Visual Prompting via Image Inpainting

Few-shot Prompting (In-context Learning)

Train a model to develop wide range of abilities, and use those abilities to work on desired downstream task.

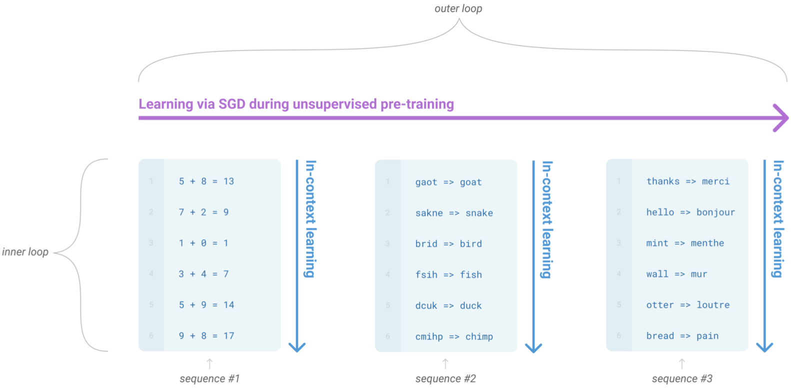

Figure 1: In-context learning

Few-shot Prompting (In-context learning)

Pretrained model receives a natural language instruction and/or a few demonstrations of the desired downstream task, and is expected to complete other instances of the task by simply predicting what comes next.

- Zero-shot : model only receives a natural language instruction.
- One-shot : model receives a natural language instruction, and a single demonstration.
- Few-shot : model receives a natural language instruction and a few (typically ≤ 10) demonstrations.

Fine-tuning

Fine-tuning

`sea otter => loutre de mer` ← example 1

↓

gradient update

↓

`peppermint => menthe poivree` ← example 2

↓

gradient update

↓

...

↓

`plush girafe => girafe peluche` ← example N

↓

gradient update

↓

`cheese =>` ← prompt

Few-shot Prompting (In-context learning)

Zero-shot

`Translate Englisth to French:`	$\leftarrow$ instruction
`cheese =>`	$\leftarrow$ prompt

One-shot

`Translate Englisth to French:`	$\leftarrow$ instruction
`sea otter => loutre de mer`	$\leftarrow$ example
`cheese =>`	$\leftarrow$ prompt

Few-shot

`Translate Englisth to French:`	$\leftarrow$ instruction
`sea otter => loutre de mer`	$\leftarrow$ example 1
`peppermint => menthe poivree`	$\leftarrow$ example 2
`plush girafe => girafe peluche`	$\leftarrow$ example 3
`cheese =>`	$\leftarrow$ prompt

Key Question

Can the in-context learning be generalized to vision tasks?

Recipe

Large capacity image inpainting models

Appropriate (large) training data

$\Longrightarrow$ Visual prompting

Inpainting

The goal of an inpainting model f is to generate an image $y \in \mathbb{R}^{H \times W \times 3}$ from given input image $x \in \mathbb{R}^{H \times W \times 3}$ and binary mask $m \in \{0,1\}^{H \times W}$:

$$y = f(x, m). \tag{1}$$

For the architecture of f, the authors propose the MAE-VQGAN model, which combines Masked AutoEncoder (MAE) and Vector Quantized GAN (VQGAN).

Masked AutoEncoder

1. Divide image into patches
2. Randomly mask certain portion of patches
3. Encode only the visible patches
4. Decode with encoded patches and masked tokens (which represent masked patches) to reconstruct the original image

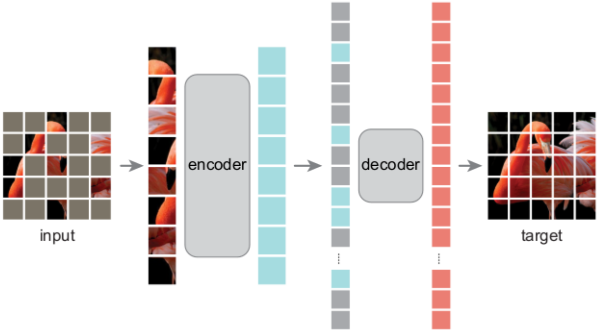

Figure 2: Masked autoencoder

Vector Quantized GAN

Similar to VQVAE, VQGAN utilizes (learnable) quantized codebook to encode latent of image:

$$x \to \hat{z} = E(x) \to z_q = q(E(x)) \to \hat{x} = G(q(E(x))).$$

However, to enrich the codebook, VQGAN jointly train the discriminator as in GAN.

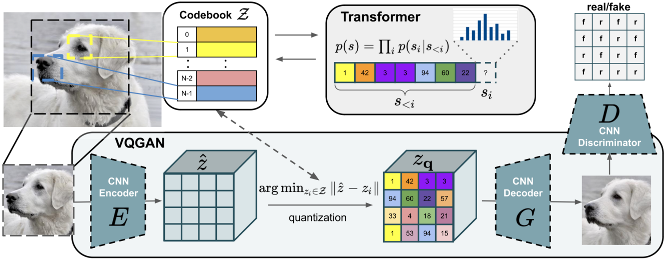

Figure 3: Vector quantized GAN

MAE-VQGAN

Given a pretrained VQGAN $E_{\text{VQ}}, G_{\text{VQ}}, \mathcal{Z}$, and a MAE E_θ, D_ϕ, MAE-VQGAN models the distribution $p_{\theta,\phi}(z_i|x,m)$, where z_i is the visual code of ith patch of x.

- Training:

$$\begin{aligned}&\mathcal{L}_{\theta,\phi}(x,m)\\ =&CE(q(E_{\text{VQ}}(x)), D_\phi(E_\theta(x*m)))*m\end{aligned} \tag{2}$$

- Inference:

$$y = G_{\text{VQ}}(q(D_\phi(E_\theta(x*m)))) \tag{3}$$

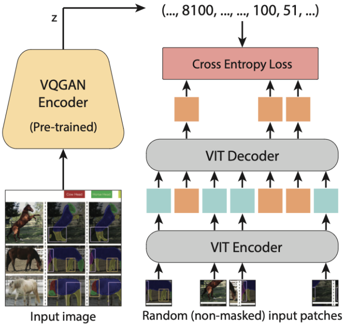

Figure 4: MAE-VQGAN

Prompting

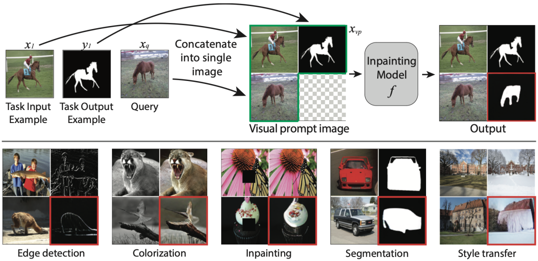

Figure 5: Visual prompting using image inpainting

Computer Vision Figures (CVF) Dataset

- Consist of 88645 images images collected from Arxiv selected from Computer-Vision partition
- Labelled 2000 images and trained a binary classifier to remove unrelated images (e.g. charts, graphs)

Figure 6: Computer vision figures

Experiments

Table 1: **Visual prompting results on computer vision tasks.** For Foreground Segmentation and Single Object Detection, we report the *mIOU* score. For Colorization, we report the *MSE*.

Model	Foreground Segmentation ↑				Single Object Detection ↑				Colorization ↓	
	Split 0	Split 1	Split 2	Split 3	Split 1	Split 2	Split 3	Split 4	MSE	LPIPS
Copy	12.92	17.90	13.52	15.29	12.14	13.50	13.03	12.38	2.63	0.75
BEiT (IN-21k)	0.38	0.93	0.90	0.95	0.24	0.32	0.19	0.10	1.25	0.73
VQGAN (IN-1k)	6.96	10.55	9.59	9.43	5.19	4.99	5.09	5.10	2.44	0.66
MAE (IN-1k)	1.92	6.76	3.85	4.57	1.37	1.98	1.62	1.62	1.13	0.87
MAE-VQGAN (IN-1k)	2.22	7.07	5.48	6.28	3.34	3.21	2.80	2.80	3.31	0.75
BEiT (Figures)	5.38	3.94	3.20	3.29	0.17	0.02	0.14	0.16	0.60	0.70
VQGAN (Figures)	12.56	17.51	14.27	15.06	2.27	2.37	2.48	1.99	1.50	0.56
MAE (Figures)	17.42	25.70	18.64	16.53	5.49	4.98	5.24	5.84	**0.43**	0.55
MAE-VQGAN (Figures)	**27.83**	**30.44**	**26.15**	**24.25**	**24.19**	**25.20**	**25.36**	**25.23**	0.67	**0.40**

Experiments

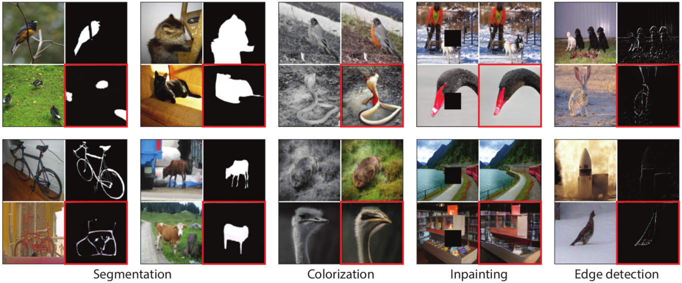

Figure 7: Examples of visual prompting (downstream tasks)

Experiments

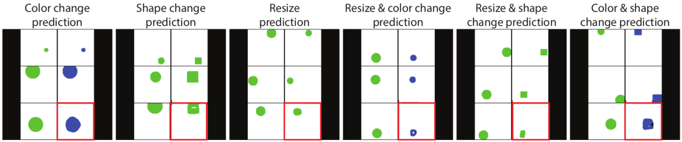

Figure 8: Examples of visual prompting (reasoning)

Experiments

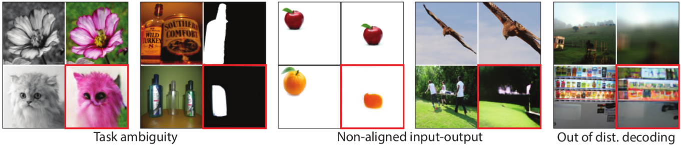

Figure 9: Failed examples

Thank You

Q & A