7/28 meeting

應名宥

compare with author (same CLIP model)


```
{"fid50k_full": 331.64653125288464}
{"fid50k_full": 201.39067457904645}
{"fid50k_full": 112.8136538562812},
{"fid50k_full": 92.62477145756523},
{"fid50k_full": 74.2568278747297},
{"fid50k_full": 65.26074268783911},
{"fid50k_full": 51.94230806043393},
{"fid50k_full": 50.06230557365797},
{"fid50k_full": 44.795518329704954}
{"fid50k_full": 40.77542432455253},
{"fid50k_full": 37.410084185962575}
{"fid50k_full": 35.47804069887169},
{"fid50k_full": 33.1603153645175},
```


original


```
{"fid50k_full": 308.6027637342727},
{"fid50k_full": 188.35161891464634},
{"fid50k_full": 102.0666766370245},
{"fid50k_full": 74.12198189088315},
{"fid50k_full": 56.47082054678368},
{"fid50k_full": 46.36170454926968},
{"fid50k_full": 39.297681694553624},
{"fid50k_full": 34.63167871020097},
{"fid50k_full": 30.801296149502548},
{"fid50k_full": 27.542499252071813},
{"fid50k_full": 25.989908475363904},
{"fid50k_full": 24.2476226967577},
{"fid50k_full": 22.25654370469862},
```


modified

during this time

{"fid50k_full": 312.4110504526001},
{"fid50k_full": 176.4415644158251},
{"fid50k_full": 104.90489318428916},
{"fid50k_full": 77.09823413780748},
{"fid50k_full": 63.81999294057303},
{"fid50k_full": 50.977956916721006},
{"fid50k_full": 44.66246608697162},
{"fid50k_full": 38.30186529643082},
{"fid50k_full": 33.58820084766524},
{"fid50k_full": 30.387231610042633},
{"fid50k_full": 28.594780898470873},
{"fid50k_full": 26.266270740788453},
{"fid50k_full": 25.097551563505984},

original

{"fid50k_full": 308.6027637342727},
{"fid50k_full": 188.35161891464634},
{"fid50k_full": 102.0666766370245},
{"fid50k_full": 74.12198189088315},
{"fid50k_full": 56.47082054678368},
{"fid50k_full": 46.36170454926968},
{"fid50k_full": 39.297681694553624},
{"fid50k_full": 34.63167871020097},
{"fid50k_full": 30.801296149502548},
{"fid50k_full": 27.542499252071813},
{"fid50k_full": 25.989908475363904},
{"fid50k_full": 24.2476226967577},
{"fid50k_full": 22.25654370469862},

modified

experiments

- symmetric loss structure (o)
- large network (o)
- different generator layer (x)
- clip mask contrastive structure (o)
- shallow mapping network (x)
- iird loss (o)
- iic loss (o)
- disable style mixing regularization (o)
- Low-Dimensional Latent Space (?)
- delayed path length regularization (?)

symmetric loss structure (o)


```
{"fid50k_full": 302.22335154211953},
{"fid50k_full": 195.7976188834841},
{"fid50k_full": 111.77636887334621},
{"fid50k_full": 81.34089919482533},
{"fid50k_full": 63.52439127743475},
{"fid50k_full": 53.91829141415987},
{"fid50k_full": 48.72026616964483},
{"fid50k_full": 41.74788527424883},
```


asymmetric


```
{"fid50k_full": 292.9706606736117 5}
{"fid50k_full": 169.1666946773363 5}
{"fid50k_full": 104.1743476125651 1}
{"fid50k_full": 77.2120401283815 2},
{"fid50k_full": 61.9587927098792 3},
{"fid50k_full": 50.0096636223623 4}
{"fid50k_full": 43.0880771668727 2},
{"fid50k_full": 37.9046400555076 5}
```


symmetric

large network (feature dim) (o)

{"fid50k_full": 312.4110504526001},
{"fid50k_full": 176.4415644158251},
{"fid50k_full": 104.90489318428916},
{"fid50k_full": 77.09823413780748},
{"fid50k_full": 63.81999294057303},
{"fid50k_full": 50.977956916721006},
{"fid50k_full": 44.66246608697162},
{"fid50k_full": 38.30186529643082},
{"fid50k_full": 33.58820084766524},
{"fid50k_full": 30.387231610042633},
{"fid50k_full": 28.594780898470873},
{"fid50k_full": 26.266270740788453},
{"fid50k_full": 25.097551563505984},

base

{"fid50k_full": 292.97066067361175},
{"fid50k_full": 169.16669467733635},
{"fid50k_full": 104.17434761256511},
{"fid50k_full": 77.21204012838152},
{"fid50k_full": 61.95879270987923},
{"fid50k_full": 50.009663622336234},
{"fid50k_full": 43.08807716687272},
{"fid50k_full": 37.904640055550765},
{"fid50k_full": 34.47229400106767},
{"fid50k_full": 29.94065800860855},
{"fid50k_full": 26.723918326413308},
{"fid50k_full": 25.046729043965996},
{"fid50k_full": 23.50986657022635},

large

different generator layer (x)

{"fid50k_full": 292.97066067361175}
{"fid50k_full": 169.16669467733635}
{"fid50k_full": 104.17434761256511}
{"fid50k_full": 77.21204012838152},
{"fid50k_full": 61.95879270987923},
{"fid50k_full": 50.00966362236234}
{"fid50k_full": 43.08807716687272},
{"fid50k_full": 37.90464005550765}

base

{"fid50k_full": 330.0083577931281},
{"fid50k_full": 169.7254208405388},
{"fid50k_full": 115.6208002052518}
{"fid50k_full": 89.0938202116907},
{"fid50k_full": 71.42279010353275},
{"fid50k_full": 61.5021694331095},
{"fid50k_full": 49.85897443889061},
{"fid50k_full": 47.95638221066022},

modified

CLIP mask (o)


```
{"fid50k_full": 292.97066067361175}
{"fid50k_full": 169.16669467733635}
{"fid50k_full": 104.17434761256511}
{"fid50k_full": 77.21204012838152},
{"fid50k_full": 61.95879270987923},
{"fid50k_full": 50.009663622336234}
{"fid50k_full": 43.08807716687272},
{"fid50k_full": 37.904640055550765}
{"fid50k_full": 34.47229400106767},
{"fid50k_full": 29.94065800860855},
{"fid50k_full": 26.723918326413308}
{"fid50k_full": 25.046729043965996}
{"fid50k_full": 23.50986657022635},
```


base


```
{"fid50k_full": 308.6027637342727},
{"fid50k_full": 188.35161891464634}
{"fid50k_full": 102.0666766370245},
{"fid50k_full": 74.12198189088315},
{"fid50k_full": 56.47082054678368},
{"fid50k_full": 46.36170454926968},
{"fid50k_full": 39.297681694553624}
{"fid50k_full": 34.63167871020097},
{"fid50k_full": 30.801296149502548}
{"fid50k_full": 27.542499252071813}
{"fid50k_full": 25.989908475363904}
{"fid50k_full": 24.2476226967577},
{"fid50k_full": 22.25654370469862},
```


modified

compare

0.0	0.05	0.1	0.2	0.5
292.97066067361175}	300.3386338709417},	302.88952940807013},	308.6027637342727},	298.6805429891298},
169.16669467733635}	180.53021873299394},	180.32552039793552},	188.35161891464634},	167.33502640428702}
104.17434761256511}	104.14130874716092},	110.30918400098402},	102.0666766370245},	95.57448838887046},
77.21204012838152},	76.4351760120521},	75.62196182655063},	74.12198189088315},	66.7350915320674},
61.95879270987923},	61.07295412898378},	62.71321848105495},	56.47082054678368},	56.96914544392126}
50.009663622336234}	49.12149438130549},	49.32456160903869},	46.36170454926968},	46.54836767819865},
43.08807716687272},	42.58373251286814},	40.99304976306831},	39.297681694553624},	41.9487491812273},
37.904640055550765}	37.131674864412574},	37.134160752974246},	34.63167871020097},	36.91568147413381},
34.47229400106767},	33.087471674861212},	32.237861314354475},	30.801296149502548},	33.0822595593487},
29.94065800860855},	28.852155483162214},	28.629787844041115},	27.542499252071813},	29.7307164629460},
26.723918326413308}	26.256436466492101},	26.11759708913024},	25.989908475363904},	27.5873104598453},

text features (cos similarity matrix)

0.85	0.64	0.54	0.38	0.45	0.45	0.58	0.68	0.57	0.73	0.63	0.55	0.57	0.44	0.58	0.56
0.48	0.81	0.42	0.36	0.50	0.62	0.59	0.54	0.54	0.56	0.50	0.53	0.57	0.38	0.56	0.51
0.65	0.59	0.71	0.35	0.50	0.52	0.58	0.68	0.47	0.60	0.78	0.53	0.70	0.46	0.63	0.52
0.34	0.39	0.31	0.94	0.28	0.43	0.42	0.39	0.35	0.28	0.36	0.41	0.31	0.41	0.42	0.34
0.46	0.57	0.51	0.34	0.81	0.51	0.53	0.53	0.53	0.48	0.59	0.52	0.48	0.38	0.50	0.41
0.40	0.70	0.33	0.40	0.41	0.76	0.59	0.49	0.45	0.47	0.43	0.50	0.44	0.40	0.46	0.41
0.46	0.49	0.49	0.43	0.45	0.68	0.72	0.49	0.49	0.42	0.47	0.52	0.46	0.49	0.56	0.41
0.56	0.49	0.62	0.40	0.42	0.51	0.49	0.77	0.58	0.53	0.57	0.47	0.52	0.38	0.53	0.46
0.40	0.46	0.37	0.41	0.43	0.46	0.55	0.49	0.81	0.40	0.41	0.37	0.38	0.36	0.45	0.34
0.75	0.71	0.56	0.35	0.49	0.48	0.59	0.70	0.56	0.89	0.67	0.57	0.65	0.45	0.62	0.66
0.44	0.44	0.55	0.34	0.47	0.43	0.44	0.54	0.51	0.48	0.76	0.36	0.41	0.35	0.41	0.34
0.52	0.68	0.45	0.40	0.47	0.56	0.63	0.60	0.48	0.54	0.55	0.88	0.57	0.48	0.57	0.54
0.63	0.65	0.61	0.42	0.48	0.51	0.61	0.79	0.58	0.62	0.65	0.55	0.78	0.44	0.68	0.56
0.48	0.57	0.46	0.47	0.42	0.54	0.56	0.56	0.45	0.49	0.53	0.59	0.51	0.79	0.56	0.52
0.49	0.41	0.43	0.35	0.29	0.46	0.43	0.43	0.34	0.43	0.43	0.37	0.51	0.42	0.81	0.47
0.55	0.45	0.49	0.32	0.28	0.40	0.40	0.42	0.36	0.56	0.39	0.42	0.47	0.53	0.51	0.80

example

- A girl sitting in a wheelchair is playing tennis.
- People on a tennis court are playing tennis.
- A group of surfboards sitting up against wooden poles.
- A bathroom sink sitting on top of a wooden cabinet.
- A person that is jumping in the sky on a snowboard.
- a group of people on skis in the snow.

contrastive loss function (clip mask)

$$M_{n\times n} = \begin{bmatrix} 1 & c & \cdots & 0 \\ c & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & c & \cdots & 1 \end{bmatrix}$$

$$S''_{n\times n} = M_{n\times n} \cdot S'_{n\times n}$$

element wise product

$$S'_{n\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

contrastive loss function (clip mask)

ϕ_i : # non-zero number in i-th row

$$R_i = (\sum_{j=1}^{n} S''_{ij})/(1 + c \cdot (\phi_i - 1))$$

$$L_{contra} = \sum_{i=1}^{n} log(\tau \cdot (R_i))$$

final loss function (clip mask)

$$S_{n\times n} = exp(cos(u_{n\times 1\times k}, v_{1\times n\times k})/\tau)$$

$$S'_{n\times n} = softmax(S_{n\times n})$$

$$S''_{n\times n} = M_{n\times n} \cdot S'_{n\times n}$$

$$R_i = (\sum_{j=1}^{n} S''_{ij})/(1 + c \cdot (\phi_i - 1))$$

$$L_{contra} = \sum_{i=1}^{n} log(\tau \cdot (R_i))$$

shallow mapping network (x)

- Mapping network depth For the “Shallow mapping” case in Figure 8a, we reduced the depth of the mapping network from 8 to 2. Reducing the depth further than 2 yielded consistently inferior results, confirming the usefulness of the mapping network. In general, we found depth 2 to yield slightly better results than depth 8, making it a good default choice for future work.

ref : [2006.06676] Training Generative Adversarial Networks with Limited Data (arxiv.org)

FFHQ (256×256)	**2k**	**10k**	**140k**
Baseline	78.80 ±2.31	30.73 ±0.48	3.66 ±0.10
PA-GAN [48]	56.49 ±7.28	27.71 ±2.77	3.78 ±0.06
WGAN-GP [15]	79.19 ±6.30	35.68 ±1.27	6.54 ±0.37
zCR [53]	71.61 ±9.64	23.02 ±2.09	**3.45** ±0.19
Auxiliary rotation [6]	66.64 ±3.64	25.37 ±1.45	4.16 ±0.05
Spectral norm [31]	88.71 ±3.18	38.58 ±3.14	4.60 ±0.19
Shallow mapping	71.35 ±7.20	27.71 ±1.96	3.59 ±0.22
Adaptive dropout	67.23 ±4.76	23.33 ±0.98	4.16 ±0.05
ADA (Ours)	**16.49** ±0.65	**8.29** ±0.31	3.88 ±0.13

Comparison methods

iird loss (o)


```
{"fid50k_full": 294.73744220867707}
{"fid50k_full": 170.53164601240377}
{"fid50k_full": 104.9150042903048},
{"fid50k_full": 72.91047187559334},
{"fid50k_full": 56.432818168206815}
{"fid50k_full": 46.39171702014412},
{"fid50k_full": 42.594051294424986}
{"fid50k_full": 36.63453660354291},
{"fid50k_full": 33.90016837601769},
```


without iird loss


```
{"fid50k_full": 308.6027637342727},
{"fid50k_full": 188.35161891464634}
{"fid50k_full": 102.0666766370245},
{"fid50k_full": 74.12198189088315},
{"fid50k_full": 56.47082054678368},
{"fid50k_full": 46.36170454926968},
{"fid50k_full": 39.297681694553624}
{"fid50k_full": 34.63167871020097},
{"fid50k_full": 30.801296149502548}
```


with iird loss

iic loss (o)

{"fid50k_full": 292.48173967961174},
{"fid50k_full": 171.76532843286873},
{"fid50k_full": 105.35854999031206},
{"fid50k_full": 77.22717405597767},
{"fid50k_full": 62.84395965257046},
{"fid50k_full": 50.43996061933641},
{"fid50k_full": 41.372698253324245},
{"fid50k_full": 36.979935086986075},

without iic loss

{"fid50k_full": 308.6027637342727},
{"fid50k_full": 188.35161891464634},
{"fid50k_full": 102.0666766370245},
{"fid50k_full": 74.12198189088315},
{"fid50k_full": 56.47082054678368},
{"fid50k_full": 46.36170454926968},
{"fid50k_full": 39.29768169455362 4},
{"fid50k_full": 34.63167871020097},

with iic loss

disable style mixing regularization (o)


```
{"fid50k_full": 310.96906962882315}
{"fid50k_full": 170.13250381594995}
{"fid50k_full": 101.64768307111036}
{"fid50k_full": 73.00872630532605},
{"fid50k_full": 58.0460023159047},
{"fid50k_full": 45.38762969198932},
{"fid50k_full": 40.00212552783568},
{"fid50k_full": 34.9307794498806},
{"fid50k_full": 31.371506506415464}
{"fid50k_full": 27.609065998706882}
```


enable


```
{"fid50k_full": 306.45143353507126}
```


disable

Low-Dimensional Latent Space (?)

Dataset	MNIST	SVHN	CIFAR-100	CelebA	CIFAR-10	MS-COCO	ImageNet
MLE (k=3)	7	9	11	9	13	22	26
MLE (k=5)	11	14	18	17	21	33	38
MLE (k=10)	12	18	22	24	25	37	43
MLE (k=20)	13	19	23	26	26	36	43
SOTA Accuracy	99.84	99.01	93.51	-	99.37	-	88.55

ImageNet's dimension estimate is around 40. Accordingly, a latent code of size 512 is highly redundant, making the mapping network's task harder at the beginning of training.

ref :
[2104.08894] The Intrinsic Dimension of Images and Its Impact on Learning (arxiv.org)
[2202.00273] StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets (arxiv.org)

example

	Configuration	FID ↓	IS ↑
A	StyleGAN3	53.57	15.30
B	+ Projected GAN & small **z**	22.98	57.62
C	+ Pretrained embeddings	20.91	35.79
D	+ Progressive growing	19.51	35.74
E	+ ViT & CNN as $\mathbf{F}_{1,2}$	12.43	56.72
F	+ CLF guidance (StyleGAN-XL)	**12.24**	**86.21**

delayed path length regularization

- Path length regularization can lead to poor results on complex datasets.
- We also observe unstable behavior and divergence when using path length regularization in practice. We found that this problem can be circumvented by only applying regularization after the model has been sufficiently trained, i.e., after 200k images.

ref :
[2202.00273] StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets (arxiv.org)

setting

- training step : 1000 kimg (~ 30hr)
- CLIP model : VIT-B16
- training set : COCO2014 training set (~ 80000 images)
- testing set : COCO2014 testing set (~ 40000 images)

tune hyper-parameters

- structure parameters ratio
  - Discriminator : 1
  - CLIP : 2
  - ResNet : 1
- image-text
  - [0.5, 1.25, 2.5, 3.75, 5]
- image-image
  - [0.5, 1.25, 2.5, 3.75, 5]
- contrastive ratio (Heterologous / Homologous)
  - [0, 0.1, 0.2, 0.5, 1, 2]
- clip mask ratio
  - [0, 0.05, 0.1, 0.2, 0.5]

base parameters

- structure parameters ratio
  - Discriminator : 1
  - CLIP : 2
  - ResNet : 1
- image-text
  - [0.5, 1.25, 2.5, 3.75, 5]
- image-image
  - [0.5, 1.25, 2.5, 3.75, 5]
- contrastive ratio (Heterologous / Homologous)
  - [0, 0.1, 0.2, 0.5, 1, 2]
- clip mask ratio
  - [0, 0.1, 0.2, 0.5]

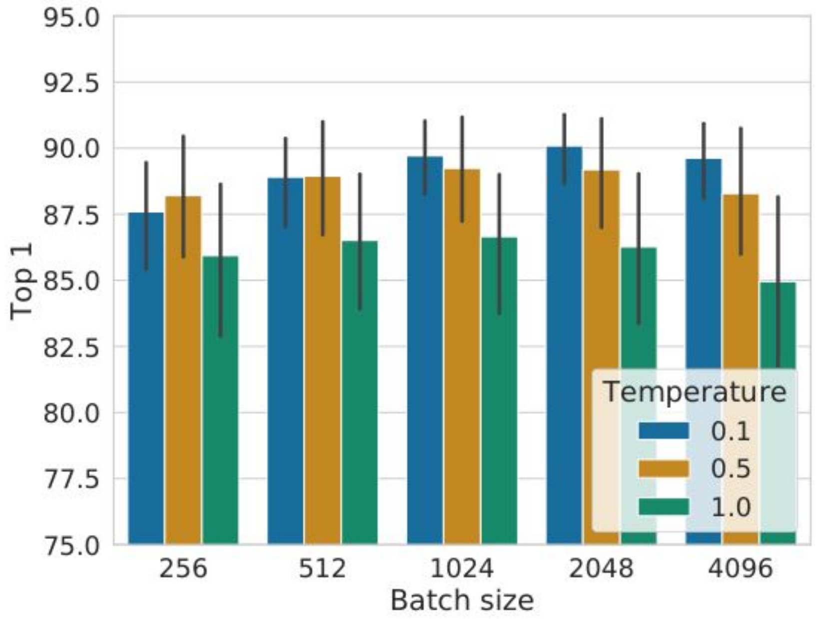
schedule

- experiments (~ 20 days) (O)
- tune hyper-parameters (~ 20 days) (in progress)

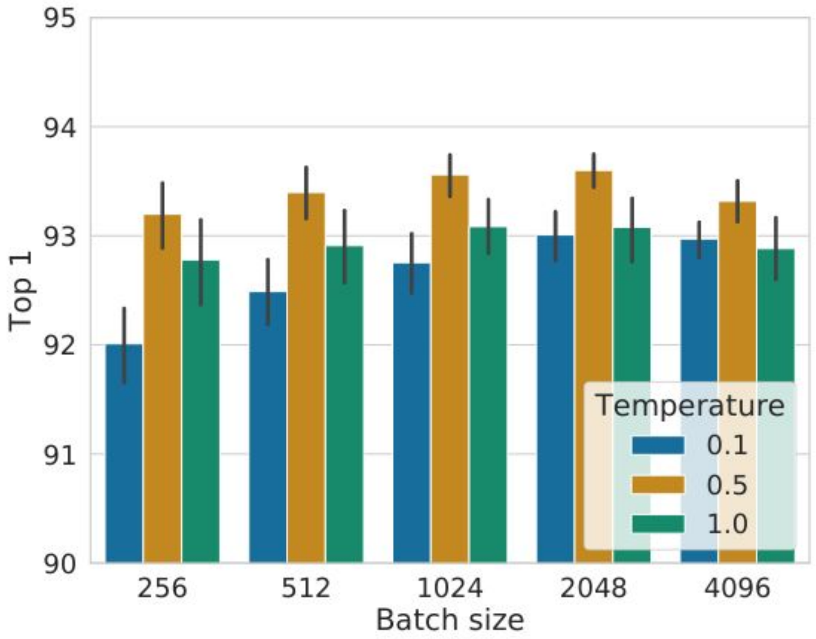
study

- Exploring Generative Adversarial Networks for Text-to-Image Generation with Evolution Strategies
  - https://arxiv.org/abs/2207.02907
- StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets
  - StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets (arxiv.org)
- Diffusion Models Beat GANs on Image Synthesis
  - 49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf (neurips.cc)

NT-Xent loss (different temperature)



(a) Training epochs ≤ 300



(b) Training epochs > 300

contra loss structure

$$Sim(u,v) = exp(cos(u,v)/\tau)$$

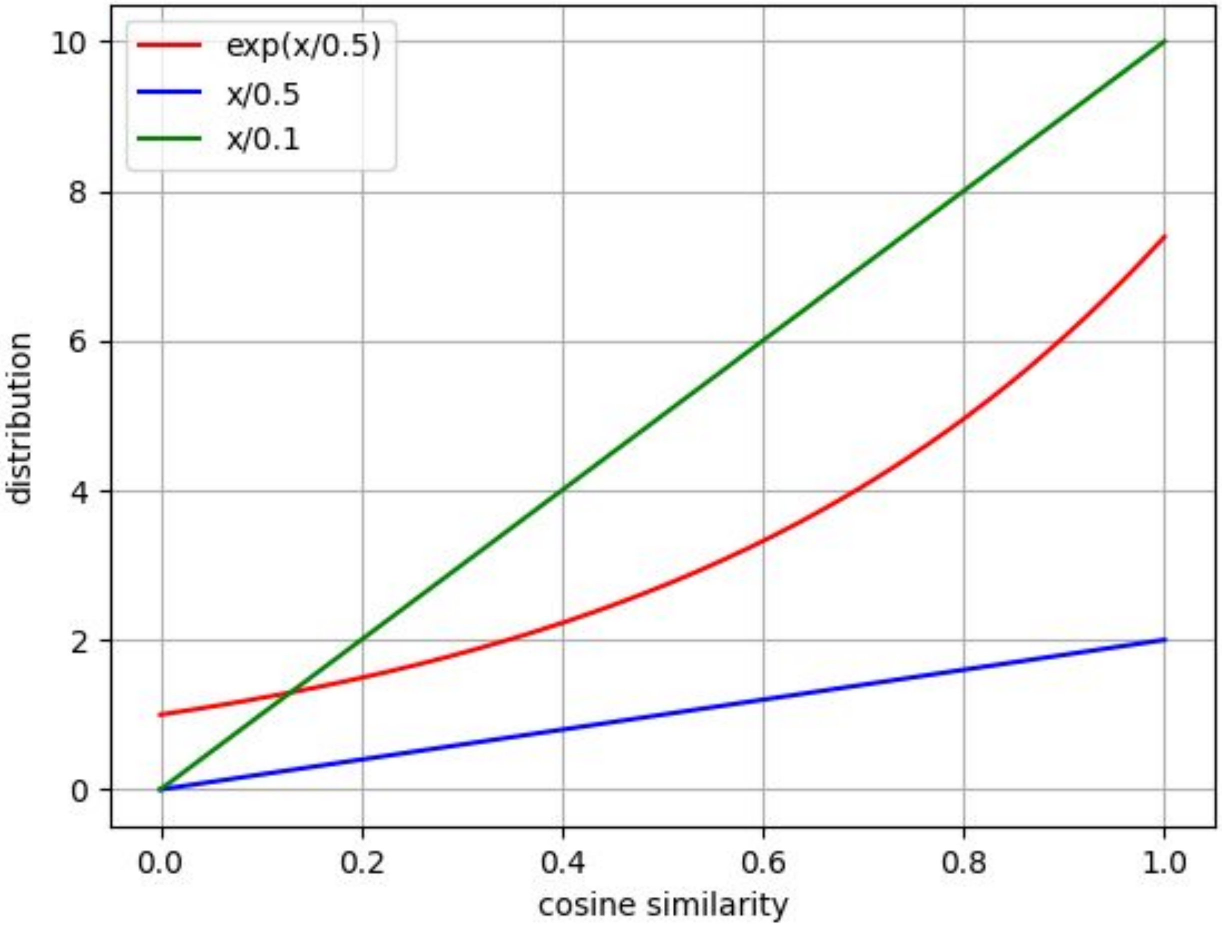
$$-\tau\sum_{i=1}^{n} log(\frac{exp(Sim(u_i,v_i))}{\sum_{j=1}^{n} exp(Sim(u_j,v_i))})$$

$$Sim(u,v) = cos(u,v)/\tau$$

$$-\sum_{i=1}^{n} log(\frac{exp(Sim(u_i,v_i))}{\sum_{j=1}^{n} exp(Sim(u_j,v_i))})$$

example

```
clip img:img postive  pair : 0.7656
clip img:img negative pair : 0.4238
clip img:txt postive  pair : 0.2957
clip img:txt negative pair : 0.0867
res  img:img postive  pair : 0.8743
res  img:img negative pair : 0.5021
```

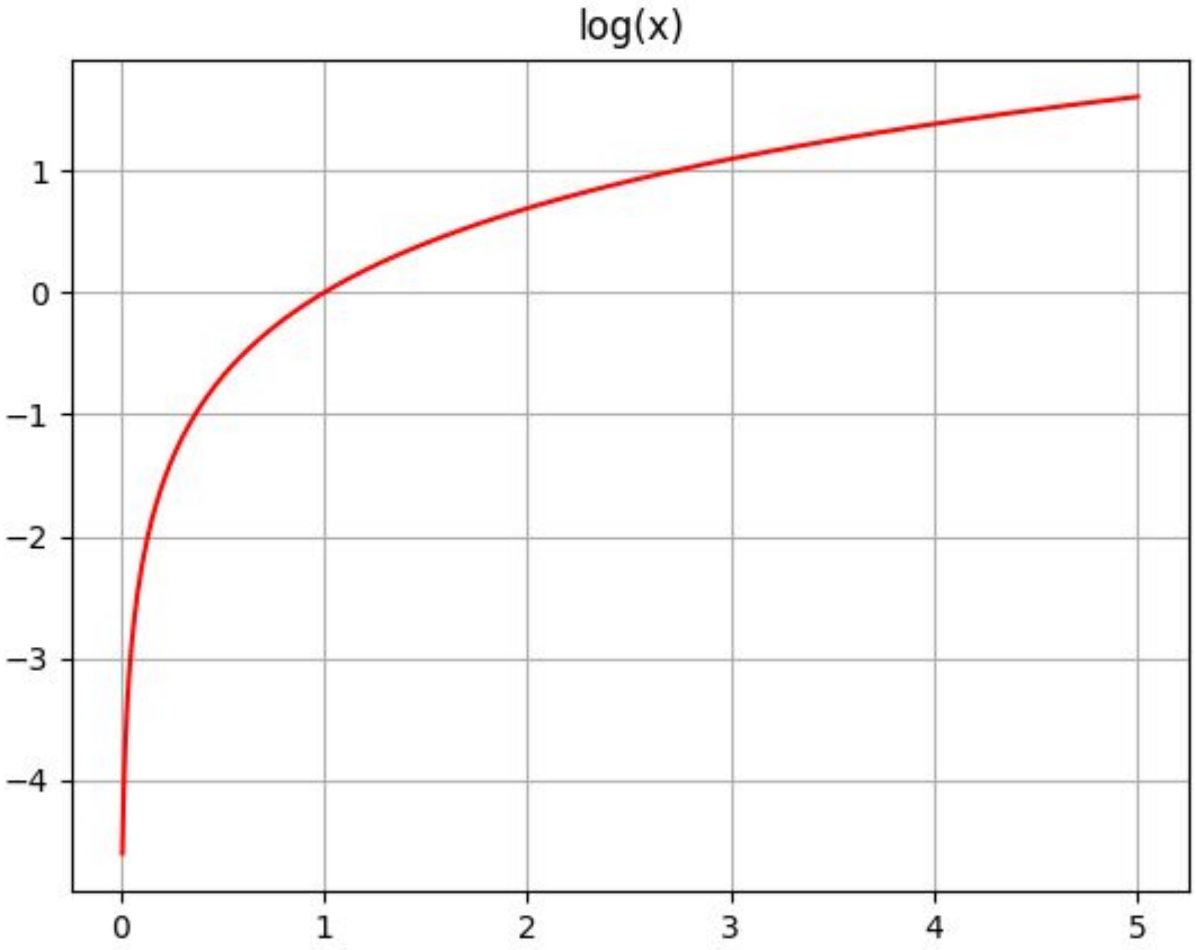

exp(x/0.5)
x/0.5
x/0.1
distribution
cosine similarity
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0

problem

```
sim = torch.cosine_similarity(mat1.unsqueeze(1), mat2.unsqueeze(0), dim=-1)
sim = torch.exp(sim/temp)
sim = torch.diagonal(F.softmax(sim, dim=1)) * temp
return torch.log(sim2)
```

$$Sim(u,v) = exp(cos(u,v)/\tau)$$

$$-\tau \sum_{i=1}^{n} log(\frac{exp(Sim(u_i, v_i))}{\sum_{j=1}^{n} exp(Sim(u_j, v_i))})$$


log(x)
1
0
−1
−2
−3
−4
0
1
2
3
4
5