OMRiDA: Omni Mathematical expression Recognition via Domain Adaptation

Sungmin Yang, Soyeon Kim, Dongjin Kim*
School of Data Science, Hanyang University

Research Motivations

1. Mathematical Expression Recognition(MER) is challenging due to **complex 2d structure and the number of symbols**

2. Recent MER studies have mainly focused on increasing performance on **CROHME** benchmark only

3. Model performing decent on CROHME showed **poor performance on other datasets** with different distributions

4. Considering varying handwriting style in the real-world, it is impossible to construct **a dataset covering all kind of handwriting styles**

Research question: How can we build a general MER model for various distributions, from the limited data distribution?

We proposed domain-invariant MER model by leveraging
1) rendered additional source domain (PME) and 2) domain-free text modal representation

Dataset

- Source dataset: CROHME, CROPME(rendered)
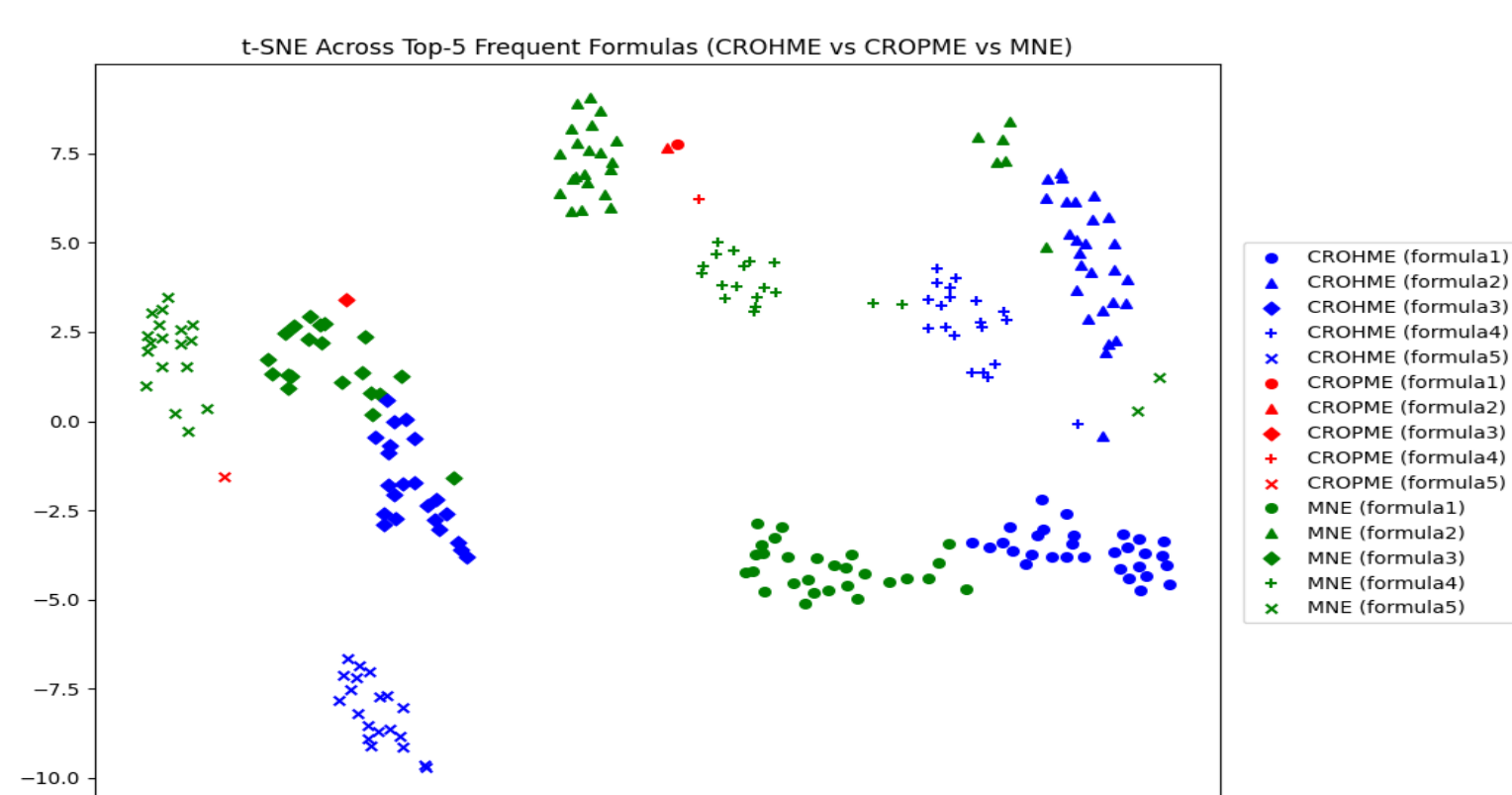- Target dataset: IM2LATEX, MNE, HME100K, MATHWRITING

▶ Different distribution of data depending on the handwriting style, device, background, and acquisition process

Methods

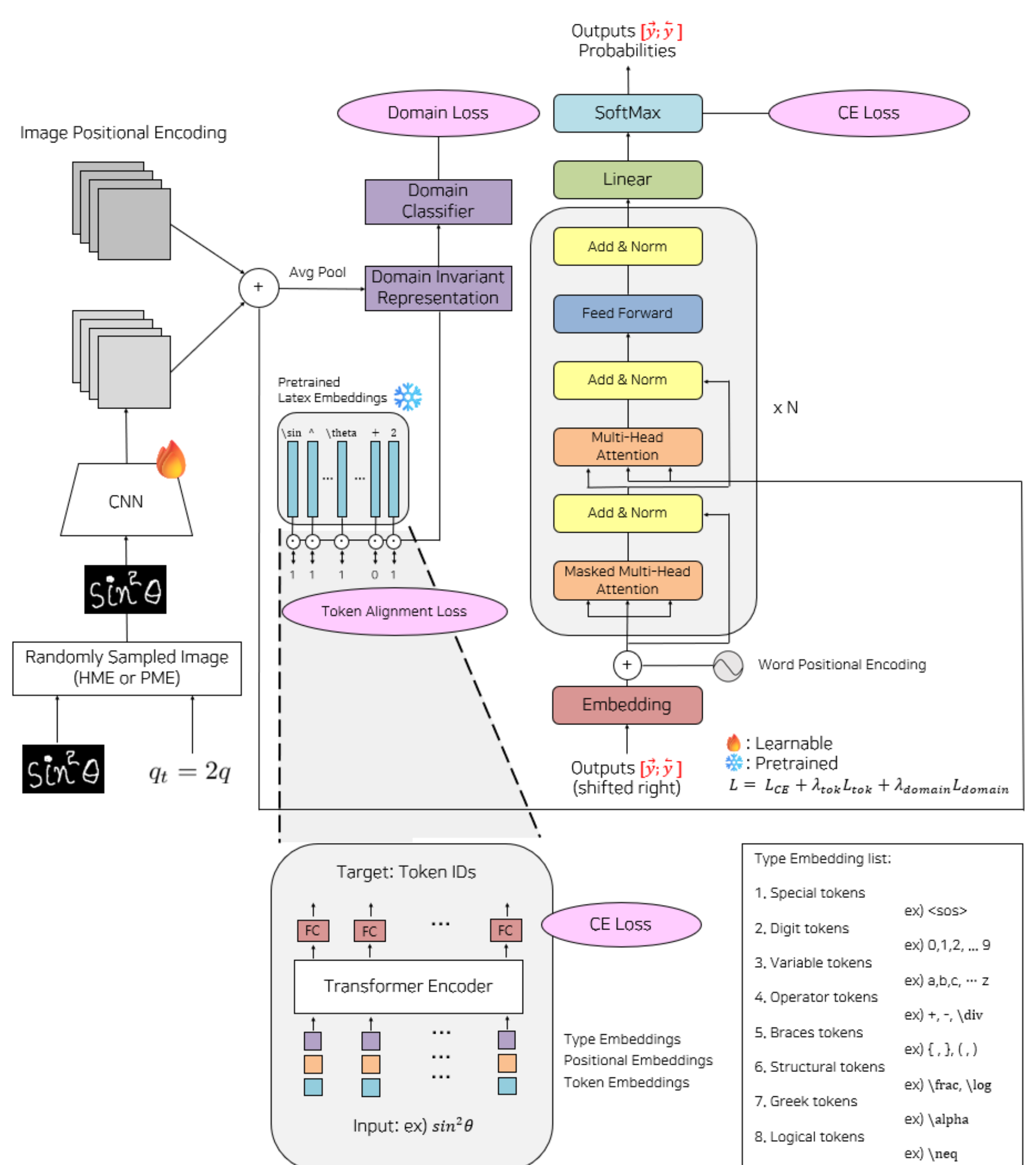## Experiments & Results

Evaluated on Expression Recognition Rate

Datasets	BTTR (HME Only)	BTTR (HME+PME)	Ours
CROHME_2014	19.98	15.82	34.28
CROHME_2016	20.84	17.18	34.35
CROHME_2019	15.76	14.93	34.95
CROPME_2014	0.00	34.38	61.76
CROPME_2016	0.00	30.34	55.97
CROPME_2019	0.00	34.11	58.72
IM2LATEX_HME	0.00	0.00	0.00
IM2LATEX_PME	14.83	12.75	1.30
MNE_N1	3.62	1.32	31.68
MNE_N2	3.76	2.94	11.84
MNE_N3	0.00	0.29	16.26
HME100K	0.00	0.00	0.00
MATHWRITING	0.00	0.00	0.00

- Our model showed performance improvements for both source and target domains
- However, for unseen domains, its predictions were still far from correct full expressions, frequently producing entirely incorrect formulas

▶ Same-label samples show similar embeddings across domains

Ablation Study

Evaluated on Word Error Rate

Datasets	BTTR +DANN	BTTR +Token Alignment	BTTR+DANN +Token Alignment
IM2LATEX_HME	1.04	1.02	0.99
IM2LATEX_PME	0.55	0.45	0.44
MNE_N1	0.33	0.16	0.19
MNE_N2	0.45	0.24	0.31
MNE_N3	0.52	0.39	0.39
HME100K	2.38	1.02	1.02
MATHWRITING	1.75	2.01	1.52

Future Work

1. Apply same approach for additional source domains
2. Apply same approach on SOTA baseline model
3. Increase LATEX token vocabulary size