Unconditional Image-Text Pair Generation With Multimodal Cross Quantizer

Hyungyung Lee[1], Sungjin Park[1], Joonseok Lee[2,3], Edward Choi[1]

[1]KAIST [2]Google Research [3]Seoul National University

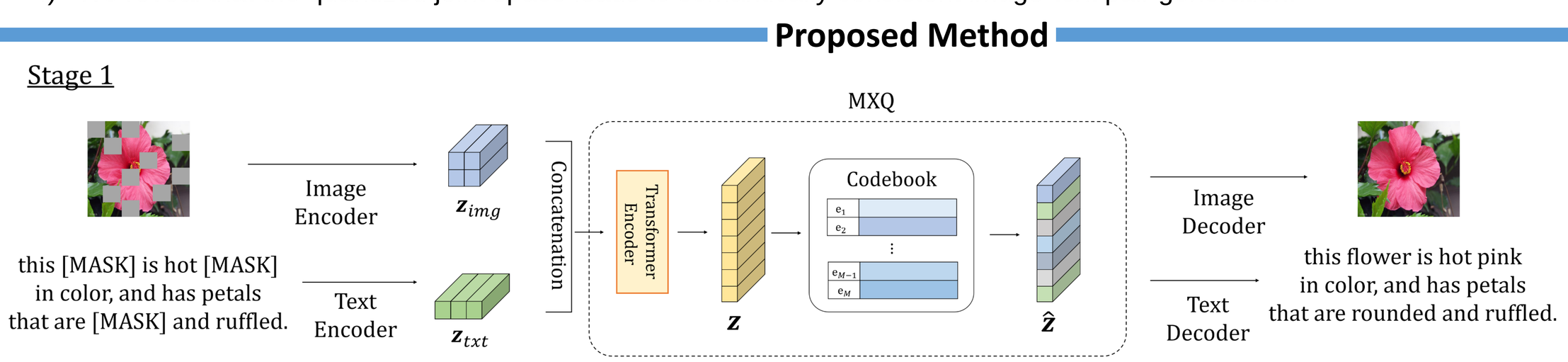
Paper link:

Overview

- Goal: we aim to generate image-text pairs simultaneously without any conditional input
- Contributions

1) We propose MXQ-VAE that learns a quantized joint representation space for unconditional image-text pair generation.
2) Experimental results reveal that MXQ-VAE generates a semantically consistent image-text pairs on multiple benchmark datasets
3) Also, MXQ-VAE learns meaningful semantic correlation between image and text in the quantized joint space.
4) We reveal that the quantized joint space leads to semantically consistent image-text pair generation.

Proposed Method



- Stage 1: Learn a joint quantized representation space (named as MXQ-VAE)

  1) MXQ-VAE takes masked image-text pairs as input, and learns a quantized joint representation space.

  2) Then, the input is converted into a unified code sequence ($\hat{z}$).

- Stage 2: Unconditional image-text pair generation with a unified discrete code sequence

  1) Autoregressive transformer models the joint distribution over the code sequence.

  2) At inference, MXQ-VAE decodes a sampled code sequence to an image-text pair.

Dataset

- Caption MNIST (text augmented MNIST)

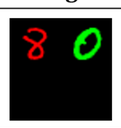
  : Each image-text pair contains several color, digit and position.

- Oxford Flower-102, CUB-200-2011, COCO
- Degree Dataset

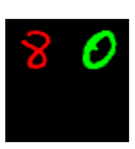
  : To evaluate the semantic correlation between image and text in the quantized joint space in Stage 1.

  : We gradually adjust the degree of the alignment between image and text by replacing the color and digit in text to other random color and digit.

Quadrant	Image	Text
Single		This is white 7.
Quad1		The 6 on the lower right is blue.

Quadrant	Image	Text
Quad2		The green 0 is on the upper right, and the top left 8 is red.
Quad3		the lower left 4 is blue, the red 0 is on the lower right, and the 2 on the top right is green.
Quad4		The bottom left 1 is blue, the upper right 7 is white, the bottom right 0 is red, and the upper left 3 is green.

Example of Caption MNIST

Image	Degree	Text
	2	The green 0 is on the upper right, and the top left 8 is red.
	1	The white 3 is on the upper right, and the top left 8 is red.
	0	The blue 1 is on the upper right, and the top left 7 is green.

Example of the Caption MNIST Quad2 Degree Dataset

Image	Degree	Text
	2	this flower has petals that are yellow and has brown stamen.
	1	this flower has petals that are red and has brown stamen.
	0	this flower has petals that are maroon and has wine stamen.

Example of the Flower Quad2 Degree Dataset

Results

Dataset	Degree	Models	Degree 4	Degree 3	Degree 2	Degree 1	Degree 0
Caption MNIST	Quad4	Only Sharing C	0.486	0.443	0.394	0.358	0.315
		MXQ-VAE w/o TC	1.0	0.975	0.951	0.929	0.906
		MXQ-VAE w/o IM	0.896	0.802	0.698	0.595	0.498
		MXQ-VAE (Ours)	0.969	0.729	0.489	0.248	0.009
Flower	Quad4	Only Sharing C	0.939	0.704	0.516	0.321	0.131
		MXQ-VAE w/o TC	1.0	0.944	0.886	0.866	0.810
		MXQ-VAE w/o IM	0.997	0.728	0.482	0.278	0.067
		MXQ-VAE (Ours)	0.996	0.737	0.490	0.250	0.014
CUB	Quad4	Only Sharing C	0.985	0.771	0.572	0.356	0.155
		MXQ-VAE w/o TC	1.0	0.948	0.894	0.825	0.748
		MXQ-VAE w/o IM	0.998	0.833	0.645	0.424	0.181
		MXQ-VAE (Ours)	0.995	0.749	0.515	0.292	0.083

Table1. Multimodal semantic correlation on the Degree datasets

Models	Single	Quad1	Quad2	Quad3	Quad4	Average
I&T	0.979	0.926	0.675	0.434	0.255	0.654
T&I	0.803	0.780	0.458	0.282	0.161	0.497
I_T_{Embd}	0.953	0.953	0.956	0.958	0.849	0.945
T_{Embd}_I	0.086	0.895	0.913	0.916	0.828	0.728
MXQ-VAE (Ours)	**0.998**	**0.997**	**0.994**	**0.996**	**0.974**	**0.992**

Table2. Semantic consistency of the generated samples on Caption MNIST

Models	Flower				CUB			
	Modified unigram precision	Sentence similarity Top-1	Top-5	Top-10	Modified unigram precision	Sentence similarity Top-1	Top-5	Top-10
Joint GAN [ ] *	0.324	0.808	0.788	0.774	-	-	-	-
MMVAE [ ]	-	-	-	-	0.262	0.707	0.682	0.667
MXQ-VAE (Ours)	**0.428**	**0.941**	**0.926**	**0.916**	**0.478**	**0.948**	**0.919**	**0.900**

Table3. Semantic consistency of the generated samples on Flower and CUB

- Enhance the alignment between image and text (Table 1)

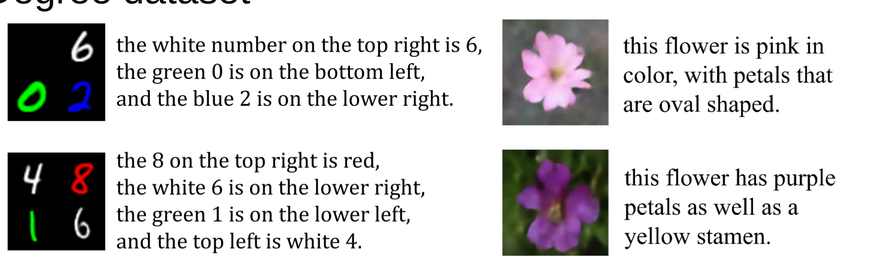
  → Measure reconstruction accuracy between the reconstructed text and the input text of the Degree dataset

  → Our model reaches near 1.0, 0.75, 0.5, 0.25, 0.0 on the Quad4 Degree dataset.

- Generate semantically consistent image-text pairs (Table 2, 3)

  → Measure semantic consistency between the generated image and text

  → Our model outperforms all baselines on three datasets.



Generated image-text pairs