Deep Denoising Models for Visual Representation Learning

Mido Assran

Meta

Representation Learning

Paper: https://arxiv.org/pdf/1206.5538.pdf

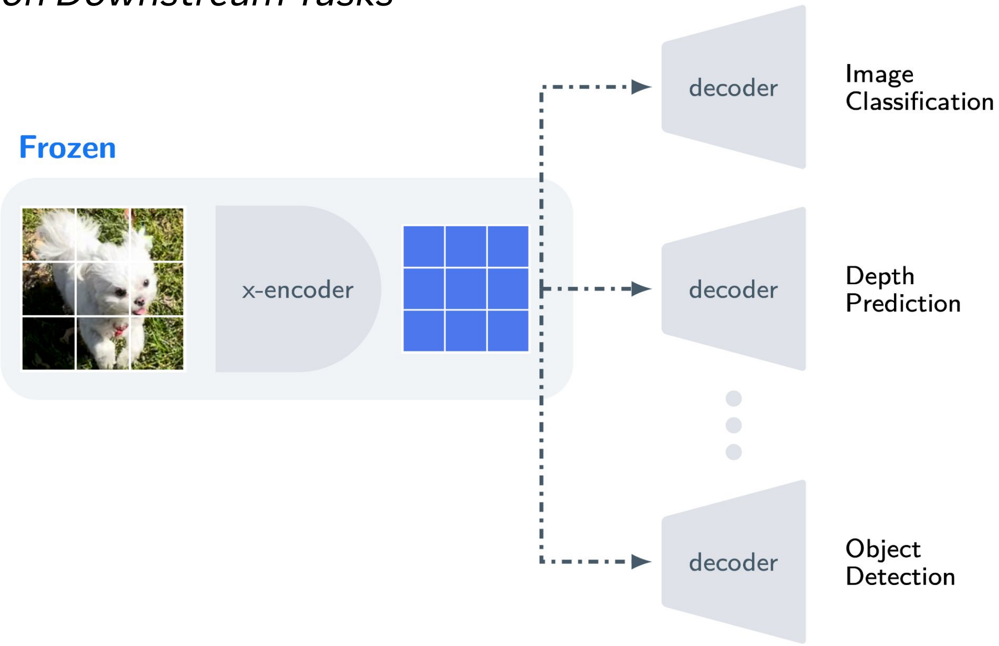
Pretraining:
How can we train a neural network to extract semantic features from unstructured data?

Evaluation:
How do we measure the quality of the learned features?

Visual Representation Learning

Evaluation:
How do we measure the quality of the learned features?

Frozen Evaluation on Downstream Tasks



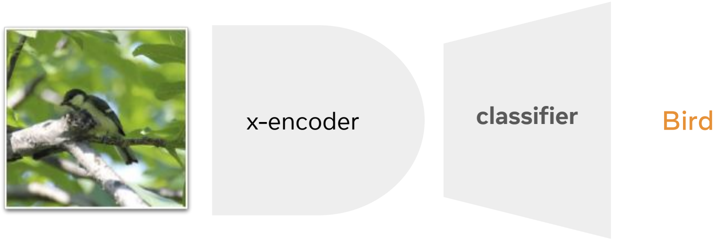
Visual Representation Learning

Pretraining:
How can we train a neural network to extract semantic features from unstructured data?

Supervised Learning

Labeled Examples

Train the image encoder by classifying labeled images

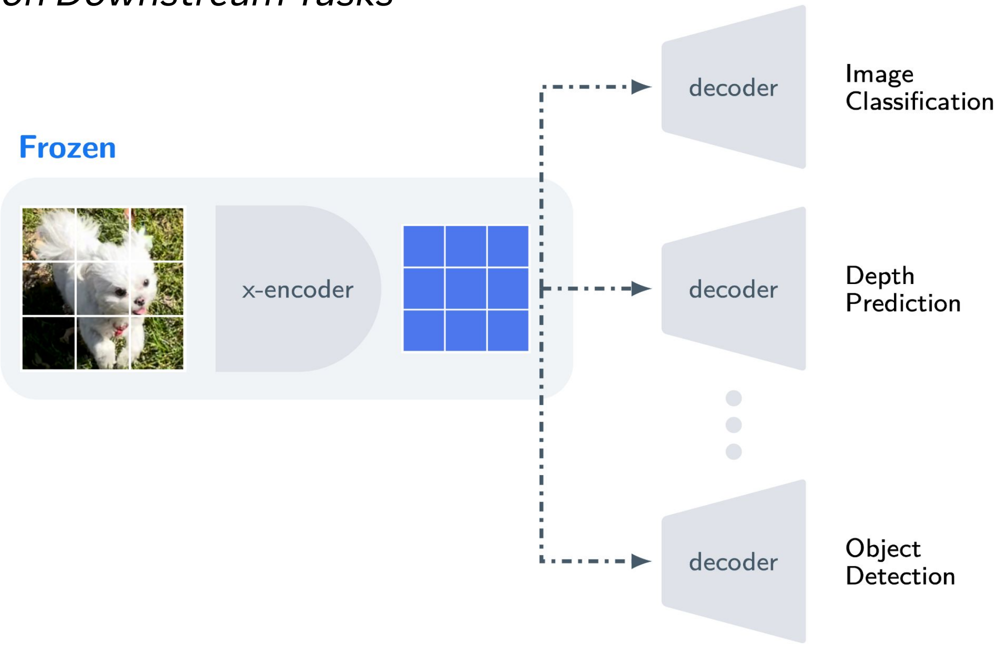


Meta

Visual Representation Learning

Evaluation:
How do we measure the quality of the learned features?

Frozen Evaluation on Downstream Tasks



Visual Representation Learning

Pretraining:
How can we train a neural network to extract semantic features from unstructured data?

Supervised Learning

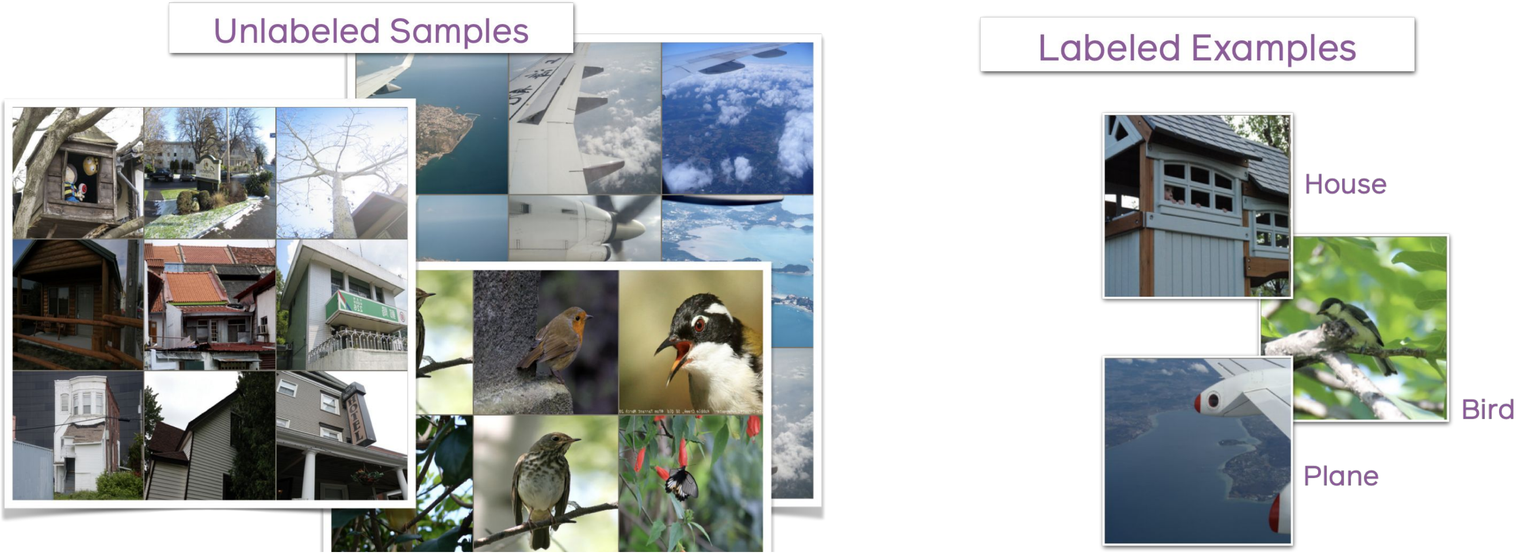
Limitations:

- Need lots of human annotated data (expensive)
- Representations that are best for image classification are not necessarily the best for other tasks

Visual Representation Learning

Pretraining:
How can we train a neural network to extract semantic features from unstructured data?

Semi-Supervised Learning



Visual Representation Learning

Pretraining:
How can we train a neural network to extract semantic features from unstructured data?

Semi-Supervised Learning

Many previous approaches train the encoder parameters Θ by minimizing a weighted sum of a supervised loss and an unsupervised loss, where $\lambda > 0$ is the relative weighting between the two losses

$$\text{minimize}_\theta \quad \ell(\theta; D_U, D_S) = \ell_{\text{unsupervised}}(\theta; D_U) + \lambda \ell_{\text{supervised}}(\theta; D_S)$$

Visual Representation Learning

<u>Pretraining:</u>
How can we train a neural network to extract semantic features from unstructured data?

Semi-Supervised Learning

Many previous approaches train the encoder parameters θ by minimizing a weighted sum of a supervised loss and an unsupervised loss, where $\lambda > 0$ is the relative weighting between the two losses

$$\text{minimize}_\theta \quad \ell(\theta; D_U, D_S) = \ell_{\text{unsupervised}}(\theta; D_U) + \lambda \ell_{\text{supervised}}(\theta; D_S)$$

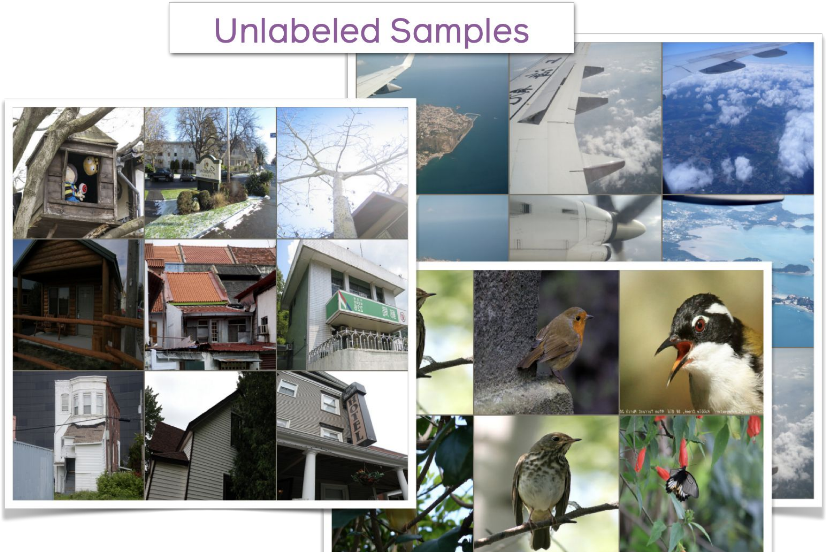
Limitations:

- Tend to overfit without enough labeled examples
- Representations that are best for image classification are not necessarily the best for other tasks

Visual Representation Learning

Pretraining:
How can we train a neural network to extract semantic features from unstructured data?

Unsupervised Learning



Meta

Visual Representation Learning

<u>Pretraining:</u>
How can we train a neural network to extract semantic features from unstructured data?

Unsupervised Learning

Underlying Hypothesis: *There exist “proxy tasks” such that an encoder trained to solve such tasks on unlabeled data has learned to produce effective visual representations.*

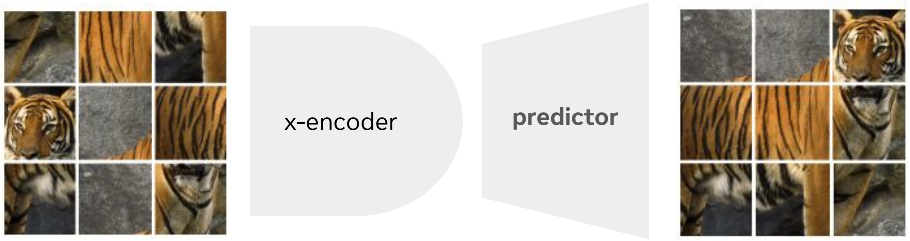
Visual Representation Learning

<u>Pretraining:</u>
How can we train a neural network to extract semantic features from unstructured data?

Unsupervised Learning

Underlying Hypothesis: *There exist "proxy tasks" such that an encoder trained to solve such tasks on unlabeled data has learned to produce effective visual representations.*

Train the image encoder by solving Jigsaw Puzzels



Paper: https://arxiv.org/pdf/1603.09246.pdf

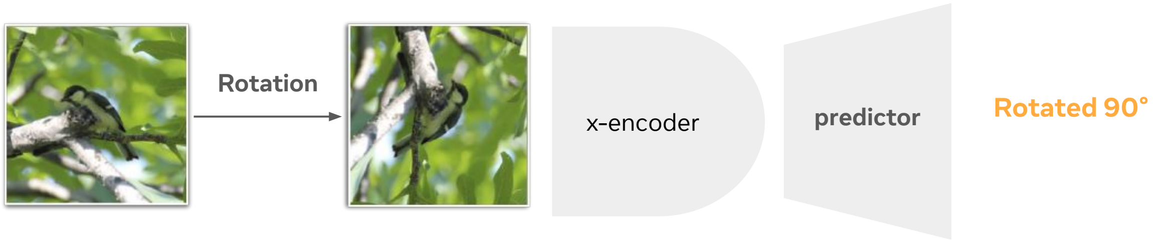
Visual Representation Learning

<u>Pretraining:</u>
How can we train a neural network to extract semantic features from unstructured data?

Unsupervised Learning

Underlying Hypothesis: *There exist "proxy tasks" such that an encoder trained to solve such tasks on unlabeled data has learned to produce effective visual representations.*

Train the image encoder by predicting image rotations



Paper: https://arxiv.org/pdf/1803.07728.pdf

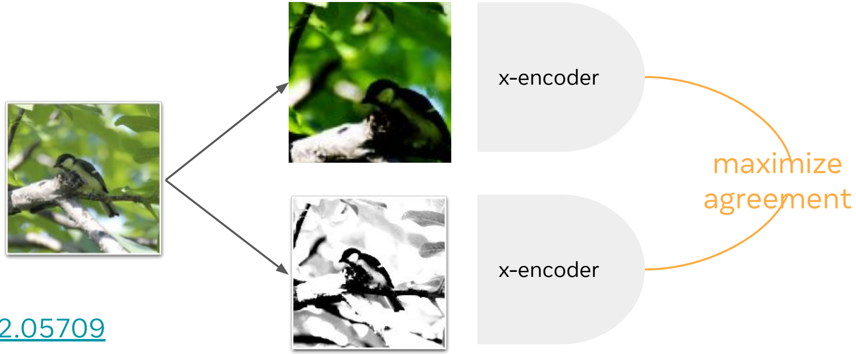
Visual Representation Learning

<u>Pretraining:</u>
How can we train a neural network to extract semantic features from unstructured data?

Unsupervised Learning

Underlying Hypothesis: *There exist "proxy tasks" such that an encoder trained to solve such tasks on unlabeled data has learned to produce effective visual representations.*

Train the image encoder by enforcing invariance to data augmentations



Paper: https://arxiv.org/abs/2002.05709

Meta

Visual Representation Learning

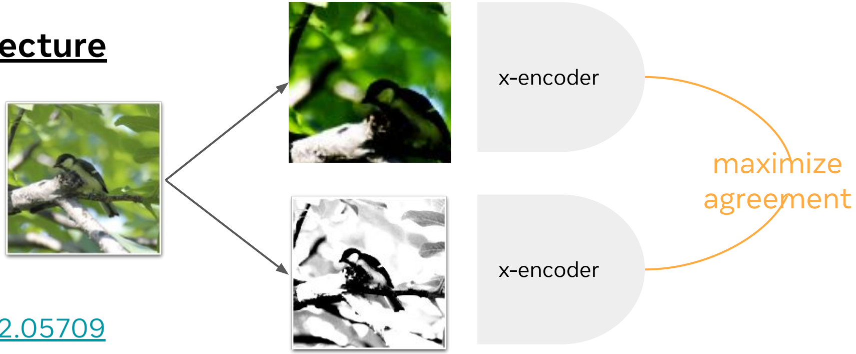
<u>Pretraining:</u>
How can we train a neural network to extract semantic features from unstructured data?

Unsupervised Learning

Underlying Hypothesis: *There exist "proxy tasks" such that an encoder trained to solve such tasks on unlabeled data has learned to produce effective visual representations.*

Train the image encoder by enforcing invariance to data augmentations

<u>Joing-Embedding Architecture</u>



Paper: https://arxiv.org/abs/2002.05709

Visual Representation Learning

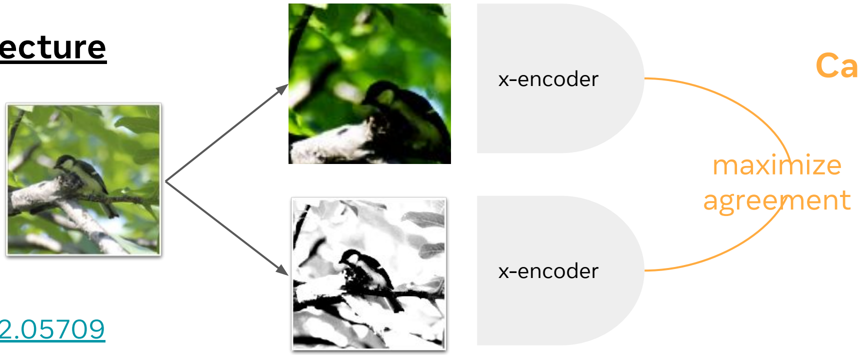
<u>Pretraining:</u>
How can we train a neural network to extract semantic features from unstructured data?

Unsupervised Learning

Underlying Hypothesis: *There exist “proxy tasks” such that an encoder trained to solve such tasks on unlabeled data has learned to produce effective visual representations.*

Train the image encoder by enforcing invariance to data augmentations

<u>Joing-Embedding Architecture</u>



Can anything go wrong here?

Paper: https://arxiv.org/abs/2002.05709

Visual Representation Learning

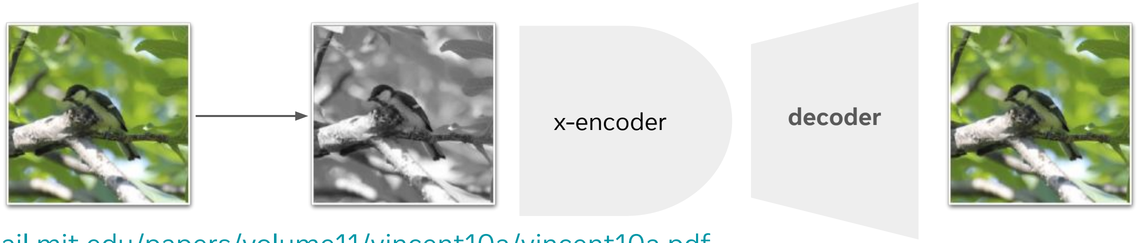
<u>Pretraining:</u>
How can we train a neural network to extract semantic features from unstructured data?

Unsupervised Learning

Underlying Hypothesis: *There exist "proxy tasks" such that an encoder trained to solve such tasks on unlabeled data has learned to produce effective visual representations.*

Train the image encoder by reconstructing/denoising corrupted images

<u>Generative Architecture</u>



Paper: https://jmlr.csail.mit.edu/papers/volume11/vincent10a/vincent10a.pdf

Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

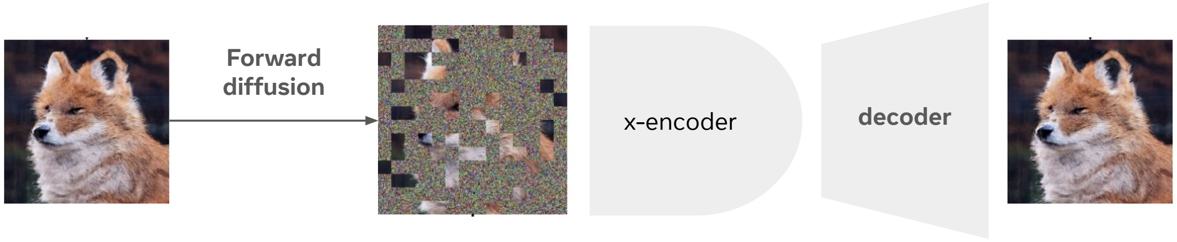
Pretraining:

How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

Train the image encoder by reconstructing/denoising corrupted images

Generative Architecture



Representation Learning by Denoising Pixels

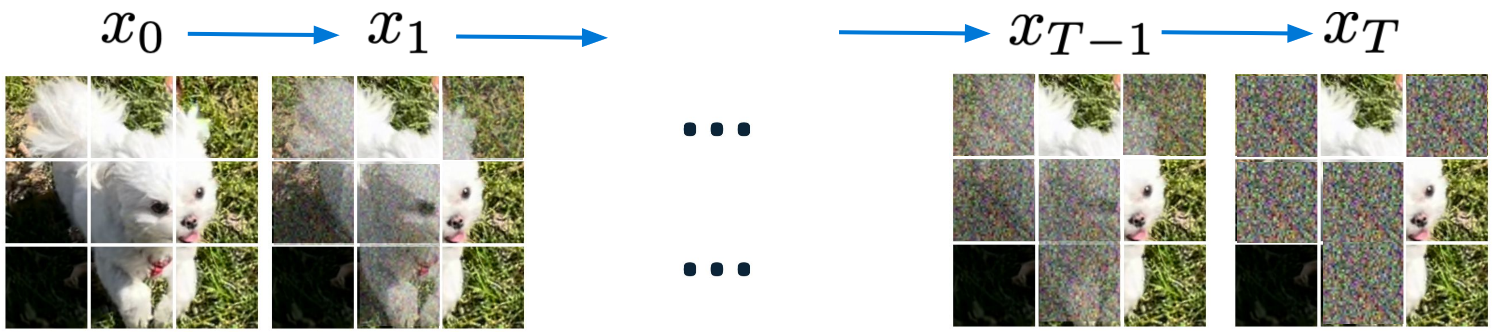
Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

Generative Architecture

1. Divide image into "visible regions" and "masked regions"
2. Forward diffusion process adds Gaussian Noise to masked regions in each step

Forward Diffusion Process

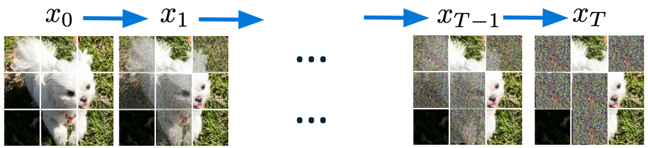


Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

Generative Architecture



1. Divide image into "visible regions" and "masked regions"
2. Forward diffusion process adds Gaussian Noise to masked regions in each step

Similar to traditional diffusion models… forward process specified by Markov Process

$$p(x_t^m | x_{t-1}^m) = \mathcal{N}(x_t^m; \sqrt{1-\beta_t} x_{t-1}^m, \beta_t \mathbf{I})$$

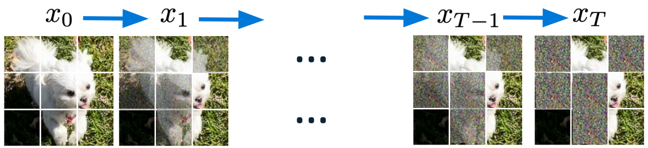
where the superscript *m* denotes a masked region of the image.

Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

Generative Architecture



1. Divide image into "visible regions" and "masked regions"
2. Forward diffusion process adds Gaussian Noise to masked regions in each step

Samples from forward distribution are also Gaussian, and can be sampled without recursion:

$$p(x_t^m | x_0^m) = \mathcal{N}(x_t^m; \sqrt{\bar{\alpha}_t} x_0^m, (1 - \bar{\alpha}_t \mathbf{I})$$

Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

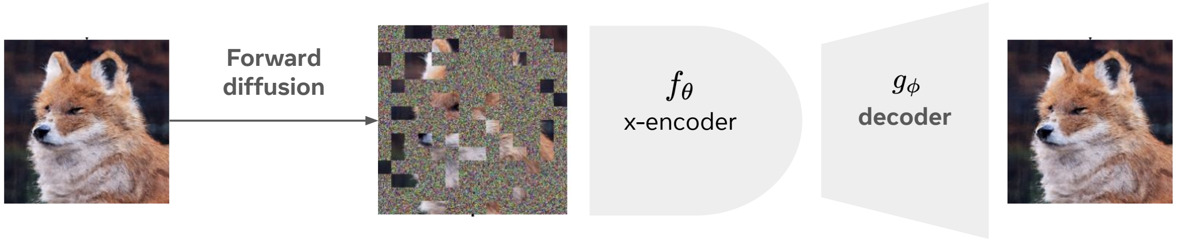
Pretraining:

How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

Train the image encoder by reconstructing/denoising corrupted images

Generative Architecture



Representation Learning by Denoising Pixels

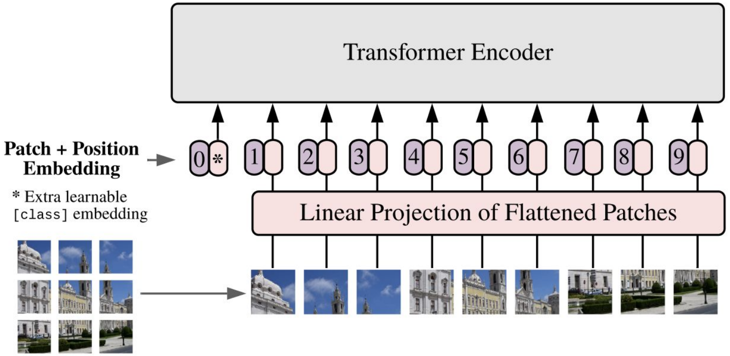
Paper: https://arxiv.org/pdf/2304.03283.pdf

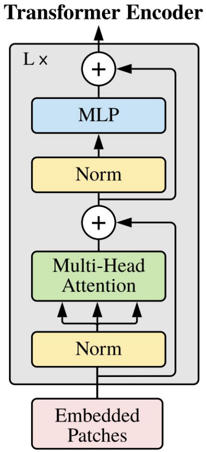
Pretraining:

How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

Encoder is a Vision Transformer





Representation Learning by Denoising Pixels

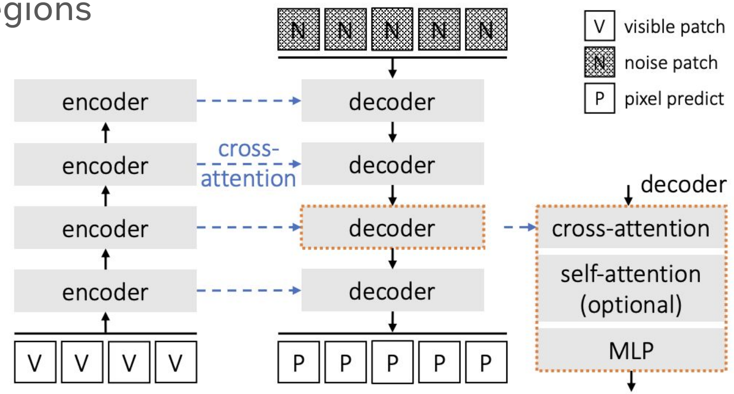
Paper: https://arxiv.org/pdf/2304.03283.pdf

Pretraining:
How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

* For efficiency, encoder f_θ only processes visible regions
* Decoder g_ϕ processes masked regions and visible regions

$$\ell(x;\theta,\phi) = \|x_0^m - g_\phi(x_t^m, f_\theta(x_0^v))\|$$

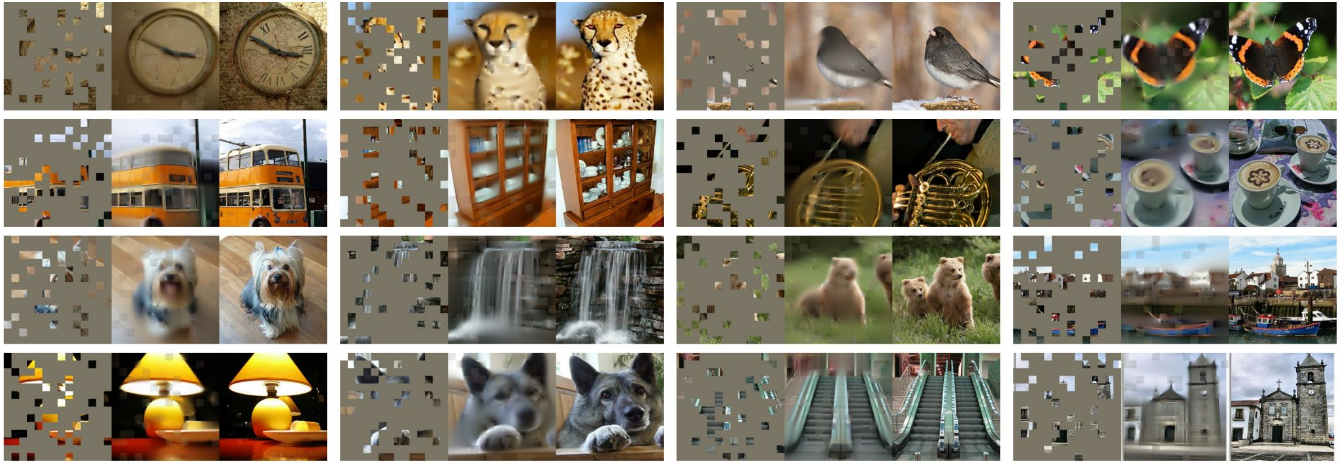


Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Pretraining:
How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**



Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Pretraining:
How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

Encoder learns effective representations for downstream image classification on the ImageNet-1K benchmark

pre-train	w/ CLIP	ViT-B	ViT-L	ViT-H
from-scratch [34]	×	82.3	82.6	83.1
MoCo v3 [11]	×	83.2	84.1	-
DINO [7]	×	82.8	-	-
iBOT [97]	×	84.0	84.8	-
BEiT [3]	×	83.2	85.2	-
MaskFeat [87]	×	84.0	85.7	-
MAE [34]	×	83.6	85.9	**86.9**
DiffMAE	×	83.9	85.8	**86.9**

Generalized Noise Patterns: Mask Noise

Paper: https://arxiv.org/pdf/2304.03283.pdf

Pretraining:

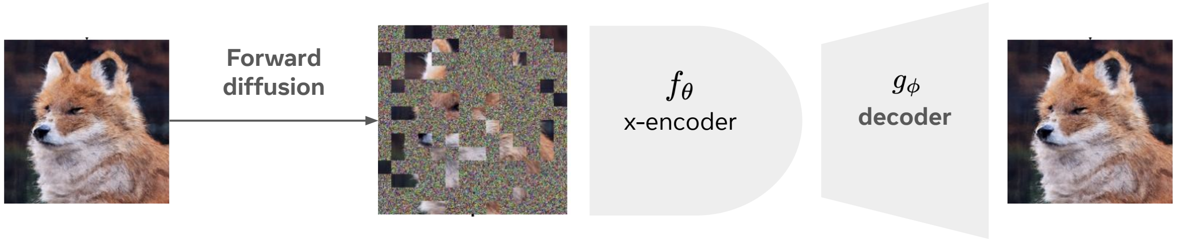
How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Gaussian Noise**

Train the image encoder by reconstructing/denoising corrupted images

Generative Architecture

Do we really need to use Gaussian Noise in Forward Process?



Meta

Generalized Noise Patterns: Mask Noise

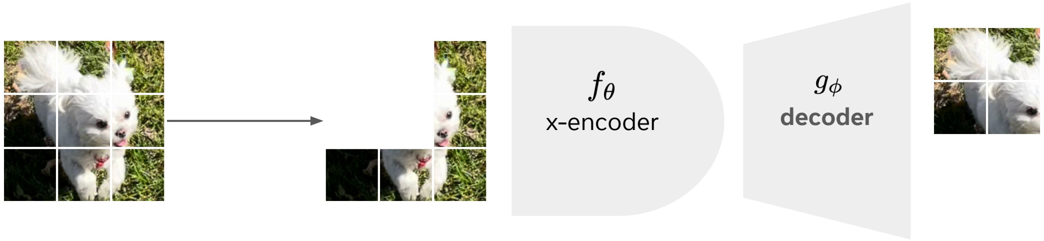
Paper: https://arxiv.org/pdf/2111.06377.pdf

Pretraining:
How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Mask Noise**

Train the image encoder by reconstructing missing patches

Generative Architecture



Meta

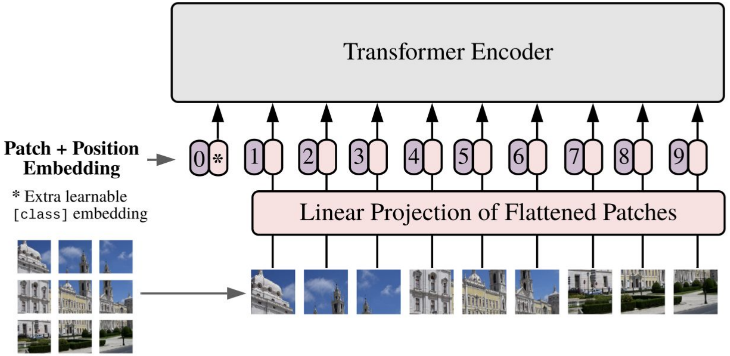
Generalized Noise Patterns: Mask Noise

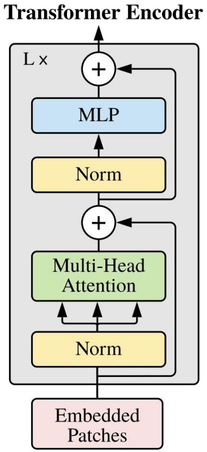
Paper: https://arxiv.org/pdf/2111.06377.pdf

Pretraining:
How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Mask Noise**

Encoder is still a Vision Transformer (processes sequence of patches)





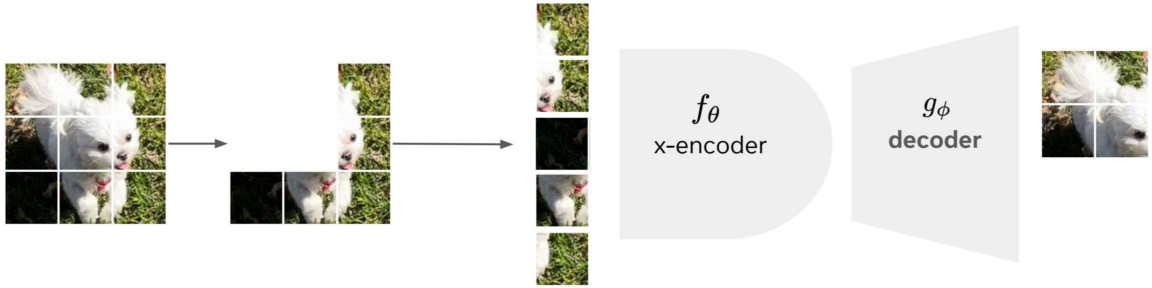
Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise**

<u>Generative Architecture</u>

1. Divide image into a sequence of patches
2. Split seq. into "visible regions" and "masked regions"
3. Drop masked patches from the input sequence



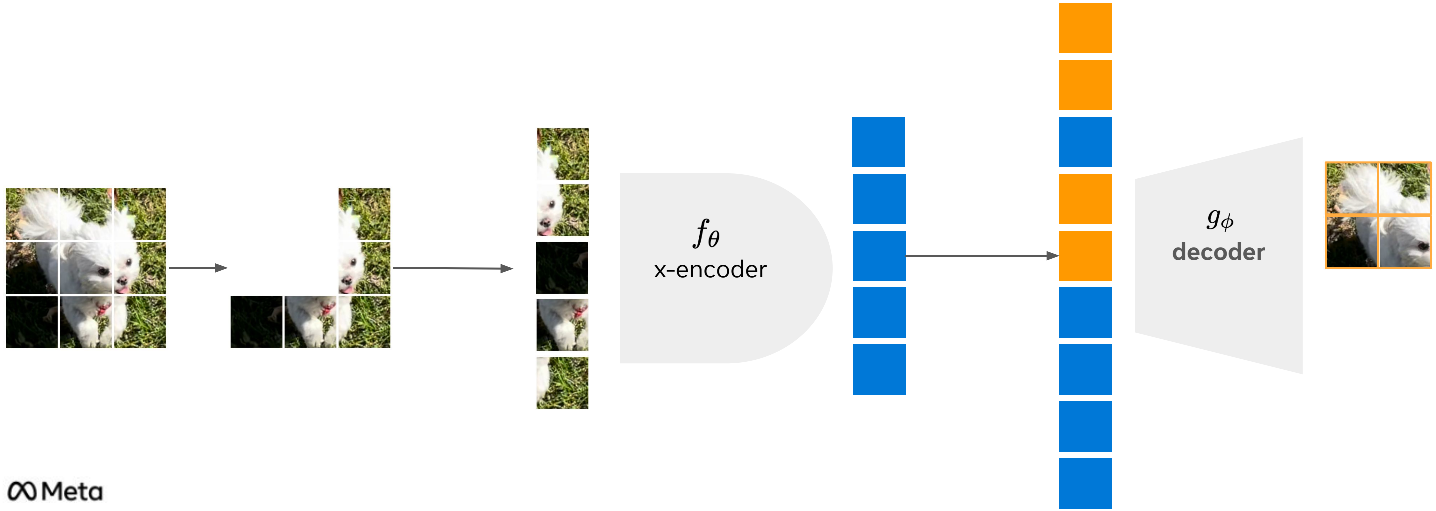
Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise**

Generative Architecture

Decoder takes "mask tokens" and patch representations to predict pixels of missing regions



Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise**

Generative Architecture

Loss is just L2 distance between predicted pixels and (normalized) ground truth pixels,

$$\ell(x;\theta,\phi) = \|x^m - g_\phi(m, f_\theta(x^v))\|$$

where *m* denotes the mask tokens.

Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise**

Generative Architecture

Encoder learns effective representations for downstream object detection and segmentation on COCO benchmark

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise**

Generative Architecture

Encoder learns effective representations for downstream semantic segmentation on ADE20K benchmark

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

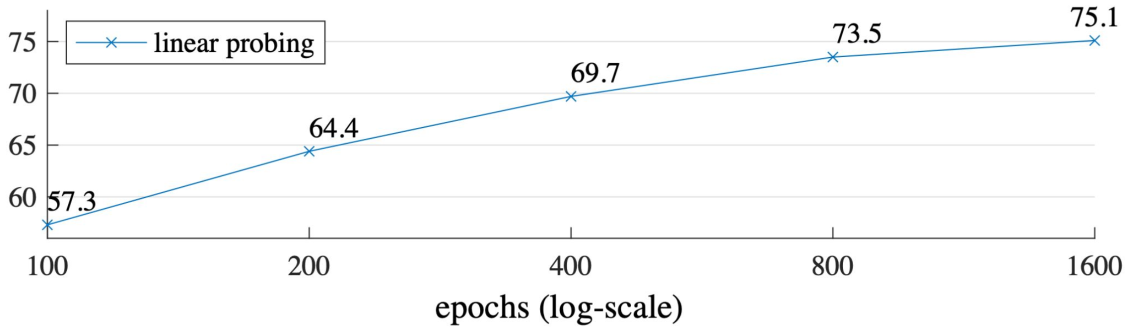
Representation Learning by Denoising Pixels

Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise**

Generative Architecture

Encoder learns effective representations for downstream image classification on the ImageNet-1K benchmark



Representation Learning by Denoising Pixels

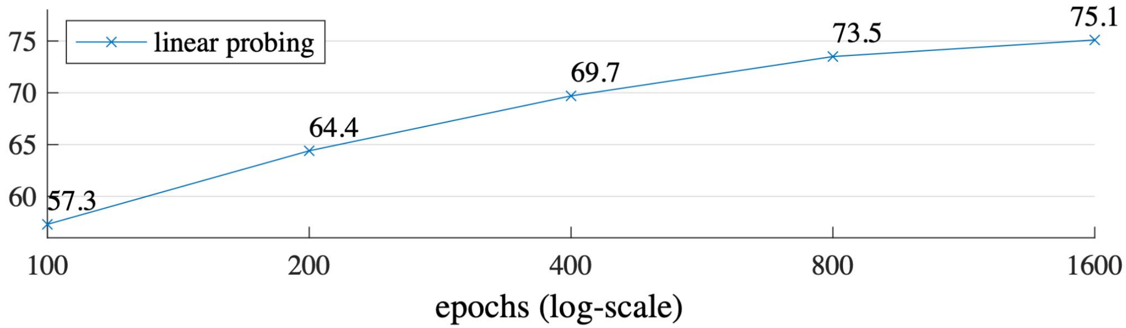
Paper: https://arxiv.org/pdf/2304.03283.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise**

Generative Architecture

Encoder learns effective representations for downstream image classification on the ImageNet-1K benchmark

**** Needs long training schedules and lots of compute...**
Is pixel prediction the most efficient approach?



Meta

Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

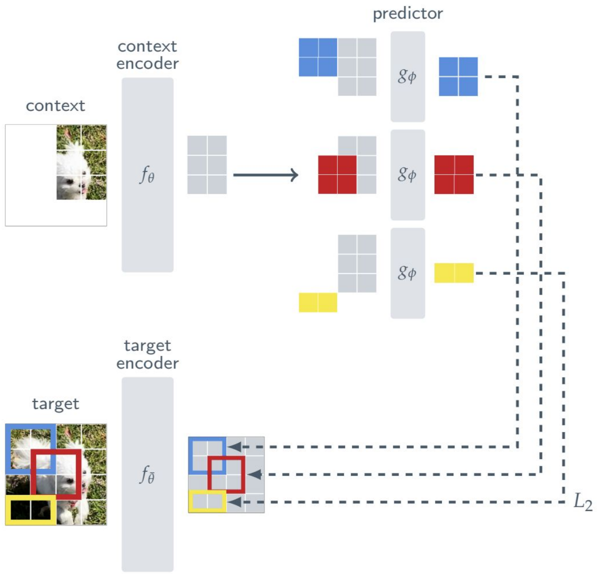
Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

Train the image encoder by predicting ***representations*** of missing patches, instead of raw pixels…

Similar to traditional latent diffusion models, idea is to improve efficiency by solving prediction task in a compressed latent space.

Intuition:
Low-level pixel details are not important for learning effective visual representations, so we abstract away irrelevant information, and solve prediction task in this new space.



Meta

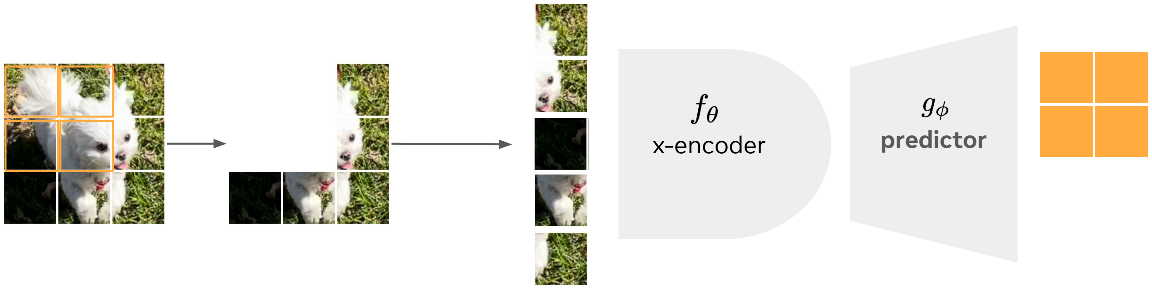
Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

1. Divide image into a sequence of patches
2. Split seq. into "visible regions" and "masked regions"
3. Drop masked patches from the input sequence



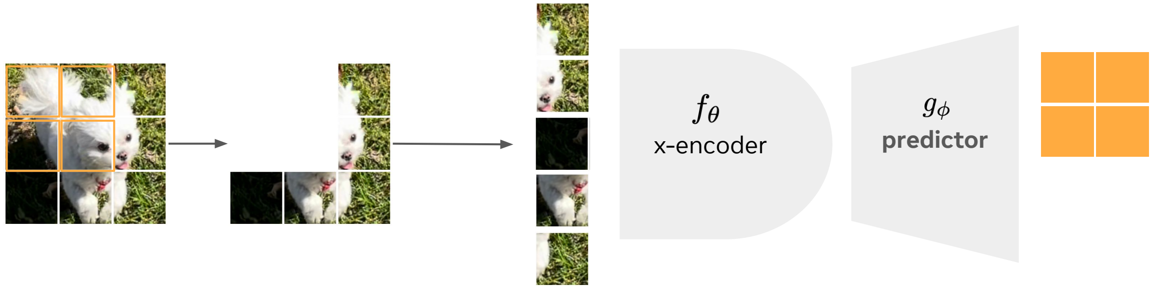
Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

1. Divide image into a sequence of patches
2. Split seq. into "visible regions" and "masked regions"
3. Drop masked patches from the input sequence



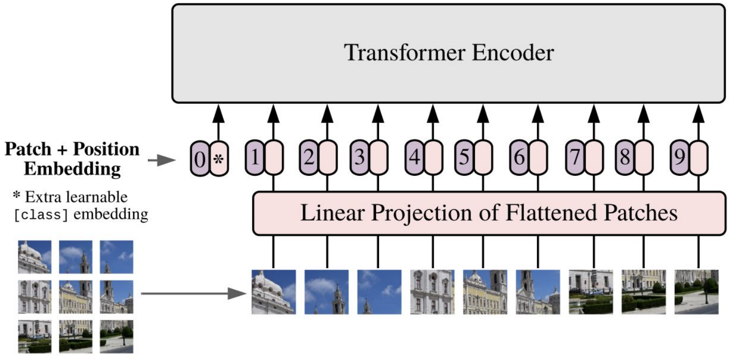
Representation Learning by Denoising in Latent Space

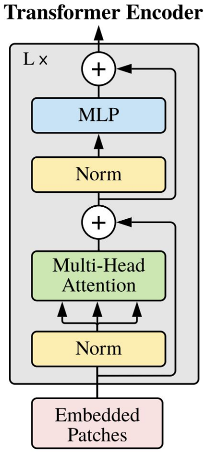
Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

Encoder is still a Vision Transformer (processes sequence of patches)





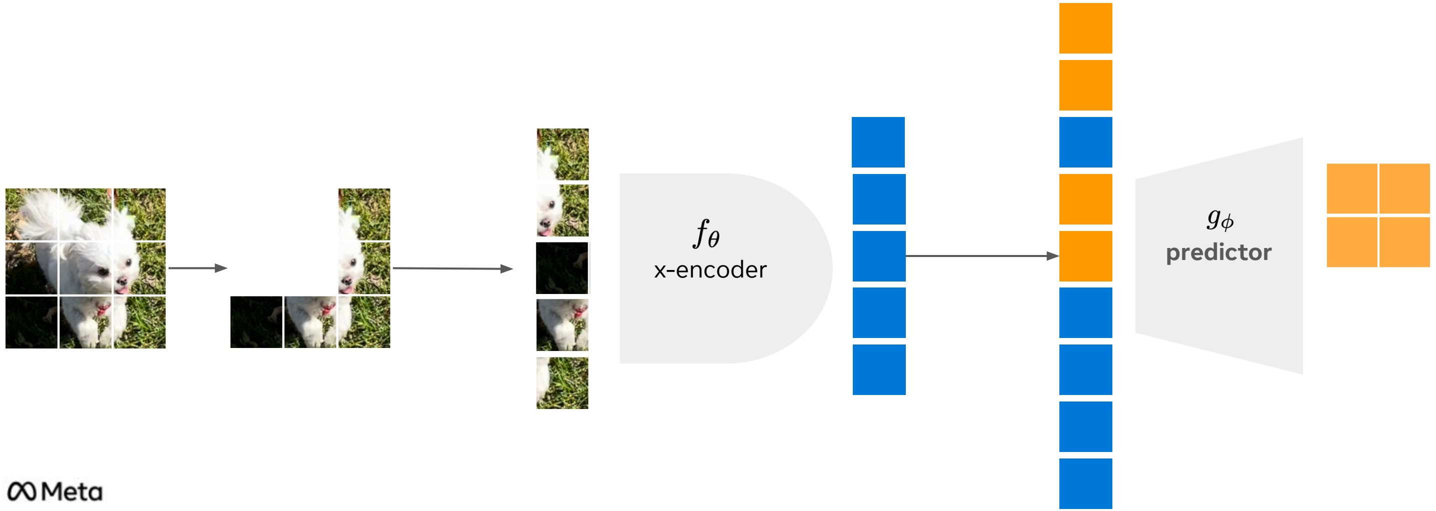
Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

Predictor takes "mask tokens" and patch representations to predict representations of missing regions



Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

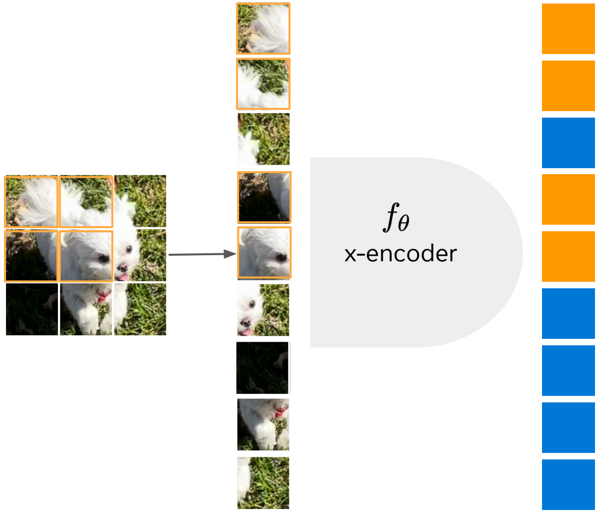
Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

Now we don't predict pixels… instead prediction representations of masked regions…

Note that target representations of masked regions are computed by processing the full image…

important for building *contextualized targets!*



Representation Learning by Denoising in Latent Space

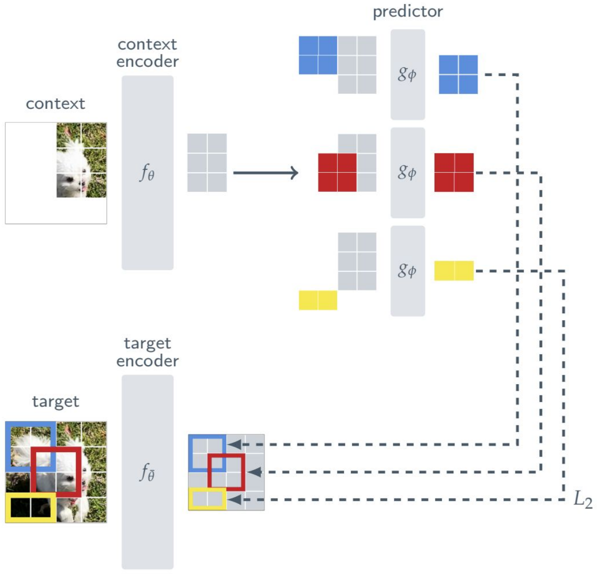
Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

Putting it all together… loss is just a simple L2

$$\ell(x;\theta,\phi) = \|f_\theta(x^m) - g_\phi(m, f_\theta(x^v))\|$$



Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

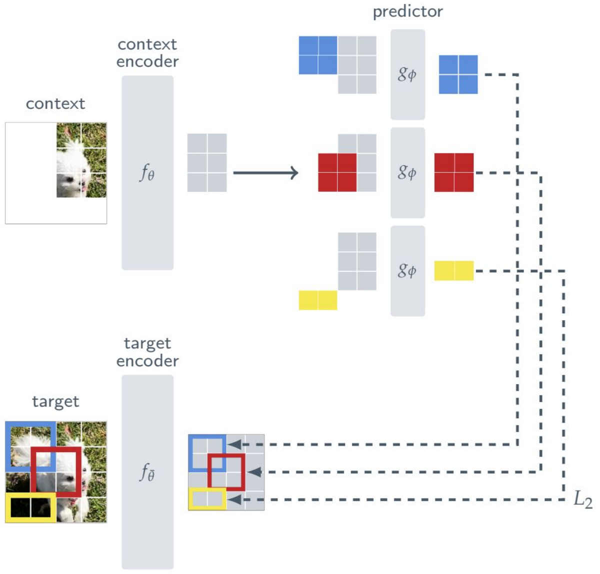
Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

Putting it all together… loss is just a simple L2

$$\ell(x;\theta,\phi) = \|f_\theta(x^m) - g_\phi(m, f_\theta(x^v))\|$$

Can anything go wrong here?



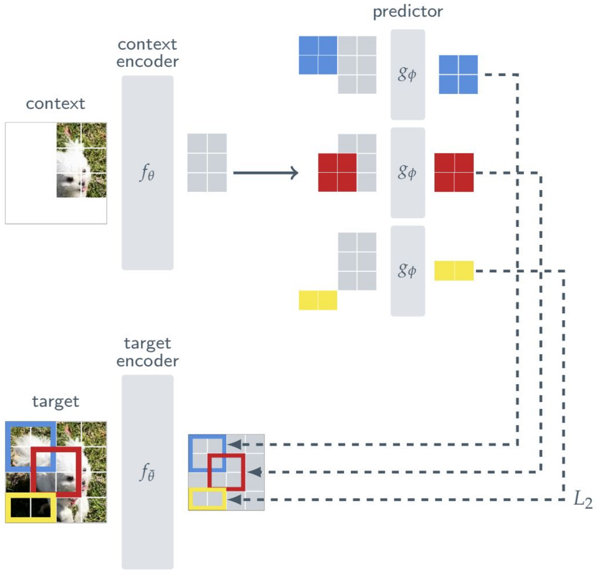
Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

To prevent collapse, add stop gradient operation $\mathrm{sg}(\cdot)$ and compute target encoder weights from an exponential moving average of context encoder weights



$$\ell(x;\theta,\phi) = \|\mathrm{sg}(\overline{f}_\theta(x^m)) - g_\phi(m, f_\theta(x^v))\|$$

Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

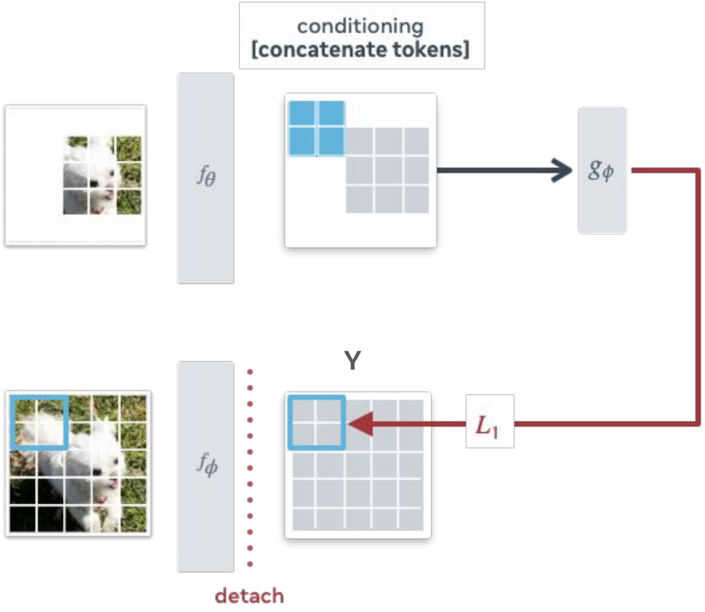
Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

$$g^\star = \mathrm{argmin}_g \mathbb{E}\|g(f_\theta(x^v)) - Y\| = \mathrm{median}(Y | f_\theta(x^v))$$

$$\nabla_\theta \mathbb{E}\|g^\star(f_\theta(x^v)) - Y\| = \nabla_\theta \mathrm{MAD}(Y | f_\theta(x^v))$$

Encoder must capture as much information about image as possible to minimize **median absolute deviation (MAD)**



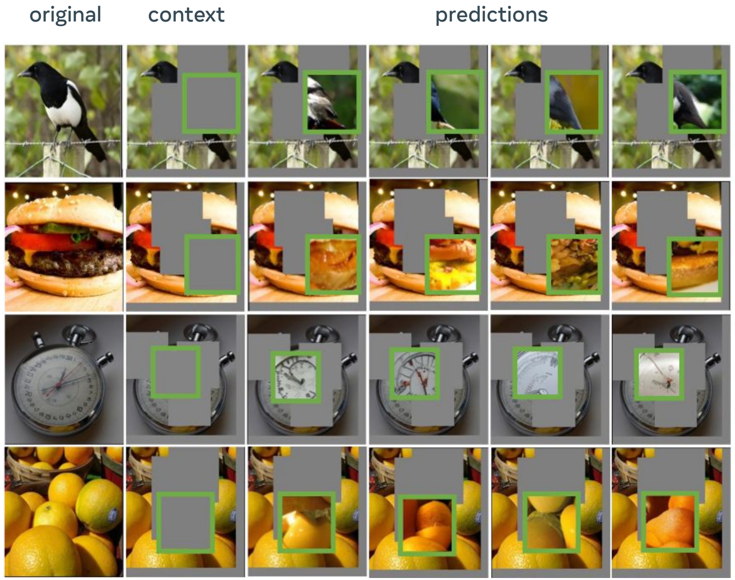
Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Joint-Embedding Predictive Architecture

Freeze pretrained encoder/predictor, and train a model to decode predictions to pixels.



Meta

Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

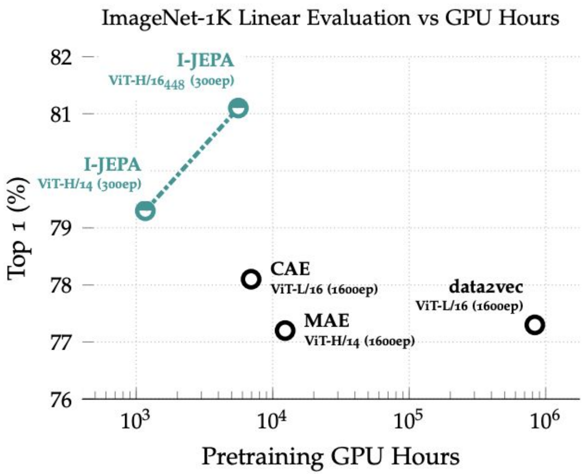
Pretraining:

How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Encoder learns effective representations for downstream image classification on the ImageNet-1K benchmark

… and with much less compute than pixel prediction methods



Meta

Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Pretraining:

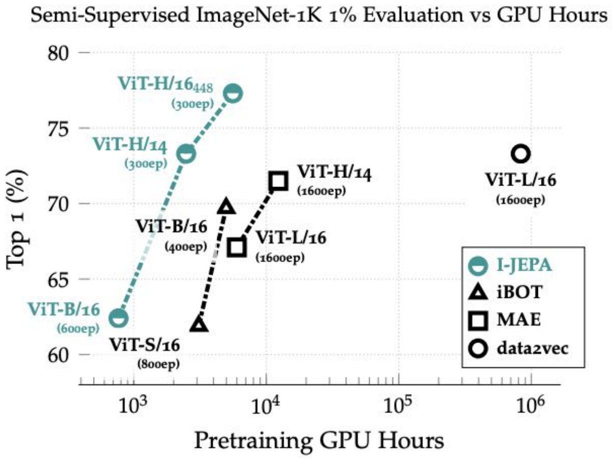
How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Encoder learns effective representations for downstream image classification on the ImageNet-1K benchmark

… and with much less compute than pixel prediction methods

… and with fewer labeled examples



Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Pretraining:

How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Encoder learns effective representations for lower-level vision tasks as well (object counting and depth prediction)

… performs similarly to pixel prediction methods

Method	Arch.	Clevr/Count	Clevr/Dist
Methods without view data augmentations			
data2vec [7]	ViT-L/16	85.3	71.3
MAE [35]	ViT-H/14	**90.5**	**72.4**
I-JEPA	ViT-H/14	86.7	**72.4**

Meta

Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Pretraining:

How can train a neural network to extract semantic features from unstructured data?

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

If we use pixels as targets, instead of representations (output of the target-encoder) the quality of the visual encoder degrades on downstream tasks

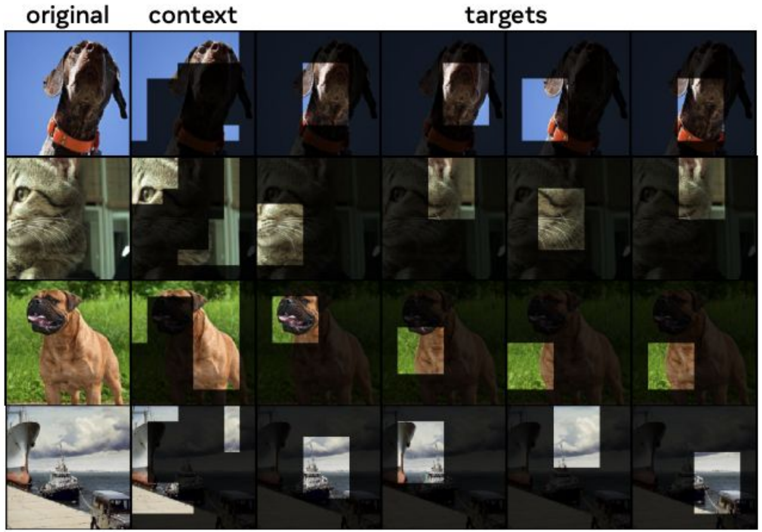
Linear Probing on ImageNet-1k with only 1% of the labels

Targets	Arch.	Epochs	Top-1
Target-Encoder Output	ViT-L/16	500	**66.9**
Pixels	ViT-L/16	800	40.7

Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**



Masking strategy is important for obtaining semantic representations…

multi-block

Representation Learning by Denoising in Latent Space

Paper: https://arxiv.org/pdf/2301.08243.pdf

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Masking strategy is important for obtaining semantic representations…

Linear Probing on ImageNet-1k with only 1% of the labels

	Targets		Context		
Mask	Type	Freq.	Type	Avg. Ratio*	**Top-1**
`multi-block`	Block(0.15, 0.2)	4	Block(0.85, 1.0) $\times$ Complement	0.25	**54.2**
`rasterized`	Quadrant	3	Complement	0.25	15.5
`block`	Block(0.6)	1	Complement	0.4	20.2
`random`	Random(0.6)	1	Complement	0.4	17.6

**Avg. Ratio is the average number of patches in the context block relative to the total number of patches in the image.*

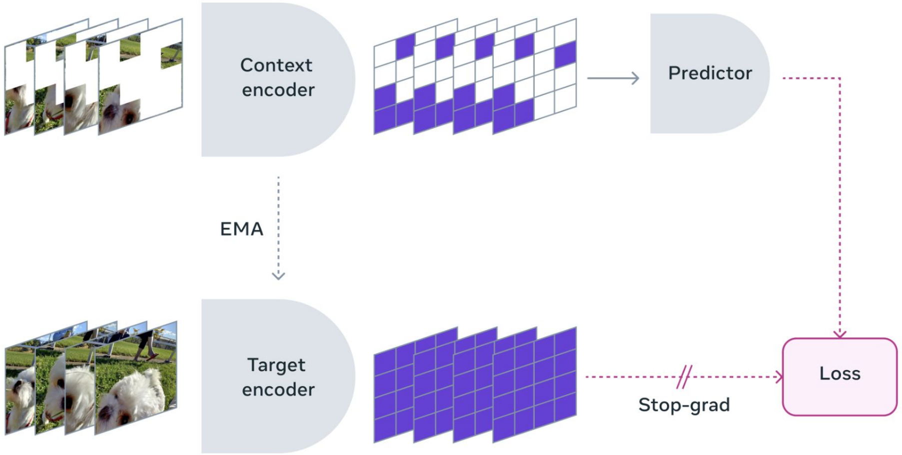
Meta

Representation Learning by Denoising in Latent Space

Paper:
https://ai.meta.com/research/publications/revisiting-feature-prediction-for-learning-visual-representations-from-video/

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Generality of the prediction task means that we can extend the learning principle to video!

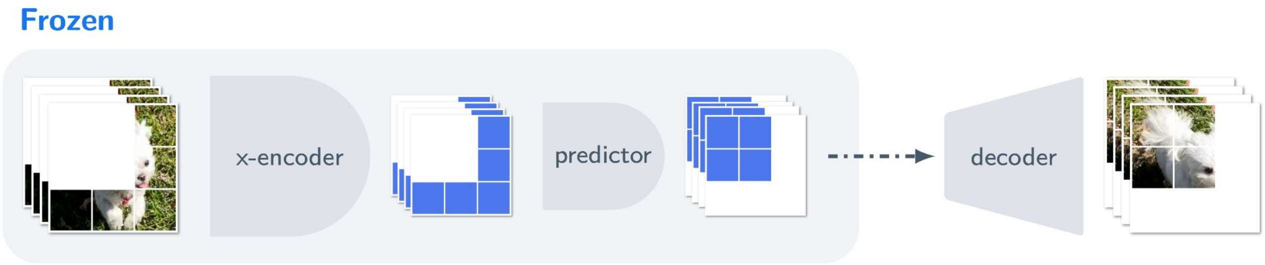


Representation Learning by Denoising in Latent Space

Paper:
https://ai.meta.com/research/publications/revisiting-feature-prediction-for-learning-visual-representations-from-video/

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

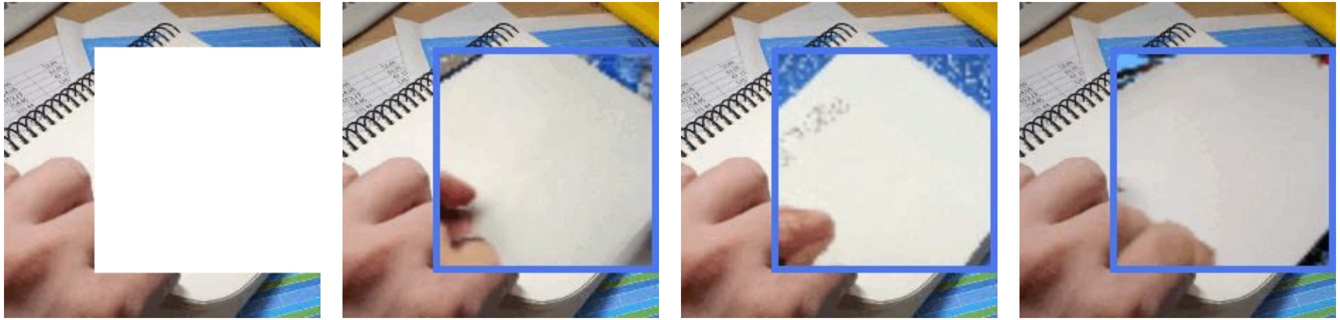
Freeze pretrained encoder/predictor, and train a model to decode predictions to pixels.



Representation Learning by Denoising in Latent Space

Paper:
https://ai.meta.com/research/publications/revisiting-feature-prediction-for-learning-visual-representations-from-video/

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

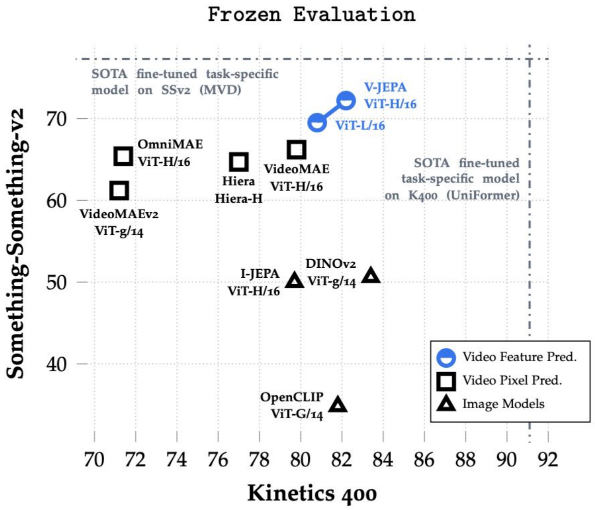


Representation Learning by Denoising in Latent Space

Paper:
https://ai.meta.com/research/publications/revisiting-feature-prediction-for-learning-visual-representations-from-video/

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Encoder learns effective representations for downstream video classification tasks



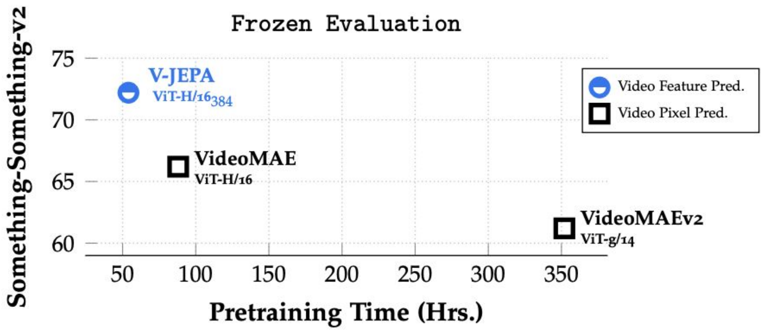
Representation Learning by Denoising in Latent Space

Paper:
https://ai.meta.com/research/publications/revisiting-feature-prediction-for-learning-visual-representations-from-video/

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Encoder learns effective representations for downstream video classification tasks

… and with much less compute than pixel prediction methods



Representation Learning by Denoising in Latent Space

Paper:
https://ai.meta.com/research/publications/revisiting-feature-prediction-for-learning-visual-representations-from-video/

Unsupervised Learning: **Denoising Pixels with Mask Noise in Latent Space**

Encoder learns effective representations for downstream video classification tasks

… and with much fewer labeled examples

		Frozen Evaluation					
		K400 (16×8×3)			**SSv2** (16×2×3)		
Method	**Arch.**	5% (∼29 samples per class)	10% (∼58 samples per class)	50% (∼287 samples per class)	5% (∼48 samples per class)	10% (∼96 samples per class)	50% (∼440 samples per class)
MVD	ViT-L/16	62.6 ± 0.2	68.3 ± 0.2	77.2 ± 0.3	42.9 ± 0.8	49.5 ± 0.6	61.0 ± 0.2
VideoMAE	ViT-H/16	62.3 ± 0.3	68.5 ± 0.2	78.2 ± 0.1	41.4 ± 0.8	48.1 ± 0.2	60.5 ± 0.4
VideoMAEv2	ViT-g/14	37.0 ± 0.3	48.8 ± 0.4	67.8 ± 0.1	28.0 ± 1.0	37.3 ± 0.3	54.0 ± 0.3
V-JEPA	ViT-H/16	67.0 ± 0.2	72.1 ± 0.1	80.2 ± 0.2	51.9 ± 0.3	57.5 ± 0.4	67.3 ± 0.2
	ViT-H/16_{384}	**68.2 ± 0.2**	**72.8 ± 0.2**	**80.6 ± 0.2**	**54.0 ± 0.2**	**59.3 ± 0.5**	**67.9 ± 0.2**