Image Annotation Using VQGAN and Backtranslation

Dong-Hyuck Im
Creative Content Research Division, ETRI
Daejoen, Korea
iammoni@etri.re.kr

Yong-Seok Seo
Creative Content Research Division, ETRI
Daejoen, Korea
yongseok@etri.re.kr

***Abstract*— The paper presents an image annotation technique based on VQGAN and backtranslation. Using the VQGAN model, it learns a quantized codebook that expresses an image in block units, encodes the image using the codebook, and then trains a back-translation model which translate image to text using small amount of text/image pair data. Using the back-translation model, we can generate synthetic text data from image only dataset. Experimental result shows that we can generate synthetic text describing Korean face images and successfully train a text-to-image generation model.**

Keywords—Image Annotation, VQGAN, Backtranslation, Txt2Img generation

I. Introduction

Text-to-image models take an input natural language description and produce an image matching that description. Quality of text-to-image models such as Dalle-2, Midjourney, and Stable Diffusion began to approach the quality of real photographs and human-drawn art.

Text-to-image models need a lot of text/image pair data for training. However, collecting such data is time-consuming and expensive. On the other hand, image-only data is easier to collect than text/image pair data. We train a back-translation model that converts images to texts using a small amount of text/image pair data, and use it to annotate the image-only data and create synthetic text/image pair data.

We present an image annotation method that uses VQGAN [1] and back-translation [2] models to create synthetic text/image data. We encode images into discrete visual tokens with VQGAN representations. The decoder can reconstruct the images with minimal distortion, even though these tokens are much smaller than the original image. We treat these latent visual tokens as a form of language and use a back-translation model to convert them into text.

II. Related works

A. VQGAN

[1, 3] introduces VQGAN, a method that learns a perceptually rich image representation and models its distribution with a transformer architecture in an autoregressive manner. It shows that the combination of CNNs' inductive bias and transformers' expressivity enables the generation of high-resolution images.

B. Backtransaltion

Neural Machine Translation (NMT) is the state-of-the-art translation method for many language pairs. However, it requires large parallel data for training. Backtranslation [2, 4] is a technique that leverages monolingual data to improve NMT. It first trains a model in one direction (e.g., target to source) and then uses it to generate synthetic source sentences from a monolingual corpus. These synthetic sentences are added to the training data and a new model is trained in the reverse direction (e.g., source to target). Backtranslation is a common practice for building state-of-the-art NMT systems, especially when new parallel text is scarce.

III. METHODS

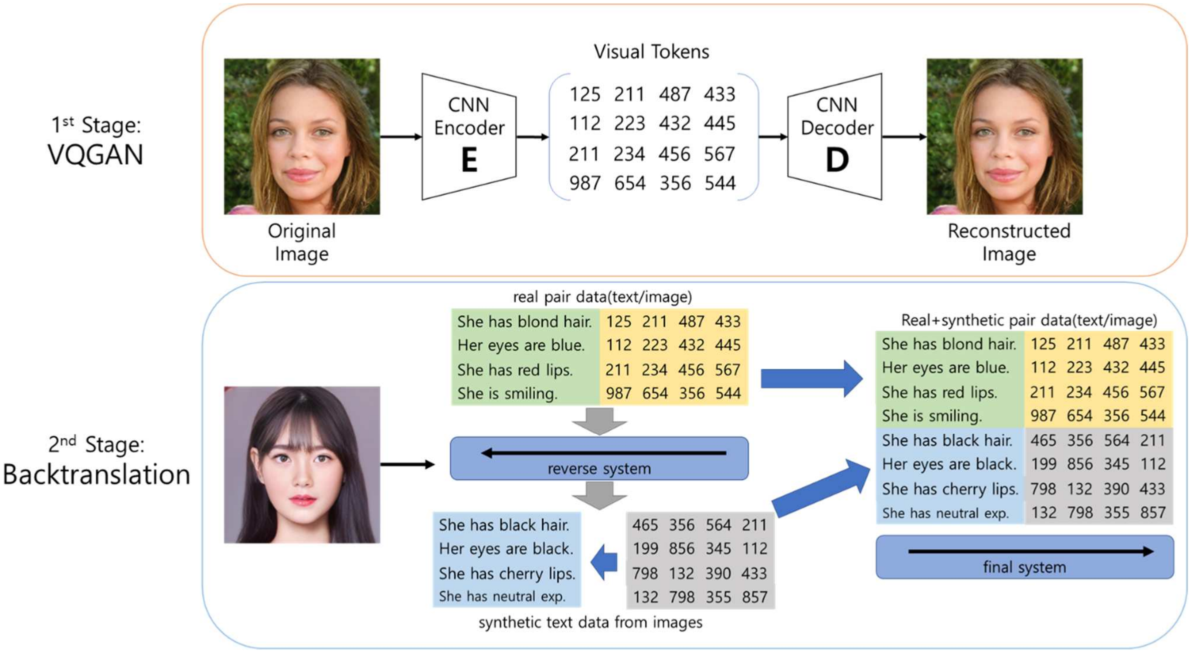
Our annotation method consists of two stages: image representation learning and back-translation. As shown in Figure 1, an input image x is encoded into contextualized vector representations. Then, the back-translation model is trained with supervised learning on these representations and text pairs.

An image can be encoded into a sequence of image tokens using a codebook, and this sequence can be viewed as a sentence. From this perspective, generating text from an image is similar to machine translation. If we consider the text as the source language and the codebook-encoded sentence as the target language, we have a lot of target language data and a little parallel data. Therefore, we will use back-translation to generate a source language (text) that matches the target language (image).

A. 1st Stage: Learning Visual Tokens

An image is represented as visual tokens using CNN encoder. In other words, an image $x \in R^{H\times W\times 3}$ is transformed into $Z = \{z_k\}_{k=1}^{K} \subset R^{h\times w\times n_z}$, where n_z is the dimensionality of

979-8-3503-1327-7/23/$31.00 ©2023 IEEE



Visual Tokens

125	211	487	433
112	223	432	445
211	234	456	567
987	654	356	544

real pair data(text/image)

She has blond hair.	125	211	487	433
Her eyes are blue.	112	223	432	445
She has red lips.	211	234	456	567
She is smiling.	987	654	356	544

She has black hair.	465	356	564	211
Her eyes are black.	199	856	345	112
She has cherry lips.	798	132	390	433
She has neutral exp.	132	798	355	857

synthetic text data from images

Real+synthetic pair data(text/image)

She has blond hair.	125	211	487	433
Her eyes are blue.	112	223	432	445
She has red lips.	211	234	456	567
She is smiling.	987	654	356	544
She has black hair.	465	356	564	211
Her eyes are black.	199	856	345	112
She has cherry lips.	798	132	390	433
She has neutral exp.	132	798	355	857

(Fig. 1) Proposed Architecture

codes. The encoder compresses images to a latent representation. Using the latent code z, the decoder learns how to restore the image x. The decoder is composed of a few residual blocks and a series of strided transposed convolutions that enlarge the representations to the original image size.

The objective function to find the best compression model Q* is as follows.

$$Q^* = \underset{E,G,Z}{\operatorname{argmin}}\ \underset{D}{\max}\ E_{x\sim p(x)}\left[L_{VQ}(E,G,Z)\right.$$
$$\left. + \lambda L_{GAN}(\{E,G,Z\},D)\right]$$

E is the encoder, G is the decoder, Z is the codebook, and D is the discriminator. L_{VQ} is a codebook learning loss that minimizes the reconstruction error in the encoding and decoding process. L_{GAN} is a GAN loss that ensures the generated image quality matches the original one. The training aims to reduce the sum of these two losses. A patch-based discriminator [5] and perceptual loss [6] preserve good perceptual quality while increasing the compression rate.

B. 2nd Stage: Backtranslation over Text and Visual Tokens pairs

We have a small parallel corpus B and a large monolingual corpus M as follows. In B, x_n is the source language text data, and y_n is the target language visual tokens that encode the image. In M, y_s is the target language visual tokens, and there is no source language because we only have image data.

$$\mathrm{B} = \{(x_n, y_n)\}_{n=1}^{N}$$
$$\mathrm{M} = \{y_s\}_{s=1}^{S}$$

Using B first, we can easily train two models. The first one $\hat{\theta}_{x\to y}$ is a model that translates text into visual tokens in a forward direction, and the second one $\hat{\theta}_{y\to x}$ is a model that translates visual tokens into text in a backward direction.

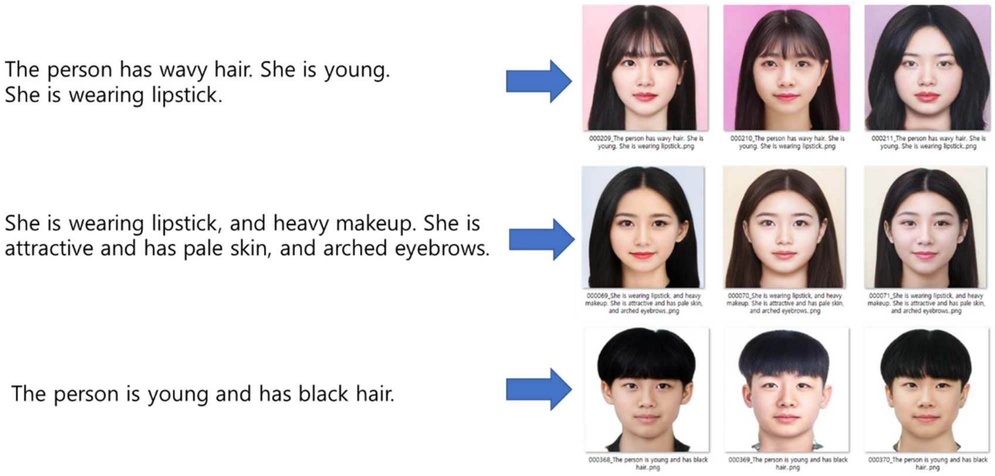
$$\hat{\theta}_{x\to y} = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{n=1}^{N} logP\left(y_n|x_n;\theta_{x\to y}\right)$$
$$\hat{\theta}_{y\to x} = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{n=1}^{N} logP\left(x_n|y_n;\theta_{y\to x}\right)$$

With the back-translation model $\hat{\theta}_{y\to x}$, we can create synthetic text data from the image data M, and form a parallel corpus $\widehat{\mathrm{M}}$ that contains synthetic text data x_s and visual tokens y_s.

$$\widehat{M} = \{(\hat{x}_s, y_s)\}_{s=1}^{S}$$
$$\text{where } \hat{x}_s = \underset{x\in\chi}{\operatorname{argmax}}\ logP\left(x|y_s;\theta_{y\to x}\right)$$

Then, we can train a new text-to-image generation model θ with the new M dataset.

$$\hat{\theta}_{x\to y} = \underset{\theta\in\Theta}{\operatorname{argmax}} \left(\sum_{n=1}^{N} logP\left(y_n|x_n;\theta_{x\to y}\right)\right.$$
$$\left. + \sum_{s=1}^{S} logP\left(y_n|\hat{x}_s;\theta_{x\to y}\right)\right)$$


The person has wavy hair. She is young.
She is wearing lipstick.
000209_The person has wavy hair. She is young. She is wearing lipstick..png
000210_The person has wavy hair. She is young. She is wearing lipstick..png
000211_The person has wavy hair. She is young. She is wearing lipstick..png
She is wearing lipstick, and heavy makeup. She is
attractive and has pale skin, and arched eyebrows.
000069_She is wearing lipstick, and heavy makeup. She is attractive and has pale skin, and arched eyebrows..png
000070_She is wearing lipstick, and heavy makeup. She is attractive and has pale skin, and arched eyebrows..png
000071_She is wearing lipstick, and heavy makeup. She is attractive and has pale skin, and arched eyebrows..png
The person is young and has black hair.
000368_The person is young and has black hair..png
000369_The person is young and has black hair..png
000370_The person is young and has black hair..png

(Fig. 2) Korean face images generated from the prompts

IV. Experimental results

We use the Multi-Modal-CelebA-HQ [7] dataset and our Korean Face Image dataset to evaluate our model. The former is a face image and text pair dataset with 30k high-resolution face images from the CelebA dataset [8]. The latter is an image-only dataset with 30k Korean face images that we collected from the web, social media, etc. We preprocessed the face images for uniform size and quality using face detection, face alignment, and super-resolution.

We train VQGAN on 256 x 256 images with dimZ = 16384 and f = 16, and back-translation with 6 transformer layers and 16 attention heads. We first train a visual tokens-to-text model on the Multi-Modal CelebA-HQ dataset, which has image and text data. The Korean face image dataset does not have text data, so we use back-translation to synthesize text data for the images. We generate 10 synthetic descriptions for each image and obtain 300k image/text pairs. With the synthetic image/text pair data, we train a text-to-image model, the Korean Face Image generator. Figure 2 shows the Korean face images generated by our models.

V. conclusion

We present an image annotation technique that leverages VQGAN and back-translation to generate synthetic text data from image-only datasets. Our method can be applied when text-to-image generation is required but only image data is available. We evaluate our approach on the Korean Face Image dataset and show its effectiveness.

Acknowledgment

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023 (Project Name: Development of artificial intelligence-based copyright infringement suspicious element detection and alternative material content recommendation technology for educational content, Project Number: CR202104003)

References

[1] Patrick Esser, et al. “Taming Transformers for High-Resolution Image Synthesis.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2021.

[2] Sergey Edunov, et al. “Understanding Back-Translation at Scale.” arXiv:1808.09381. 2018

[3] Dong-Hyuck Im and Yong-Seok Seo. "Generating Face Images Using VQGAN and Sparse Transformer." Proceedings of IEEE International Conference on Information and Communication Technology Convergence, 2021

[4] Isaac Caswell, et al. “Tagged Back-Translation.” Proceedings of the Fourth Conference on Machine Translation, 2019

[5] Phillip Isola, et al. “Image-to-Image Translation with Conditional Adversarial Networks.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017

[6] Alexey Dosovitskiy and Thomas Brox. “Generating Images with Perceptual Similarity Metrics based on Deep Networks.” In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, NeurIPS, 2016

[7] Weihao Xia, et al. “TediGAN: Text-Guided Diverse Face Image Generation and Manipulation.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021

[8] Tero Karras, et al. "Progressive growing of GANs for Improved Quality, Stability, and variation." Proceedings of the International Conference on Learning Representations. 2018