Image and Video Tokenization with Binary Spherical Quantization

Yue Zhao[1], Yuanjun Xiong[2], Philipp Krähenbühl[1]

[1]UT Austin [2]Adobe Firefly

Learning discrete visual tokenization

- Useful in visual compression, recognition, and generation

Main drawback:

- (1) Architectural changes between image and video tokenizers
- (2) VQ [1] scales poorly with codebook size

Our contributions

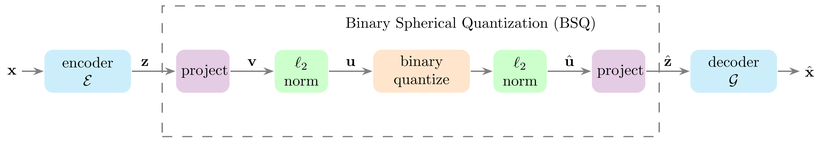
(1) Encoder-decoder with Causal Vision Transformer

- Computationally more efficient than CNNs
- Support variable-length videos as input

(2) A novel quantization method (Binary Spherical Quantization),

- Parameter-efficient with an explicit codebook
- Scalable to arbitrary token dimensions
- Compact: achieve a good rate-distortion trade-off

Overall Architecture



Training objectives

- Reconstruction loss: $\mathcal{L}_{\text{MSE}}(\hat{\mathbf{x}}, \mathbf{x})$
- Quantization loss (entropy term): $\mathcal{L}_{\text{q}} = \mathbb{E}_{\mathbf{u}}\big[H\big(q(\mathbf{u})\big)\big] - \gamma H(\mathbb{E}_{\mathbf{u}}[q(\mathbf{u})])$
- Perceptual loss: $\mathcal{L}_{\text{LPIPS}}(\hat{\mathbf{x}}, \mathbf{x})$
- Adversarial loss: $\mathcal{L}_{\text{GAN}}(\hat{\mathbf{x}}, \mathbf{x})$

Advantages

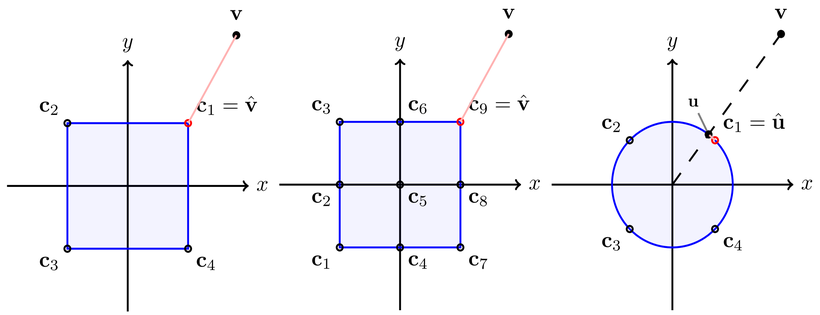
- Efficient entropy computation: $O(2^L) \rightarrow O(L)$, where $\mathbf{u} \in \mathbb{S}^{L-1}$

$$\mathbb{E}_{\mathbf{u}}\big[H\big(q(\mathbf{c}|\mathbf{u})\big)\big] = \mathbb{E}_{\mathbf{u}}\left[\sum_{d=1}^{L} H\big(q(\mathbf{c}_d|\mathbf{u}_d)\big)\right]$$

$$H(\mathbb{E}_{\mathbf{u}}[q(\mathbf{c}|\mathbf{u})]) \approx \sum_{d=1}^{L} H\big(\mathbb{E}_{\mathbf{u}_d}[q(\mathbf{c}_d|\mathbf{u}_d)]\big)$$

- Bounded quantization error: $\mathbb{E}_{\mathbf{u}}[d(\mathbf{u}, \hat{\mathbf{u}})] < \sqrt{2 - 2/\sqrt{L}}$

Illustration of BSQ in the 2D case



(a) Lookup-Free Quantization (LFQ) [2] (b) Finite Scalar Quantization (FSQ) [3] (c) Binary Spherical Quantization (BSQ, ours)

Image Reconstruction

Methods	Params.	# bits	Throughput (img/sec)	ImageNet-1k val			
				PSNR ↑	SSIM ↑	LPIPS ↓	rFID ↓
DALL-E dVAE	98M	13	34	25.46	.7385	.3127	36.84
SD-VAE 1.x (VQ)	68M	14	22.4	22.82	.6354	.0912	1.23
SD-VAE 1.x (kl)	68M	64	22.4	21.99	.6285	.0980	1.35
SD-VAE 2.x (kl)	84M	64	18.9	25.08	.7054	.0731	0.78
BSQ-ViT (**Ours**)	174M	18	45.1	25.36	.7578	.0761	1.14
BSQ-ViT (**Ours**)	174M	36	**45.1**	**28.14**	**.8448**	**.0400**	**0.45**

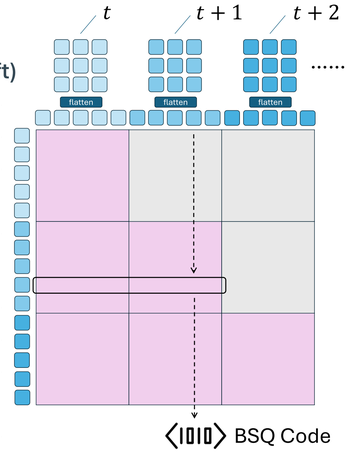
Causal Video Transformer

- Full attention *vs.* Block-wise attention (left)

	UCF-101 val			
	PSNR ↑	SSIM ↑	LPIPS ↓	rFVD ↓
Full Attn.	31.88	.9410	.0241	14.17
Block-wise Attn.	31.49	.9357	.0254	10.57

- Image-only *vs.* finetune

	UCF-101 val			
	PSNR ↑	SSIM ↑	LPIPS ↓	rFVD ↓
Frame only	25.83	.8259	.1108	342.
Video Finetune	31.49	.9357	.0254	10.57

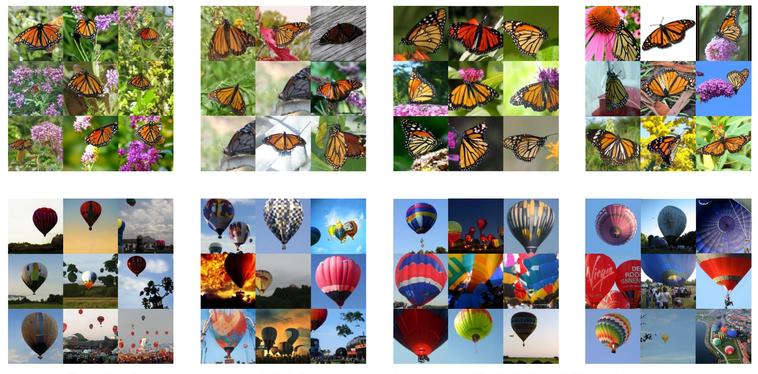


Quantitative Results

Methods	Generation, ImageNet-1k			
	FID ↓	IS ↑	Prec. ↑	Rec. ↑
BigGAN	6.02	**145.8**	**0.86**	0.35
ADM	5.91	93.3	0.70	0.65
MaskGIT + VQ	9.4			
MaskGIT + FSQ	8.5			
MaskGIT + BSQ	**5.44**	139.6	0.80	**0.50**

Methods	Compression, Kodak		
	BPP ↓	MS-SSIM (dB) ↑	LPIPS ↓
JPEG2000	.2986	11.574	.1892
WebP	.2963	12.193	.1655
MAGVIT2	.2812	8.103	.1260
BSQ (Ours)	.2812	**12.852**	**.0823**
BSQ + AC (Ours)	**.2073**	**12.852**	**.0823**

Qualitative Results: Generation on ImageNet



(a) BigGAN (b) ADM (c) MaskGIT+BSQ-ViT (d) Groundtruth

Qualitative Results: Compression on Kodak

(a) Original Image (b) JPEG 2000 (c) WebP (d) BSQ (Ours)

References

[1] Van den Oord, et al. Neural discrete representation learning. *NeurIPS 2017*

[2] Yu et al. "Language Model Beats Diffusion--Tokenizer is Key to Visual Generation." *ICLR 2024*

[3] Mentzer et al. "Finite scalar quantization: VQ-VAE made simple". *ICLR 2024*