AdapTex: Image-to-LaTeX Equation OCR Model with Hybrid Vision Transformer and Transfer Learning with Adapter

Dong Ik Lee*†*, Seong Ju Lee*†*, Jae Hyung Sim, Yeon Jun Jung

†*contributed equally*

AIFFEL, ModuLabs Inc.

모두의연구소

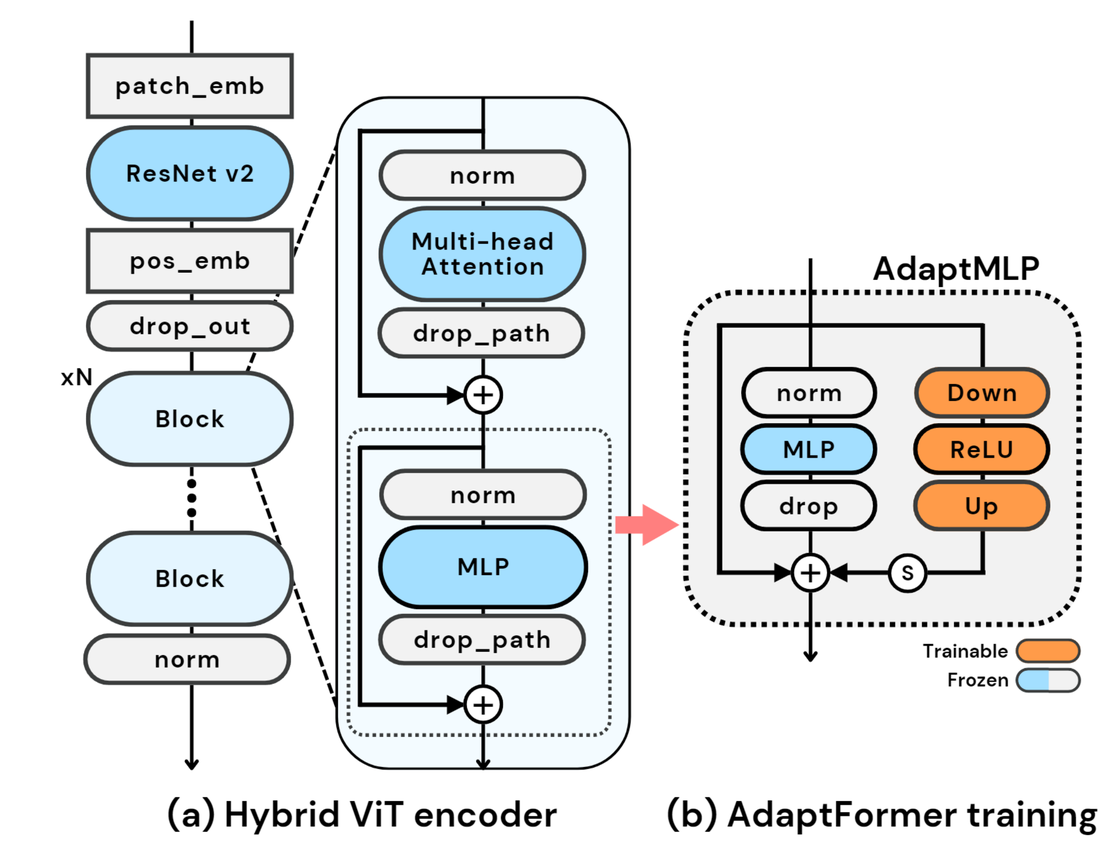
Information

Team project at AI Engineer education program(AIFFEL: Research course)
Submission approved: KIISE, 2023.11.17.
Subjects: Computer Vision and Pattern Recognition, Optical Character Recognition

Abstract

Optical character recognition (OCR) is the task of converting text in an image to a machine-readable format. On the other hand, the process of converting mathematical equations to LaTeX format for representation in electronic documents in the academic and educational fields is a cumbersome task. **This study aims to automate this process through a deep learning model to improve the work efficiency of researchers, educators, and students. To this end, we proposed a transformer-based model that takes in a mathematical equation image as input and outputs a sequence in LaTeX format. And we constructed additional 87% dataset, and performace improved average 35% compare to pre-trained model.** We also verified efficient learning and performance improvement through the addition of MLP adapters.

Proposed Network



Encoder We used a hybrid structure encoder consisting of ResNetV2 backbone with 2, 3, and 7 layer blocks and ViT(a).

Decoder The decoder uses the same decoder as the existing transformer decoder, and outputs a sequence of LaTeX tokens. The tokens are generated using the LaTeX ground truth tokens from the constructed dataset.

Adapter The adapter used in this study, AdaptFormer, is a method that keeps the pre-trained weights frozen and only trains the newly added adapter part. It shows performance better than full fine-tuning with very few parameters(b).

Datasets

Dataset	name	img type	size
PDF	P1	printed	234,884(235k)
AIHUB_pdf	P2	printed	24,559(25k)
CROHME	H1	handwritten	10,846(11k)
AIHUB_handwritten	H2	handwritten	88,605(88k)
AIDA	H3	handwritten	100,000(100k)
CROHME_symbol	S	handwritten	375,974(376k)

We collected a dataset of mathematical equation images labeled in LaTeX format. We then constructed a unified dataset by balancing the data between printed and handwritten formats. In the integration process, we refined the LaTeX tokens to unify the ground truth LaTeX expression for a specific equation.

Method

The proposed model was fine-tuned using the pre-trained model on the PDF dataset, which is in printed format. Since the formats of printed images and handwritten images are similar, we verified through PoC experiments that the full fine-tuning method, which trains all of the model's weights, is effective.
For improved generalization performance, we checked the continued improvement of the recognition performance for both printed and handwritten images by gradually adding the dataset.

Results and Conclusions

	Printed type			Handwritten type		
Dataset	BLEU	Token acc	Edit dist↓	BLEU	Token acc	Edit dist↓
P1*	0.878	0.586	0.092	-	-	-
(+)H1	0.889	0.596	0.086	0.498	0.457	2.910
(+)H2	0.877	0.579	0.097	0.786	0.674	0.228
(+)P2	**0.917**	**0.736**	0.056	0.784	0.685	0.191
(+)H3**	0.916	0.732	**0.053**	**0.912**	**0.867**	**0.077**
(+)S	0.895	0.685	0.069	0.752	0.612	0.241

* Pre-trained model trained by P1 dataset, ** Encoder depth 6

Datasets finetuning The performance improved as the dataset was gradually added, and eventually the performance for both printed and handwritten data showed similar scores. On the other hand, we confirmed that the performance decreased unexpectedly when the symbol dataset was used.

	Printed type			Handwritten type		
Model	BLEU	Token acc	Edit dist↓	BLEU	Token acc	Edit dist↓
1	0.917	0.736	0.056	0.784	0.685	0.191
2	0.914	0.729	0.057	**0.806**	**0.704**	0.195
3	0.887	0.666	0.075	0.674	0.523	0.387
4	**0.916**	**0.740**	**0.056**	0.805	0.700	**0.180**
5	0.906	0.701	0.062	0.672	0.547	0.342

Parameter searching Learning rate is a key hyperparameter that can have a significant impact on the performance of a model. To find the optimal learning rate, we conducted ablation studies. The basic settings used StepLR scheduler with an initial learning rate of 1.0E-3 and a gradient clipping threshold of 1.0 (Model 1). **Using CAWR scheduler instead of StepLR showed better performance, with a minimum-maximum range of 1.0E-7 to 1.0E-3 (Model 2), and a gradient clipping threshold of 0.5 (Model 4)**. (In the case of Model 3, the CAWR scheduler was set from 1.0E-6 to 1.0E-4.), However, adjusting the encoder-decoder learning rate ratio (10:1) (Model 5) did not improve the performance as expected.

	Printed type			Handwritten type		
Model	BLEU	Token acc	Edit dist↓	BLEU	Token acc	Edit dist↓
1	0.917	0.736	0.056	0.784	0.685	0.191
2	**0.921**	**0.745**	0.055	0.901	0.840	0.085
3	0.917	0.739	**0.054**	**0.933**	**0.887**	**0.059**
4	0.895	0.685	0.069	0.752	0.612	0.241
5	0.914	0.735	0.057	0.929	0.880	0.067

AdaptFormer training(AdapTex) We added AdaptFormer to the **base model (Model 1)** that was fine-tuned in the previous experiment, and found that the **underfitting was improved (Model 2)**. **And changing the depth of the encoder block from 4 to 6, the recognition performance of handwritten data improved(Model 3)**. We also found that the Symbol dataset model(Model 4), which had been performing poorly on the training set(Model 5). This suggests that the transfer-learning by adapters can improve performance, after freezing the previous weights.

Conclusions In this study, (1)we improved the mathematical equation recognition performance of a pre-trained model. (2)We also developed a model that can recognize handwritten mathematical equations, which was a limitation of the existing model. (3)We proposed a transformer-based network that uses a ViT and AdaptFormer, and explored best learning rate. (4)We verified that performance improvement and efficient learning by adding adapters.

Demo results

이미지 input	모델 output(LaTeX 수식 렌더링 시)
$\Psi \approx \psi_1(q)\|(x' + \frac{1}{2}p_1'^2 t, k')\rangle + \psi_2(q)\|(x' + \frac{1}{2}p_2'^2 t, k')\rangle.$	$\Psi \approx \psi_1(q)\|(x' + \frac{1}{2}p_1'^2 t, k')\rangle + \psi_2(q)\|(x' + \frac{1}{2}p_2'^2 t, k')\rangle.$
$x^n = \sum_{k=1}^{n} \left\{ n \atop k \right\} x^{\underline{k}} = \sum_{k=1}^{n} \left\{ n \atop k \right\} (-1)^{n-k} x^{\overline{k}},$	$x^n = \sum_{k=1}^{n} \left\{ n \atop k \right\} x^{\underline{k}} = \sum_{k=1}^{n} \left\{ n \atop k \right\} (-1)^{n-k} x^{\overline{k}},$
$\frac{d}{dt} e^{\alpha X(t)} = \int_0^1 e^{(1-\alpha)X(t)} \frac{dX(t)}{dt} e^{\alpha X(t)} d\alpha$	$\frac{d}{dt} e^{\alpha X(t)} = \int_0^1 e^{(1-\alpha)X(t)} \frac{dX(t)}{dt} e^{\alpha X(t)} d\alpha$
$\int \frac{\sin(x)+1}{\sqrt{\cos^3(x)+\tan(x)}} dx$	$\int \frac{\sin(x)+1}{\sqrt{\cos^3(x)+\tan(x)}} dx$