \log p_t(x|y) - \nabla_x \log p_t(x))$$

$$= u_t^{target}(x) - (w b_t x + w a_t \nabla_x \log p_t(x)) + (w b_t x + w a_t \nabla_x \log p_t(x|y))$$

$$= u_t^{target}(x) + w(u_t^{target}(x|y) - u_t^{target}(x))$$

All of these are vector fields.

$$= w\, u_t^{target}(x|y) + (1-w)\, u_t^{target}(x)$$

$u_t^{target}(x|y)$: guided V.F. $(1-w)\, u_t^{target}(x)$: unguided V.F.

$w > 1$

There are still 2 models

Made with Goodnotes

Empty token ϕ : No prompts

Missing prompts

$$\tilde{u}_t^w(x|y) = w\, u_t^\theta(x|y) + (1-w)\, u_t^\theta(x|\phi)$$

We do the strides up through the sampling procedure.

* CFG is a heuristic!

We need 2 N.N. calls

1 for conditional

1 for unconditional

Made with Goodnotes