5/19 meeting

應名宥

Schedule

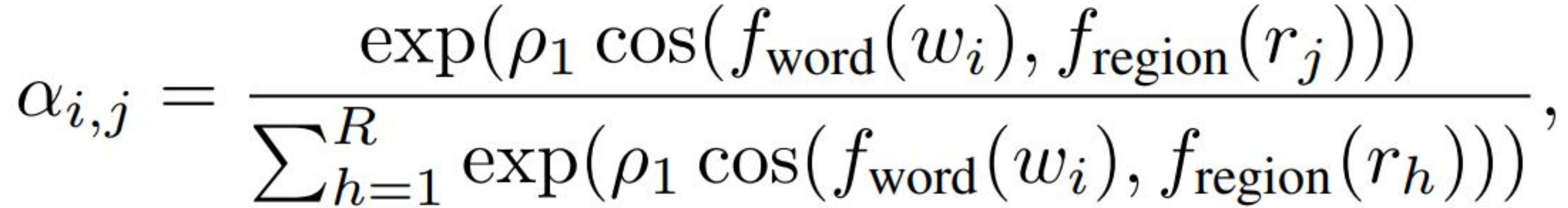
- speed up inference and training (✓) (30、2)
- region - word contrastive loss (✓) (XMC-GAN)
- img - img contrastive loss (✓) (XMC-GAN)
- text - label loss (✓)

contrastive loss component used in XMC-GAN

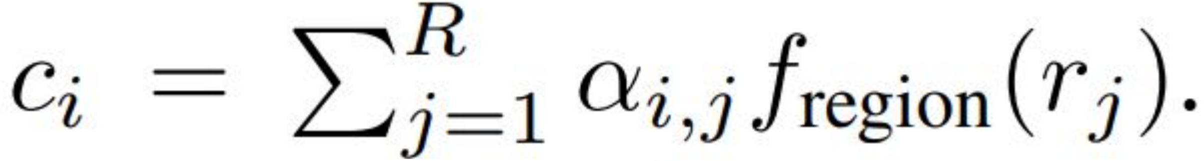
- image - sentence
- region - word
- image - image using discriminator
- image - image using VGG-19 (pretrained on ImageNet)

region - word loss

XMC - GAN

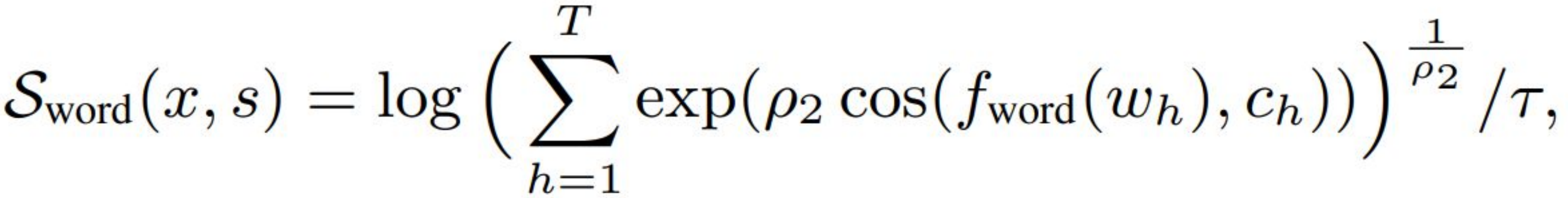


$$\alpha_{i,j} = \frac{\exp(\rho_1 \cos(f_{\text{word}}(w_i), f_{\text{region}}(r_j)))}{\sum_{h=1}^{R} \exp(\rho_1 \cos(f_{\text{word}}(w_i), f_{\text{region}}(r_h)))},$$

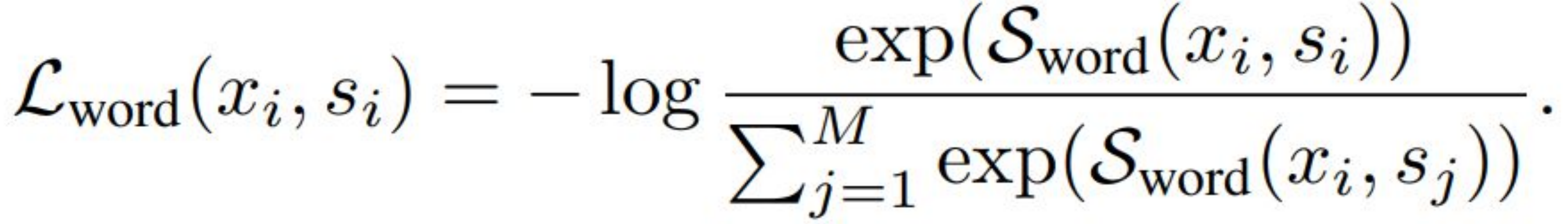


$$c_i \ = \ \sum_{j=1}^{R} \alpha_{i,j} f_{\text{region}}(r_j).$$

most aligned region feature



$$\mathcal{S}_{\text{word}}(x, s) = \log\Big(\sum_{h=1}^{T} \exp(\rho_2 \cos(f_{\text{word}}(w_h), c_h))\Big)^{\frac{1}{\rho_2}} / \tau,$$



$$\mathcal{L}_{\text{word}}(x_i, s_i) = -\log \frac{\exp(\mathcal{S}_{\text{word}}(x_i, s_i))}{\sum_{j=1}^{M} \exp(\mathcal{S}_{\text{word}}(x_i, s_j))}.$$

region - word problem

- linear combination region features
- 實作是否正確

text - label loss

vgg - 19 bn

word - label loss

1. same prompts
2. different prompts
3. no prompts

cos sim	same prompts	different prompts	no prompts
tree - leaf	0.83	0.73	0.80
tree - blood	0.8159	0.74	0.7832

text - label loss

- prompts : ‘photo of a leaf’
- prompts : ‘photo of a blood’
- text : ‘A bird that is on a tree’

cos sim	-
tree - leaf	0.824
tree - blood	0.7578

text - label loss

dist	-	-	-	-	-
case 1	0.2	0.2	0.2	0.2	0.2
case 2	0.80	0.01	0.01	0.01	0.17

cos sim	-	-	-	-	-
case 1	0.75	0.77	0.73	0.67	0.80
case 2	0.95	0.24	0.30	0.12	0.53

score 1 : 3.72

score 2 : 2.14

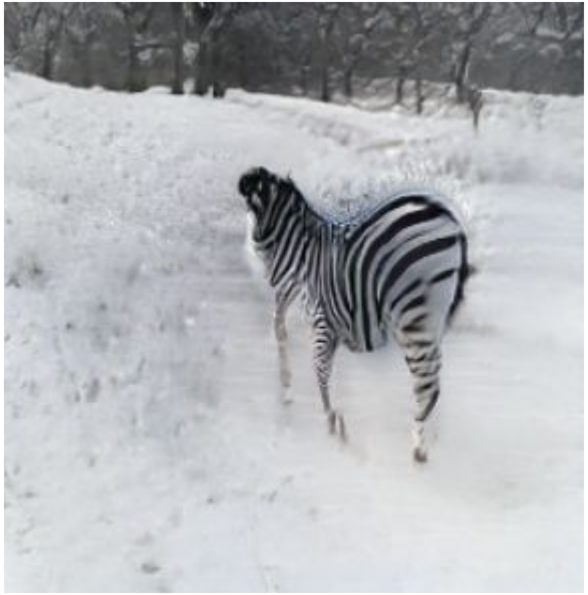
score 1 : 0.744

score 2 : 0.856

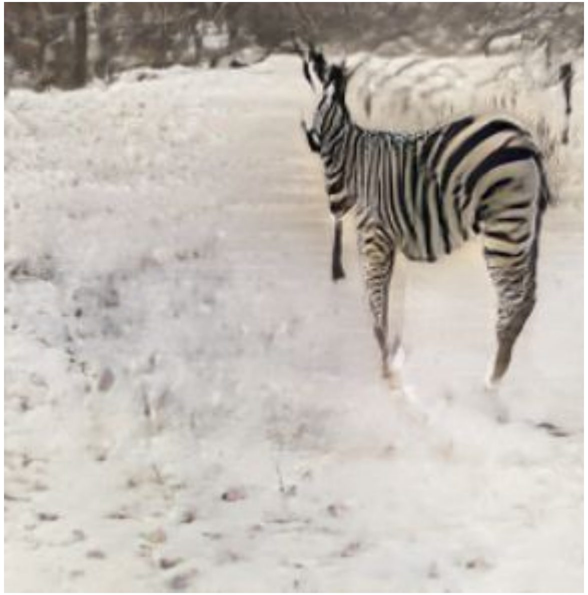
scores

1. text image score
2. image real score
3. region word score
4. text label score

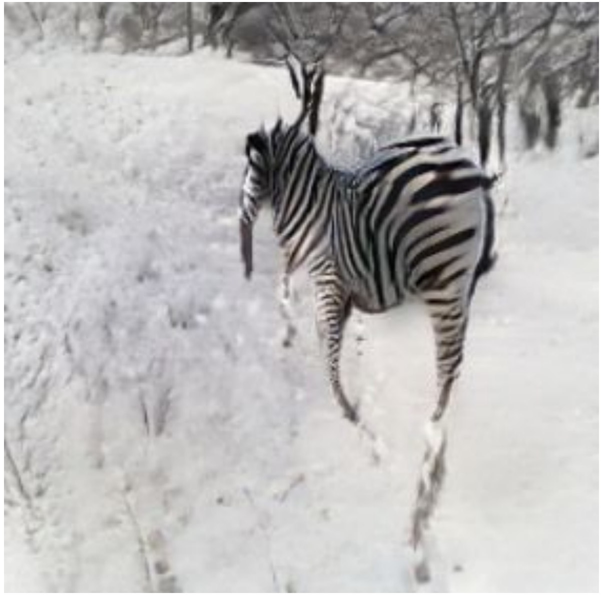
result (add new score)



normal



before



after

todo list

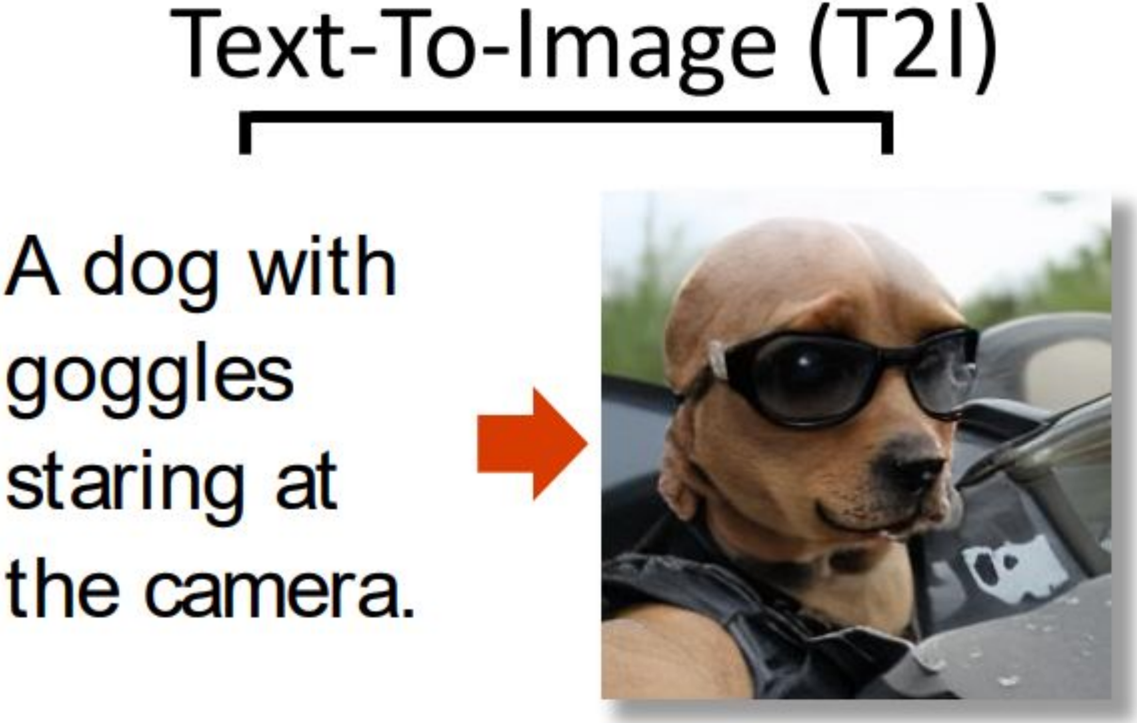
- attentional self-modulation layer (XMC-GAN)
- testing loss function
- noise memory data
- study diffusion model

zero shot

- As an example of a zero-shot learning setting, consider the problem of having. a learner read a large collection of text and then solve object recognition problems.It may be possible to recognize a specific object class even without having seen an image of that object if the text describes the object well enough. For example,having read that a cat has four legs and pointy ears, the learner might be able to guess that an image is a cat without having seen a cat before.

ref : https://www.deeplearningbook.org/contents/representation.html

zero shot in t2i task



Text-To-Image (T2I)

A dog with goggles staring at the camera.