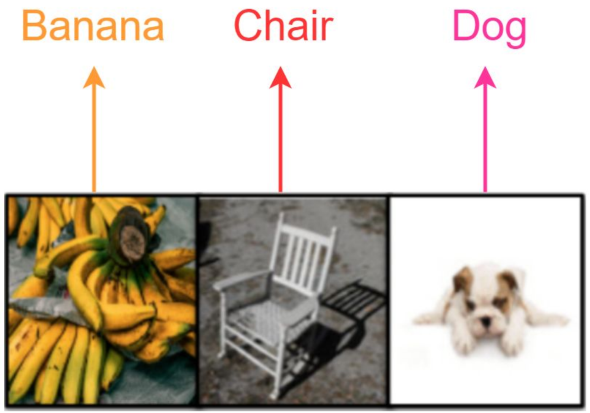
Pretraining Transformers for Domain Adaptation

NeurIPS 2021

Burhan Ul Tayyab and Nicholas Chua

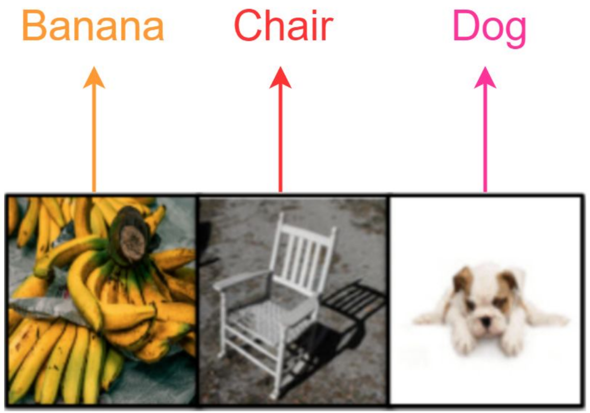
SHIRLEY ROBOTICS

Problem Statement

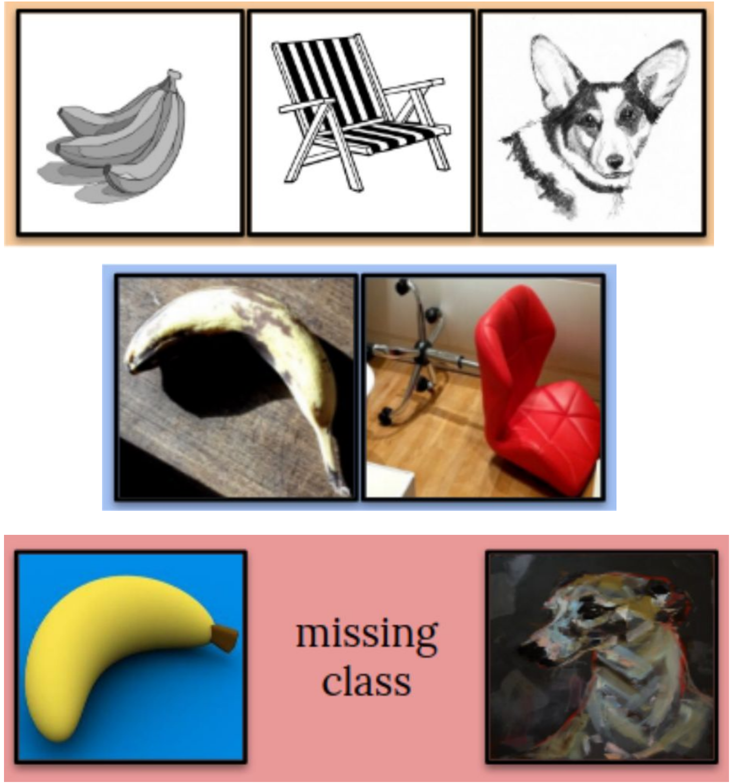


Constrained Environment

Problem Statement



Constrained Environment



Unconstrained Environment

Current Solutions

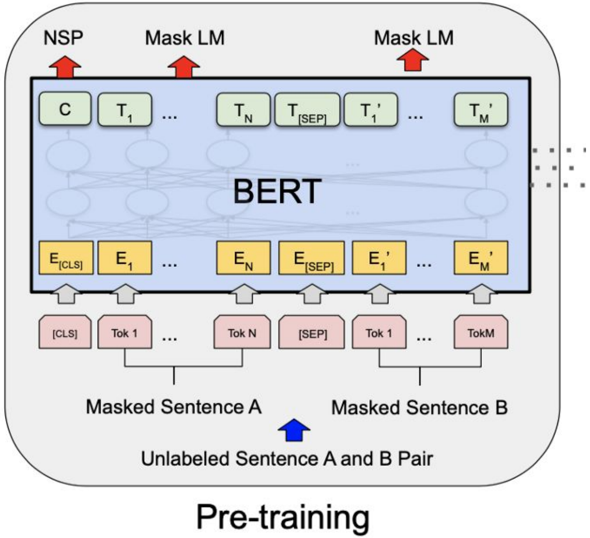
Aligning the distributions of source and target domains by learning domain-invariant representations by either

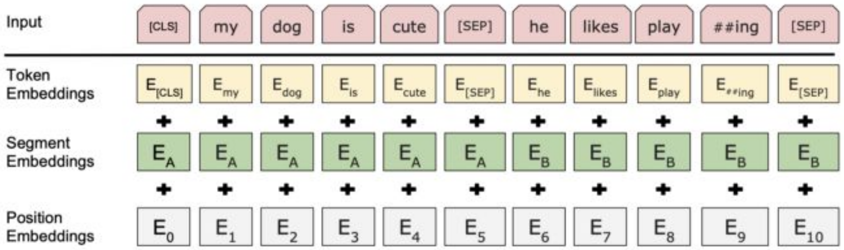
1. Moment Alignment Methods (max mean discrepancy [1, 2], second-order correlation [3, 4] or other distance metrics calculated on task-specific representations)
2. Adversarial Learning Methods (ADDA [5], CyCADA [6], MCD [7], CDAN [8] and GVB [9])

References:

1. Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In International conference on machine learning, pages 97–105, 2015.
2. Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 2208–2217. JMLR. org, 2017.
3. Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In European Conference on Computer Vision, pages 443–450. Springer, 2016
4. Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. Deep unsupervised convolutional domain adaptation. In Proceedings of the 25th ACM international conference on Multimedia, pages 261–269. ACM, 2017.
5. Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7167–7176, 2017.
6. Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In International conference on machine learning, pages 1989–1998, 2018.
7. Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3723–3732, 2018.
8. Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In Advances in Neural Information Processing Systems, pages 1647–1657, 2018.
9. Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Tian Qi. Gradually vanishing bridge for adversarial domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020

Inspiration



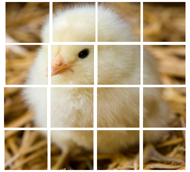


Pretraining

Original Image

x

Image Patches



224 x 224 (shape)
16 x 16 patches
14 x 14 (Each patch size)

$$\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$$

$$\downarrow$$

$$\{\boldsymbol{x}_i^p\}_{i=1}^N$$

$$\boldsymbol{x}^p \in \mathbb{R}^{N \times (P^2 C)}$$

$$N = HW / P^2$$

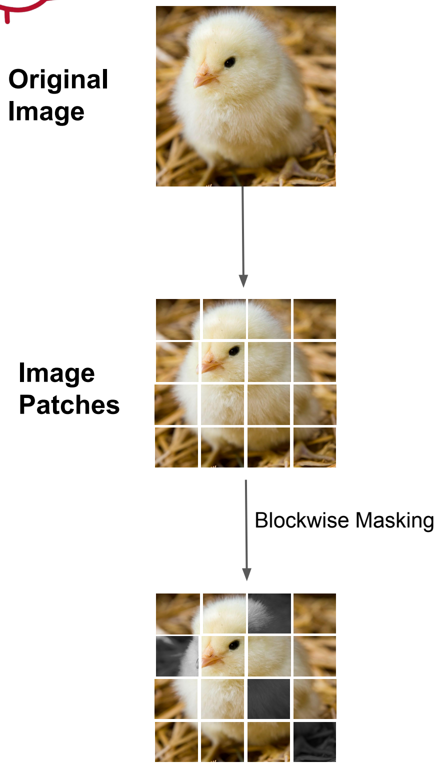
H denotes height

W denotes width

C denotes channels

P denotes resolution of each patch

Pretraining



$$\mathcal{M} \in \{1, \ldots, N\}^{0.4N}$$

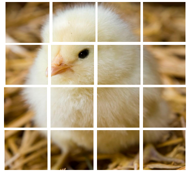
Algorithm 1 Blockwise Masking

Input: $N(= h \times w)$ image patches
Output: Masked positions $\mathcal{M}$
$\mathcal{M} \leftarrow \{\}$
repeat
  $s \leftarrow \mathrm{Rand}(16, 0.4N - |\mathcal{M}|)$ ▷ *Block size*
  $r \leftarrow \mathrm{Rand}(0.3, \frac{1}{0.3})$ ▷ *Aspect ratio of block*
  $a \leftarrow \sqrt{s \cdot r}; b \leftarrow \sqrt{s/r}$
  $t \leftarrow \mathrm{Rand}(0, h - a)\ ; l \leftarrow \mathrm{Rand}(0, w - b)$
  $\mathcal{M} \leftarrow \mathcal{M} \bigcup \{(i, j) : i \in [t, t + a), j \in [l, l + b)\}$
until $|\mathcal{M}| > 0.4N$ ▷ *Masking ratio is 40%*
return $\mathcal{M}$

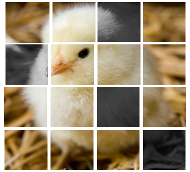
Pretraining

Original Image

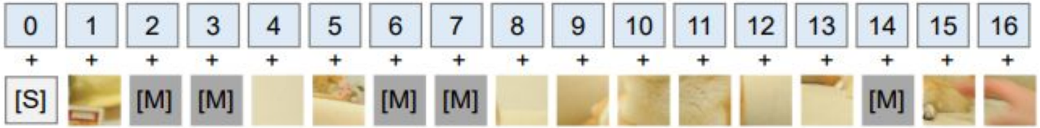
Image Patches



Blockwise Matching



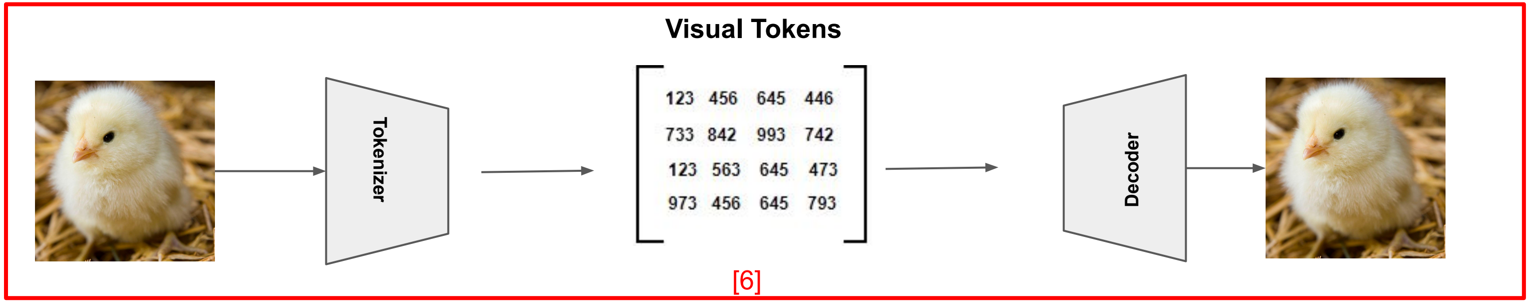
Flatten



$$\boldsymbol{E} \in \mathbb{R}^{(P^2 C) \times D}$$

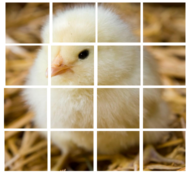
Patch Embeddings

Original Image



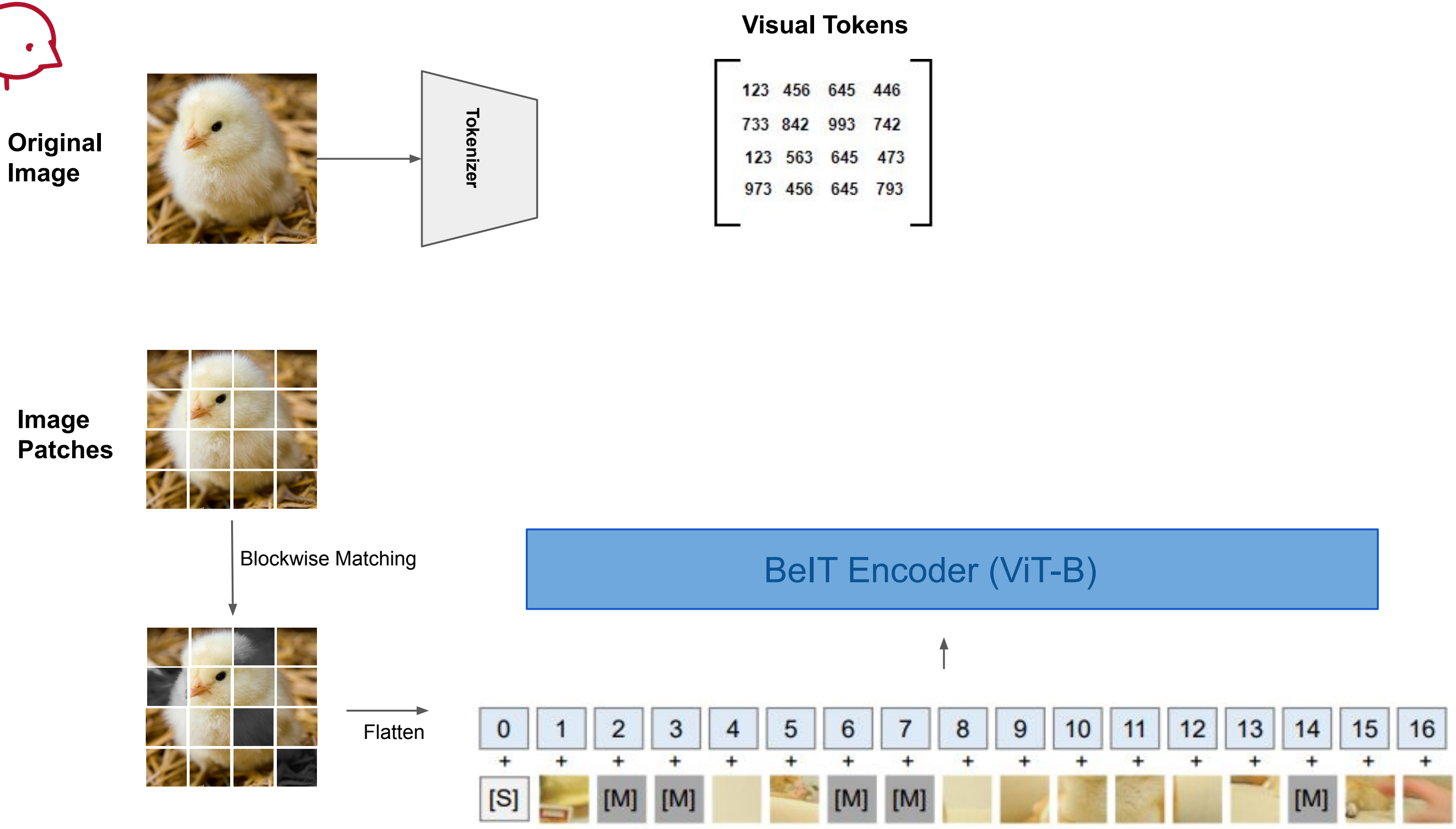
14 x 14 visual tokens
Vocabulary Size: 8192

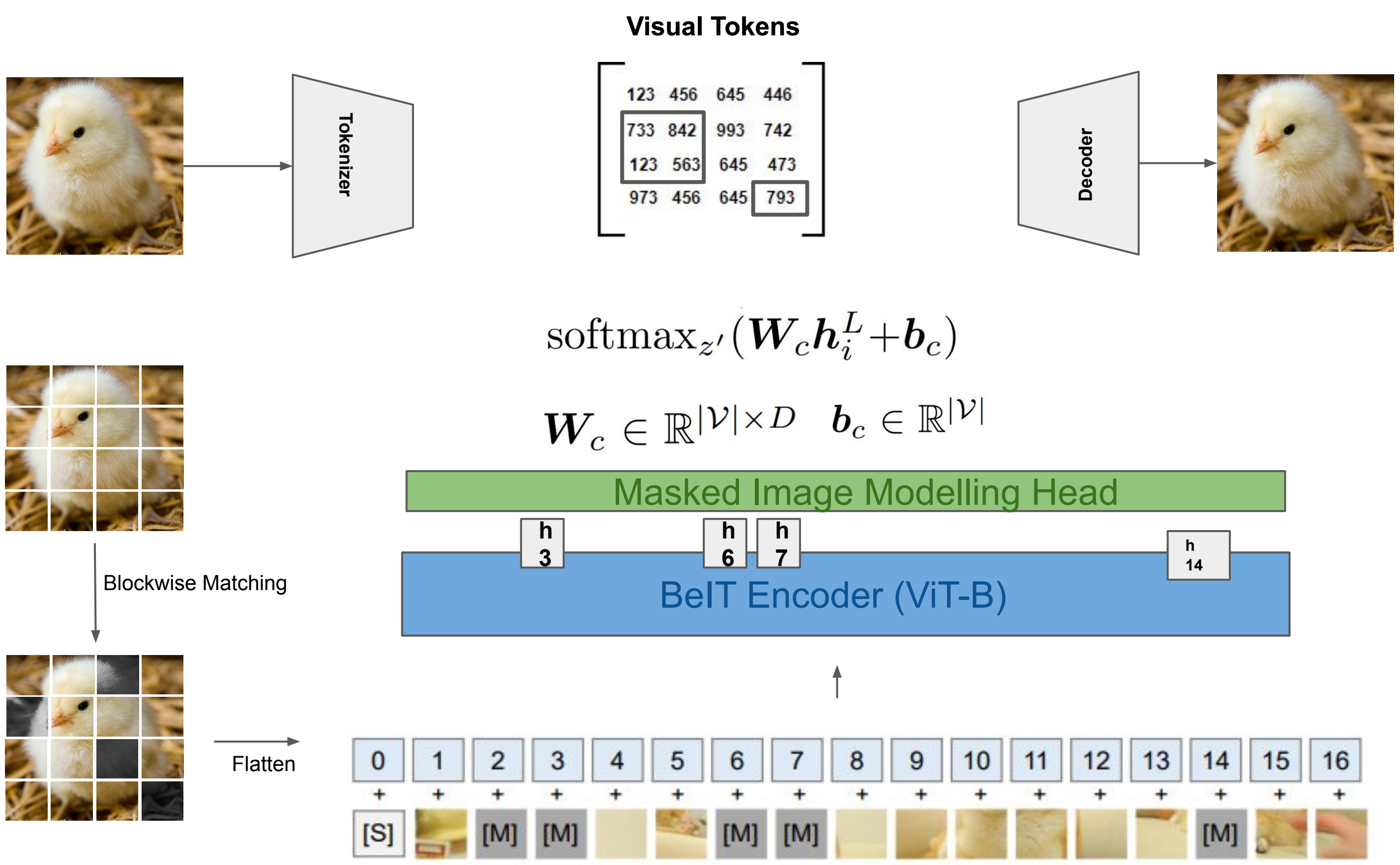
Image Patches

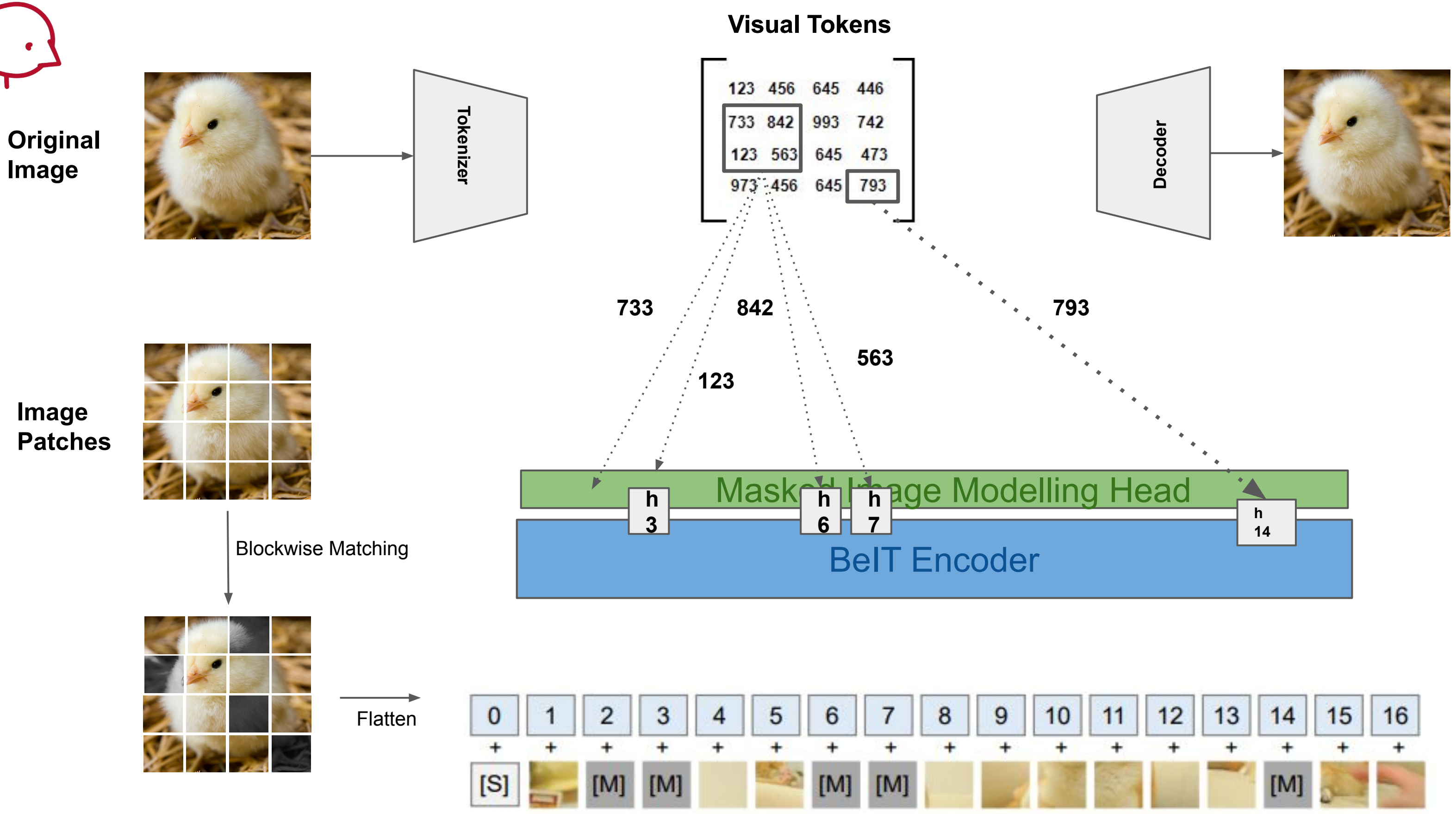


References:

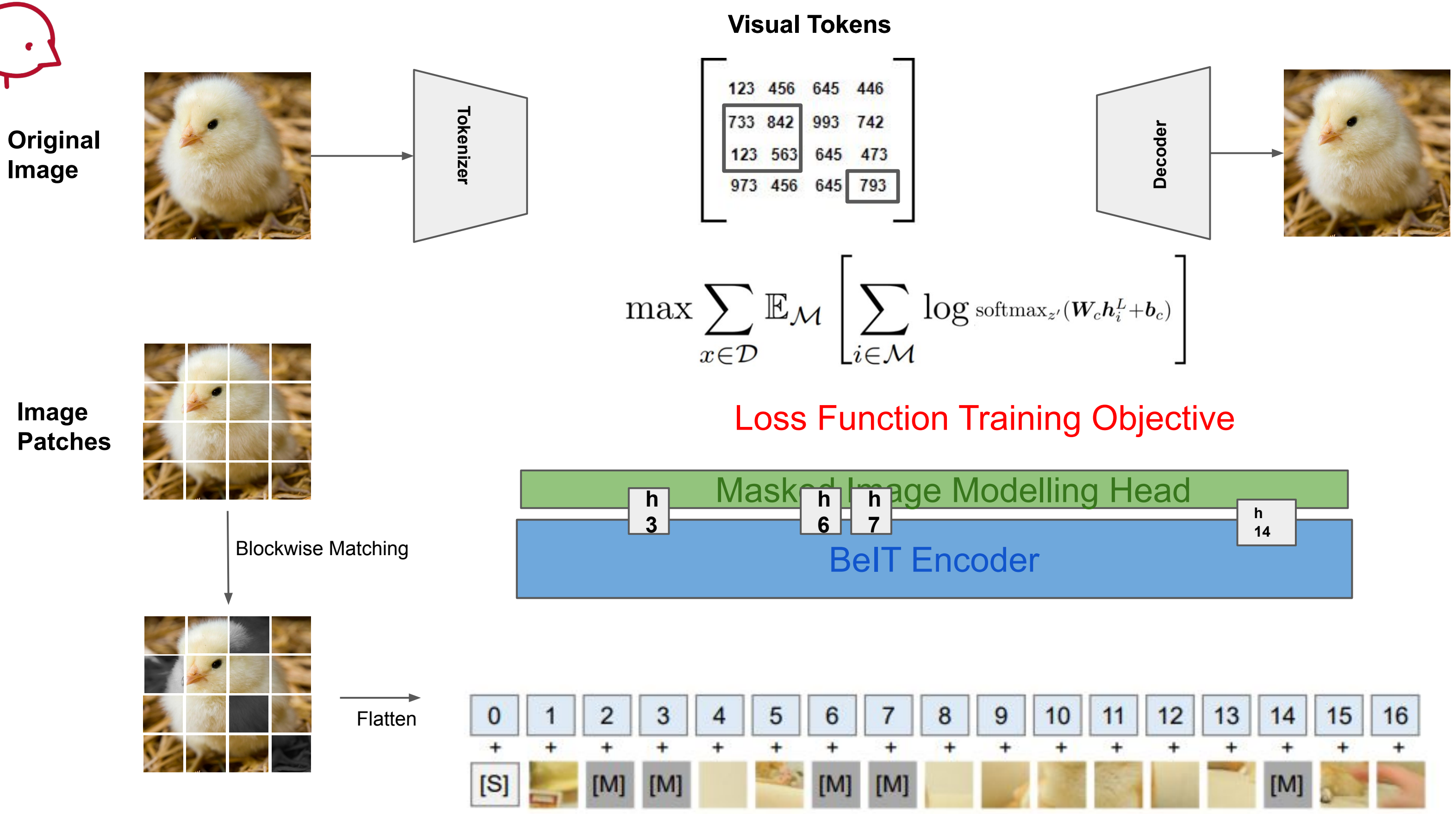
6. A. Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. ArXiv, abs/2102.12092, 2021.


Visual Tokens
Original Image
Tokenizer
123 456 645 446
733 842 993 742
123 563 645 473
973 456 645 793
Image Patches
Blockwise Matching
BeIT Encoder (ViT-B)
Flatten
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
[S] [M] [M] [M] [M] [M]


Visual Tokens
Original Image
Tokenizer
123 456 645 446
733 842 993 742
123 563 645 473
973 456 645 793
Decoder
softmax$_{z'}(\boldsymbol{W}_c \boldsymbol{h}_i^L + \boldsymbol{b}_c)$
$\boldsymbol{W}_c \in \mathbb{R}^{|\mathcal{V}| \times D}$ $\boldsymbol{b}_c \in \mathbb{R}^{|\mathcal{V}|}$
Image Patches
Masked Image Modelling Head
h 3
h 6
h 7
h 14
BeIT Encoder (ViT-B)
Blockwise Matching
Flatten
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
[S] [M] [M] [M] [M] [M]

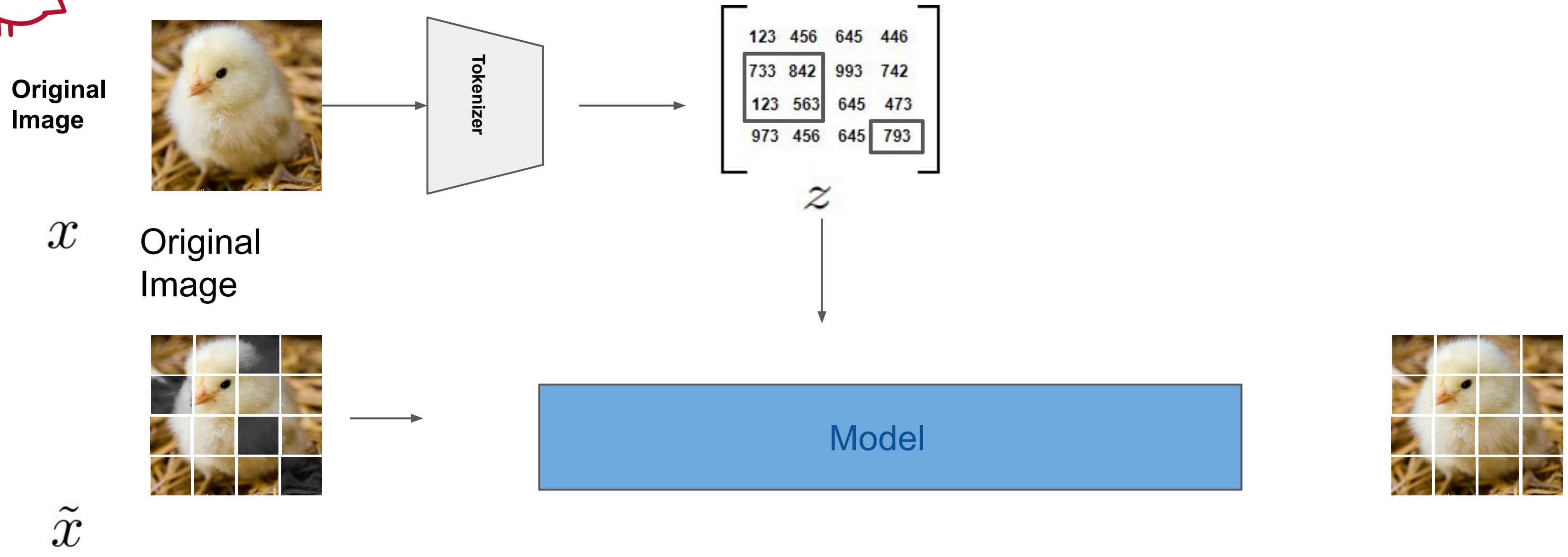


Visual Tokens
123 456 645 446
733 842 993 742
123 563 645 473
973 456 645 793
Original Image
Tokenizer
Decoder
733
842
123
563
793
Image Patches
Blockwise Matching
Masked Image Modelling Head
h 3
h 6
h 7
h 14
BeIT Encoder
Flatten
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
[S] [M] [M] [M] [M] [M]

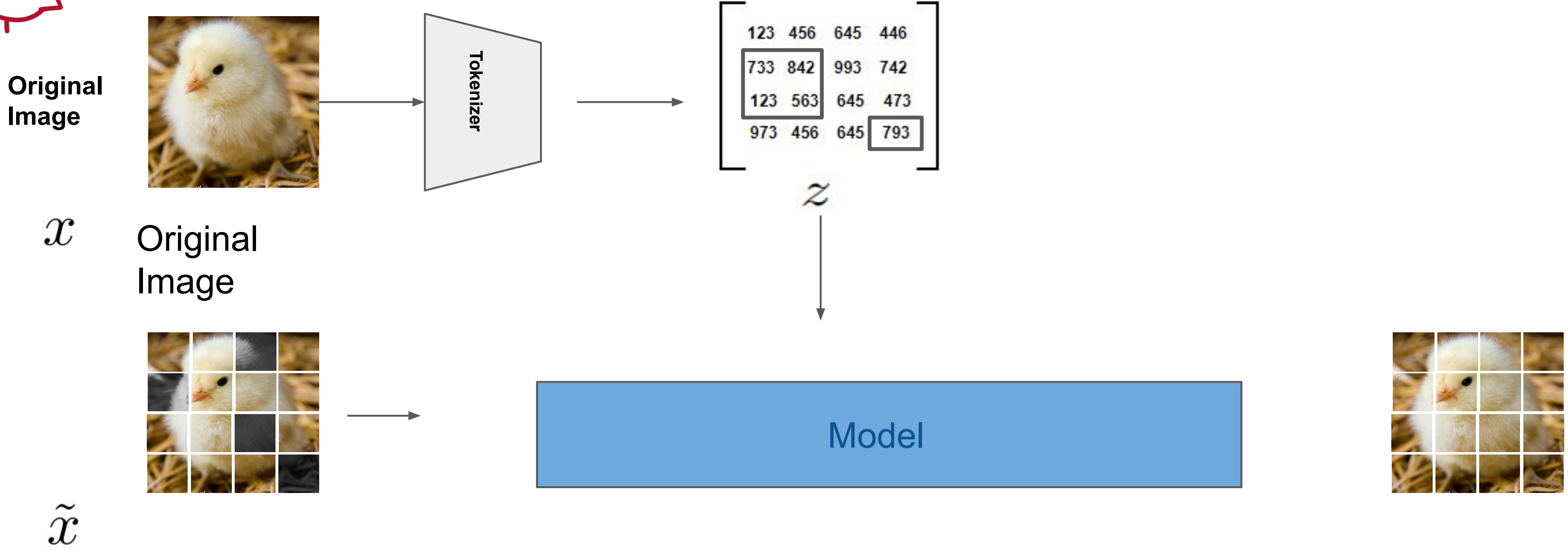


Original Image
Tokenizer
Visual Tokens
123 456 645 446
733 842 993 742
123 563 645 473
973 456 645 793
Decoder
$\max \sum_{x \in \mathcal{D}} \mathbb{E}_{\mathcal{M}} \left[\sum_{i \in \mathcal{M}} \log \mathrm{softmax}_{z'}(\boldsymbol{W}_c \boldsymbol{h}_i^L + \boldsymbol{b}_c) \right]$
Loss Function Training Objective
Image Patches
Blockwise Matching
Masked Image Modelling Head
h 3
h 6
h 7
h 14
BeIT Encoder
Flatten
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
[S] [M] [M] [M] [M] [M]

Pretraining

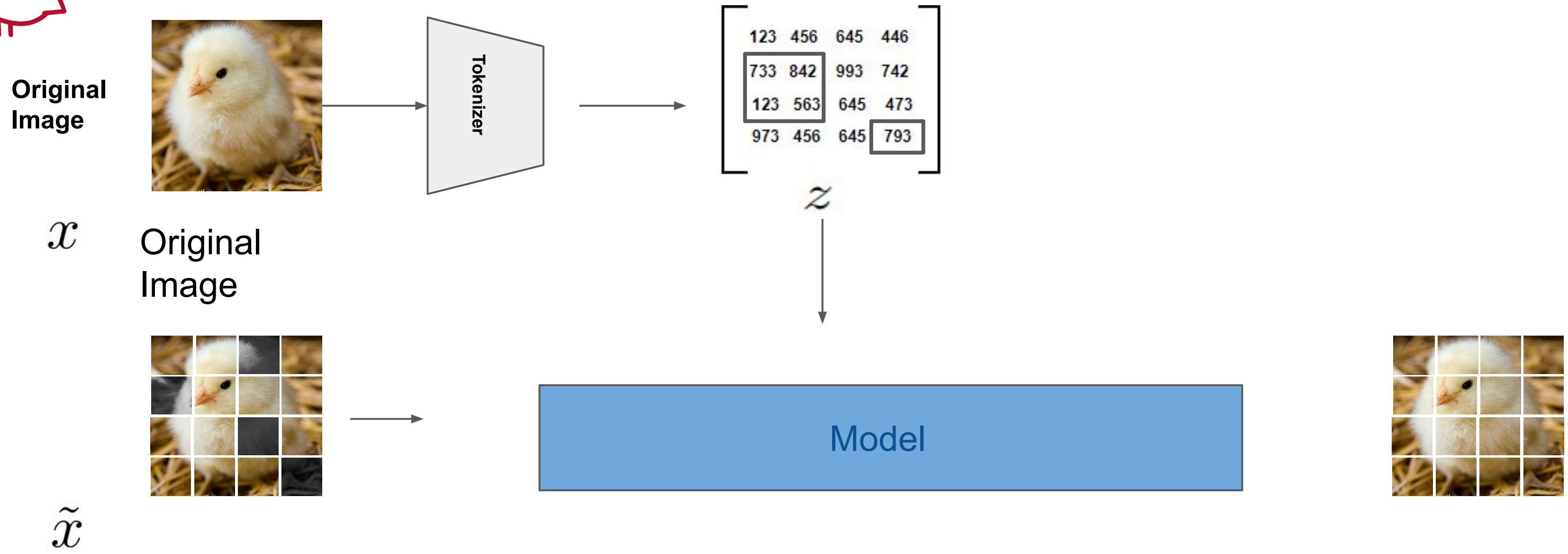


Pretraining



$$\sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \log p(x_i|\tilde{x}_i) \geq \sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \boxed{\Big(\mathbb{E}_{z_i\sim q_\phi(\mathbf{z}|x_i)}[\log p_\psi(x_i|z_i)]} - D_{\mathrm{KL}}[q_\phi(\mathbf{z}|x_i), p_\theta(\mathbf{z}|\tilde{x}_i)]\Big)$$

Pretraining

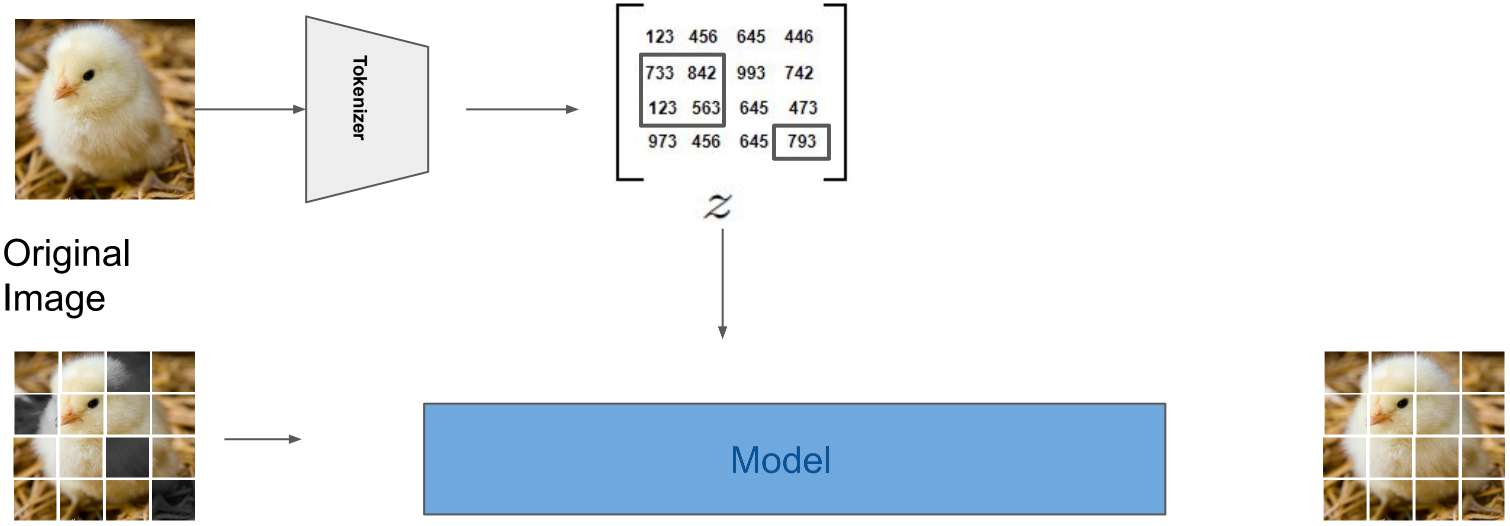


$$\sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \log p(x_i|\tilde{x}_i) \geq \sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \left(\mathbb{E}_{z_i\sim q_\phi(\mathbf{z}|x_i)}[\log p_\psi(x_i|z_i)] - D_{\mathrm{KL}}[\boxed{q_\phi(\mathbf{z}|x_i)}, p_\theta(\mathbf{z}|\tilde{x}_i)]\right)$$

Pretraining

Original Image

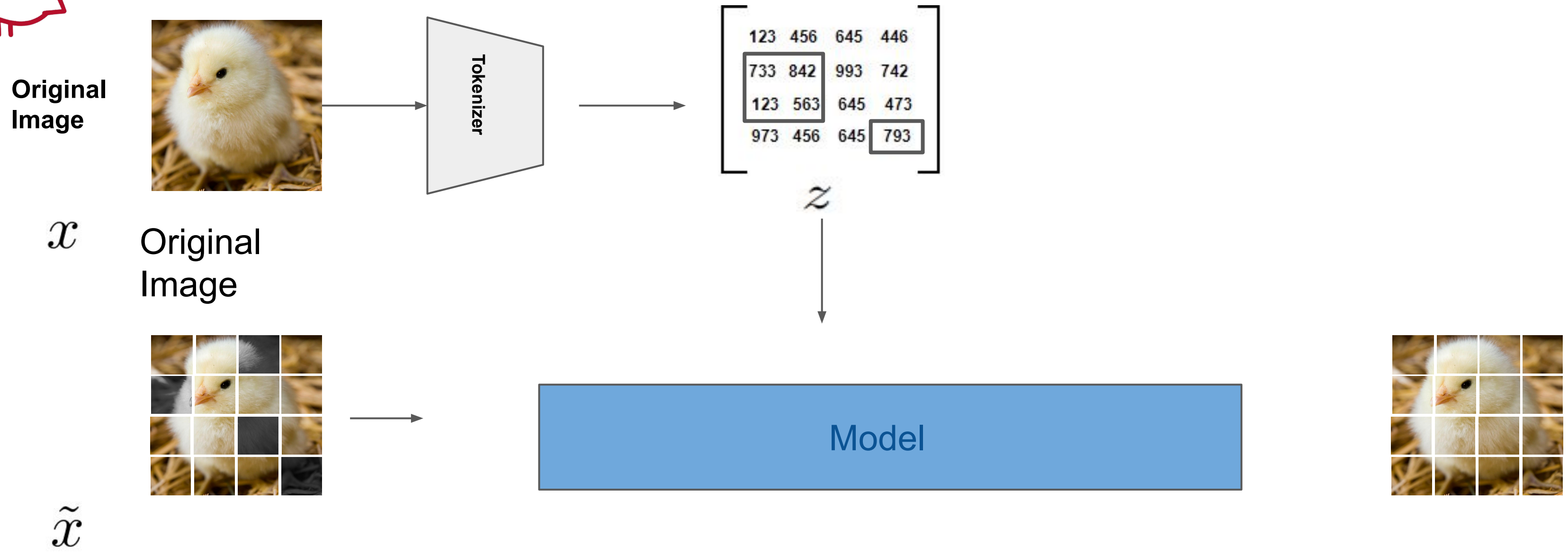
x



$\tilde{x}$

$$\sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \log p(x_i|\tilde{x}_i) \geq \sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \left(\mathbb{E}_{z_i\sim q_\phi(\mathbf{z}|x_i)}[\log p_\psi(x_i|z_i)] - D_{\mathrm{KL}}[q_\phi(\mathbf{z}|x_i), \boxed{p_\theta(\mathbf{z}|\tilde{x}_i)}]\right)$$

Pretraining

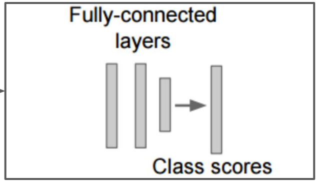


$$\sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \Big(\underbrace{\mathbb{E}_{z_i\sim q_\phi(z|x_i)}[\log p_\psi(x_i|z_i)]}_{\text{Stage 1: Visual Token Reconstruction}} + \underbrace{\log p_\theta(\hat{z}_i|\tilde{x}_i)}_{\text{Stage 2: Masked Image Modeling}} \Big)$$

Finetuning

Image

Model



$$\mathrm{softmax}(\mathrm{avg}(\{\boldsymbol{h}_i^L\}_{i=1}^N \boldsymbol{W}_c))$$

output: black_baby_chick (83.4%)

Parameters

Augmentation	Color Jitter, Horizontal Flipping, and Random Resized Cropping
Dataset	ImageNet-1k
Batch Size	1000
Epochs (Pretraining)	500
Epochs (Fine-tuning)	500
Learning Rate	1.5e-3 with cosine learning decay
Optimizer	Adam with B1: 0.9 and B2: 0.999
Parameters	86M
Image Size	224 x 224 (Pretraining); 384 × 384(Last 100 Epochs Finetuning)
Architecture	ViT-Base (12 layer transformer, 768 hidden size, 12 attention heads)

Dataset

Dataset	Number of Images	Number of Classes	Note*
ImageNet(source)	1.4M	1000	
ObjectNet	50,000	313	Only 113 classes are the same as the source
ImageNet-R	30,000	200	Different texture/style
ImageNet-C	1.4M	1000	Corrupted

Accuracy

Methods	Parameters	ACC
ImageNet-1K(pretraining)	86M	82.4%
ImageNet-1K(pretraining + fine-tuning)	86M	83.0%

Methods	ACC(Adapted Model)	AUC(Adapted Model)	ACC(Source model)	AUROC(Source model)
Babychick(ours)	56.29	69.79	56.29	69.79
liaohaojin	48.56	70.72	41.25	64.48
chamorajg	48.49	76.86	0.07	50.00
DXM-DI-AI-CV-TEAM	48.60	68.29	25.70	62.43
fomenxiaoseng	45.23	78.76	40.22	60.43

Results

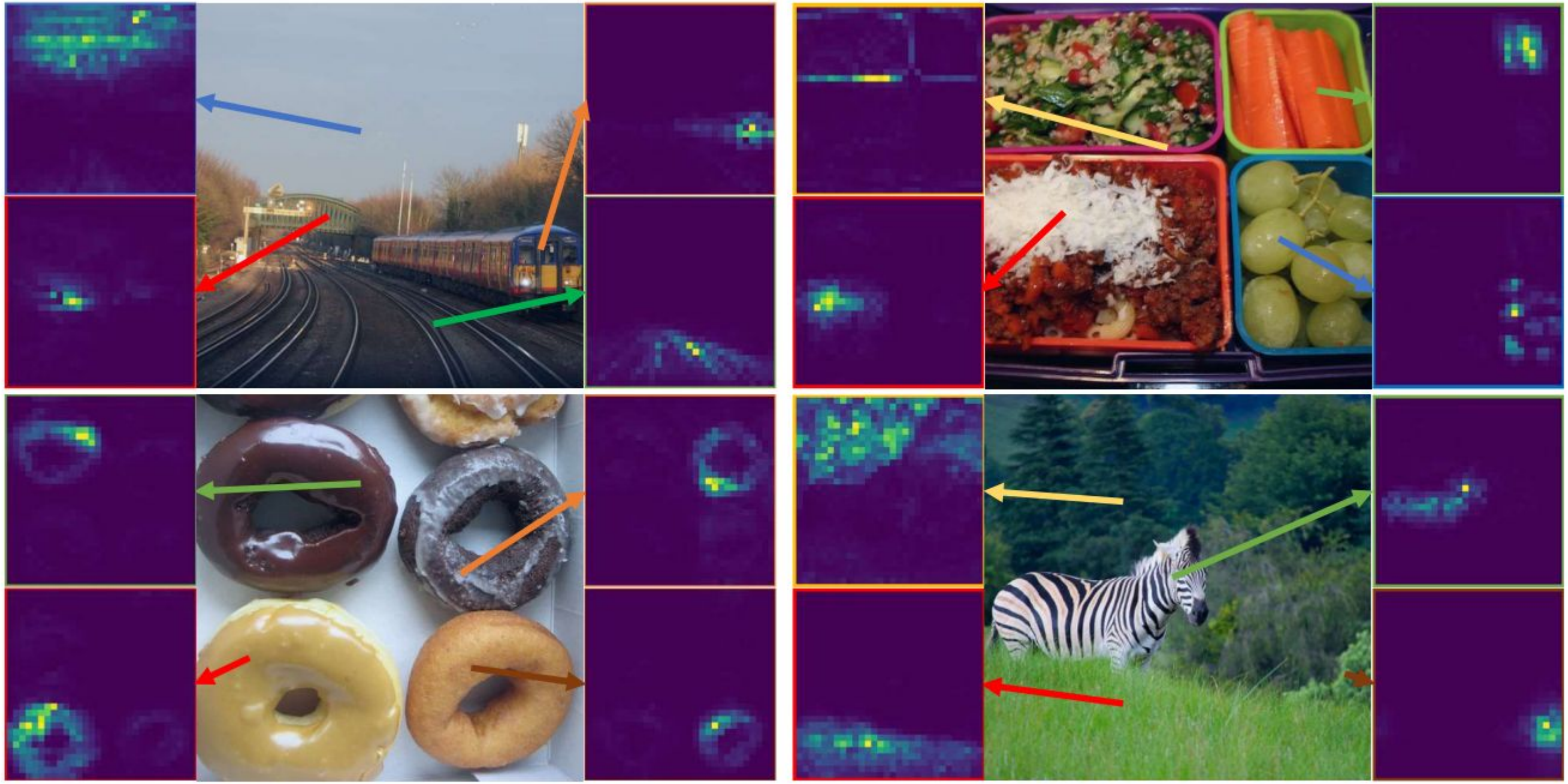


Image taken from:

7. A. Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. ArXiv, abs/2102.12092, 2021.

Results

Future Works

- Multi-attribute domain adaptation (where source and target datasets have almost no correlation)
- Vision Reconstruction in Humans / Animals via fMRI data (partially achieved)
- Reaching OpenAI's CLIP moment for Object Detection

Thanks

Questions?