Modality-Agnostic Self-Supervised Learning with Meta-Learned Masked Auto-Encoder

Huiwon Jang[A]*, Jihoon Tack[A]*, Daewon Choi[B], Jongheon Jeong[A], Jinwoo Shin[A]

[A]KAIST, [B]Korea University

*Equal contribution

TL; DR. Interpreting MAE through **meta-learning** and applying advanced meta-learning techniques to improve unsupervised representation of MAE on **arbitrary modalities**.

Introduction

Modality-agnostic SSL learns representation without modality-specific inductive bias, allowing pretraining for new domains. They often construct patch-level pretext tasks (ShED) or utilize mask (Capri, MAE) [1-2].

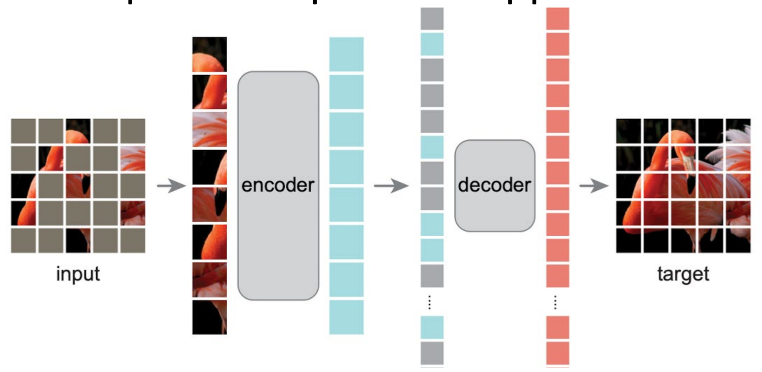
Masked Auto-Encoder (MAE) is a powerful SSL for various domains without needing domain-specific bias: mask prediction task.

- Image (MAE [3]), Language (BERT [4]), Tabular (Met [5]), …

Research Question 1: Is MAE indeed a modality-agnostic?

Observation: MAE with a proper decoder size outperforms previous approaches

decoder size	EuroSAT	Pfam	LibriSpeech
prev. best	**87.4**	54.7	60.2
0	86.3	44.7	33.3
2	86.7	**61.4**	68.1
4	**87.4**	61.3	64.1
6	86.7	**61.4**	**74.1**



Research Question 2: How to improve MAE in a modality-agnostic manner?

Key idea: Interpreting MAE as an amortization-based meta-learner and leveraging the advances of meta-learning.

- Gradient-based meta-learning on latent to improve the task adaptation process
- Task contrastive learning to better encode the task knowledge

Summary of Contribution

We propose **MetaMAE**, an effective modality-agnostic self-supervised learning framework. We interpret mask reconstruction task of MAE as a meta-learning to suggest an integration with advanced modality-agnostic meta-learning methods. Extensive experiments demonstrate that

1. **MetaMAE** significantly improves the performance of modality-agnostic SSL across a diverse range of modalities
2. **MetaMAE** can extend toward multi-modal scenarios

References

[1] Tamkin et al., DABS: A Domain-agnostic Benchmark for Self-supervised Learning, NeurIPS Datasets and Benchmarks 2021
[2] Tamkin et al., DABS 2.0: Improved Datasets and Algorithms for Universal Self Supervision, NeurIPS Datasets and Benchmarks 2022
[3] He et al., Masked Autoencoders are Scalable Vision Learners, CVPR 2022
[4] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019
[5] Majmundar et al., Met: Masked encoding for tabular data, Arxiv 2022

Interpreting MAE through meta-learning

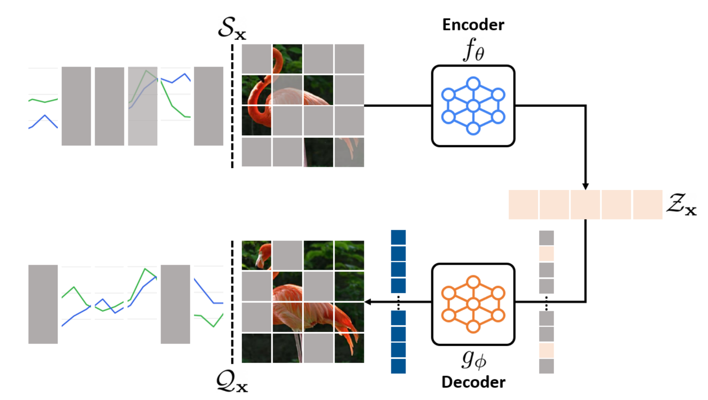
Notation for meta-learning. $\mathcal{S} \cup \mathcal{Q} \sim \mathcal{T}$ where $\mathcal{S}$ is a Support set (or train data) and $\mathcal{Q}$ is a Query set (or test data) for a sampled task $\mathcal{T}$.

Amortization-based meta-learning utilizes model (or memory) for meta-leaner:

- $\mathcal{S} \cup \mathcal{Q} = \{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^{N} \sim \mathcal{T}$: Sampling task (# task = 1)
- $\mathcal{Z} = f_\theta(\mathcal{S})$: Memory
- $\mathrm{y}^{(q)} = g_\phi(\mathbf{x}^{(q)}; \mathcal{Z})$

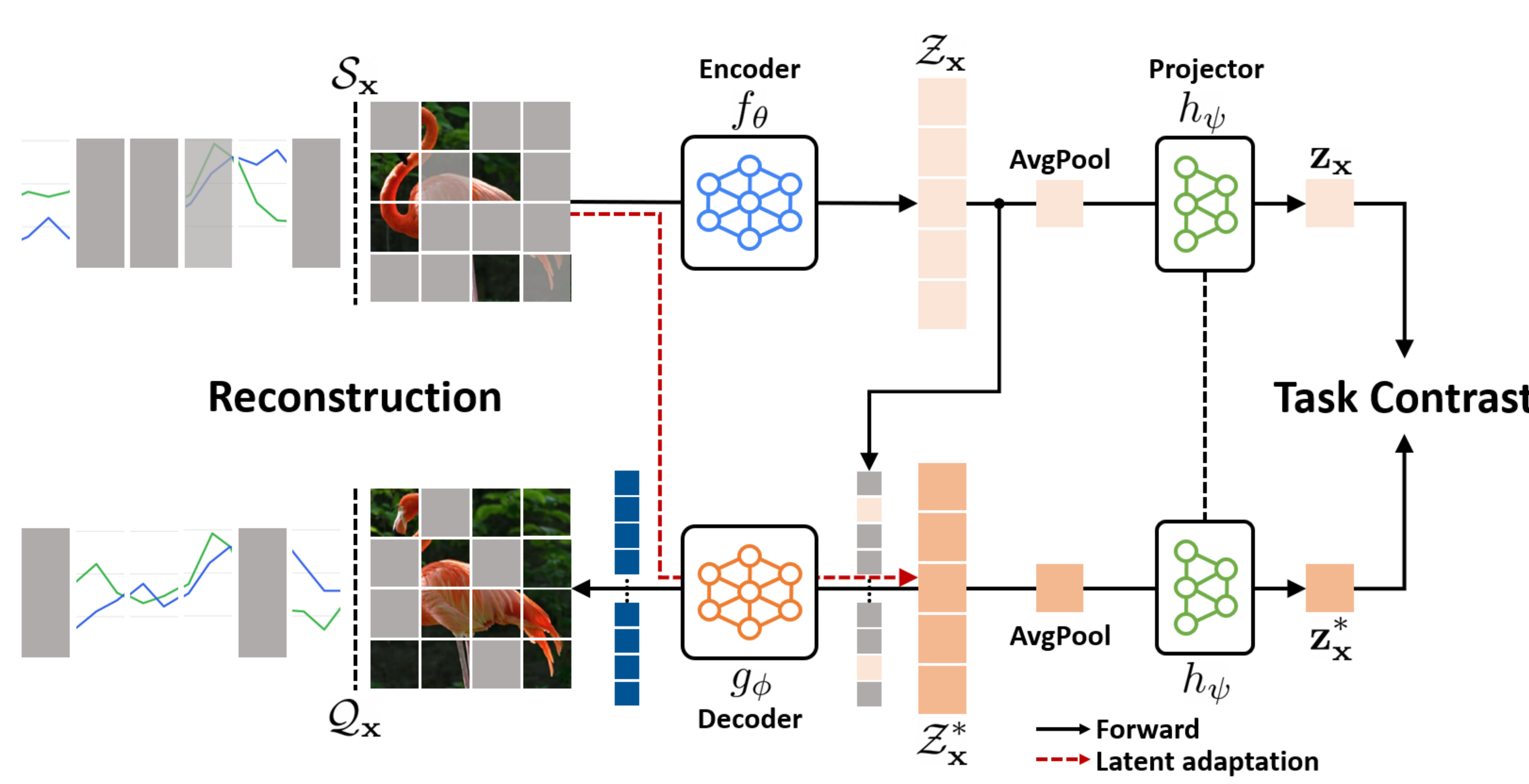
Task formulation of **MAE** with batch size 1:

- $\mathrm{Tokenize}(\mathbf{x}) \coloneqq \{(m, \bar{\mathbf{x}}^{(m)})\}_{m=1}^{M} = \mathcal{S}_\mathbf{x} \cup \mathcal{Q}_\mathbf{x}$
- $\mathcal{Z}_\mathbf{x} = f_\theta(\mathcal{S}_\mathbf{x})$
- $\bar{\mathbf{x}}^{(q)} = g_\phi^{(q)}(\mathcal{Z}_\mathbf{x}) \coloneqq g_\phi(q; \mathcal{Z}_\mathbf{x})$



Method: MetaMAE

Integration of two advanced meta-learning techniques to enhance MAE:



1. Latent adaptation via gradient-based meta-learning.

Reconstructing $\mathcal{Q}_\mathbf{x}$ from task-specific latent: $\mathcal{Z}^*_\mathbf{x} = \mathcal{Z}_\mathbf{x} - \alpha \nabla_{\mathcal{Z}_\mathbf{x}} \mathcal{L}_{MAE}(\theta, \phi; \tilde{\mathcal{S}}_\mathbf{x})$ where $\tilde{\mathcal{S}}_\mathbf{x} = \mathcal{S}_\mathbf{x} \cup \mathcal{N}(\mathcal{S}_\mathbf{x}; r)$ and $\mathcal{N}(\mathcal{S}_\mathbf{x}; r)$ bridges the gap between the latents.

- $\mathcal{L}_{grad}(\mathbf{x}, \theta, \phi) = \sum_{(q, \bar{\mathbf{x}}^{(q)}) \in \mathcal{Q}_\mathbf{x}} d(\bar{\mathbf{x}}^{(q)}, g_\phi^{(q)}(\mathcal{Z}^*_\mathbf{x}))$

2. Task contrastive learning.

Contrastive learning on prototype representation of tasks.

- $\mathcal{L}_{task-con}(\mathbf{x}, \theta, \phi) = \frac{1}{2}[l_{con}(\mathbf{z}_\mathbf{x}; \mathbf{z}^*_\mathbf{x}, \mathcal{T} \backslash \{\mathbf{z}^*_\mathbf{x}\}) + l_{con}(\mathbf{z}^*_\mathbf{x}; \mathbf{z}_\mathbf{x}, \mathcal{T} \backslash \{\mathbf{z}_\mathbf{x}\})]$

where $\mathcal{T} = \cup_\mathbf{x}\{\mathbf{z}_\mathbf{x}, \mathbf{z}^*_\mathbf{x}\}$ is a collection of all representations of tasks

Learning objective: $\mathcal{L}_{grad}(\mathbf{x}, \theta, \phi) + \lambda \mathcal{L}_{task-con}(\mathbf{x}, \theta, \phi)$

Experiment

MetaMAE consistently and significantly outperforms prior modality-agnostic SSL in (a) in-domain and (b) cross-domain linear evaluation.

(a)

Modality	Time-series	Tabular	MS Image	Token		Speech	RGB Image
Dataset	PAMAP2	HIGGS	EuroSAT	Genom	Pfam	Libri	WaferMap
Random initialization							
Baseline	69.8†	54.8†	62.3†	37.2†	30.1	17.1*	77.7†
Self-supervised learning Framework							
e-Mix	80.1	65.7	87.4	40.5	31.3	60.2	92.6
ShED	85.2	68.0†	61.5†	33.6	54.7	34.8*	92.4†
Capri	-	-	67.4†	23.5†	27.4	25.4	92.5†
MAE	85.3†	70.0†	86.3†	53.6	44.7	46.0	93.9†
MetaMAE	**89.3**	**71.5**	**88.5**	**69.4**	**62.3**	**79.8**	**95.5**

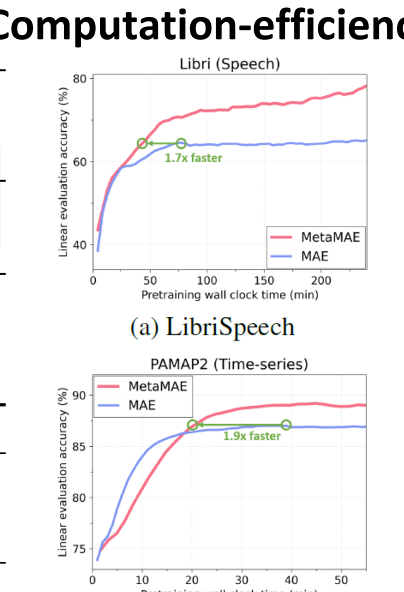
(b)

Pretrain data	Transfer data	Baseline	SSL Framework				
			e-Mix	ShED	Capri	MAE	MetaMAE
Genomics	Genomics-OOD	8.6	9.7	7.3	5.5	22.2	**37.2**
Pfam	SCOP	8.0	5.7	10.7	2.0	7.9	**11.8**
	Secondary	52.4	53.7	**67.6**	49.5	62.5	65.9
	Stability	0.31	0.39	**0.53**	0.26	0.40	**0.53**
	Fluorescence	0.04	0.20	0.27	0.06	0.06	**0.31**
LibriSpeech	Audio MNIST	33.1*	80.4*	67.3*	53.6	45.1	**89.5**
	Fluent Loc	62.1*	60.9*	60.2*	59.8	61.7	**66.7**
	Fluent Act	26.2*	29.9*	30.5*	28.3	26.8	**38.4**
	Fluent Obj	30.1*	39.9*	39.4*	33.1	32.0	**49.3**
	Google Speech	4.9*	19.2*	20.7*	13.7	9.5	**46.8**
	VoxCeleb1	0.6*	2.4*	2.8*	1.6	1.6	**7.4**
ImageNet32	CIFAR-10	24.2*	39.4*	39.6*	48.7	46.0	**59.2**
	CUB	1.6*	3.9*	3.0*	3.7	3.1	**6.3**
	VGG Flowers	9.0*	26.0*	13.0*	18.6	22.2	**36.3**
	DTD	7.4*	8.8*	18.4*	14.7	14.2	**20.9**
	Traffic Sign	14.3*	65.1*	27.5*	28.0	32.0	**67.1**
	Aircraft	2.7*	10.2*	5.6*	6.4	5.9	**16.4**

MetaMAE can extend toward multi-modal scenarios

Pretrain data	Transfer data	Baseline	SSL Framework				
			e-Mix	ShED	Capri	MAE	MetaMAE
MSCOCO	VQA	53.4	57.6	53.1	52.9	54.2	**69.7**
	Mismatched-caption	49.8	50.1	50.6	49.6	49.3	**70.5**

Computation-efficiency



(a) LibriSpeech

(b) PAMAP2

Component ablation shows the importance of each components.

Decoder	Gradient-based	Task contrast	PAMAP2	Genomics	EuroSAT	LibriSpeech	HIGGS	Pfam
✗	✗	✗	85.3	53.6	86.3	33.3	70.0	44.7
✓	✗	✗	86.5	65.2	87.4	64.1	70.5	61.3
✓	✓	✗	88.3	**69.4**	87.4	64.5	71.1	61.3
✓	✓	✓	**89.3**	**69.4**	**88.5**	**79.8**	**71.5**	**62.3**

MetaMAE shows robust performance regardless of hyperparameter selection

Modality	Time-series	Tabular	MS Image	Token		Speech	RGB Image
Dataset	PAMAP2	HIGGS	EuroSAT	Genom	Pfam	Libri	WaferMap
MetaMAE (sharing 3 HPs)	89.1	71.0	**88.5**	55.4	62.2	77.1	95.4
MetaMAE (sharing 2 HPs)	89.1	71.1	**88.5**	66.7	62.2	77.1	95.4
MetaMAE (reported)	**89.3**	**71.5**	**88.5**	**69.4**	**62.3**	**79.8**	**95.5**