Cluster Character Loss with Transformer in Scene Text Recognition

Nguyễn Khắc Thái [1,2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam

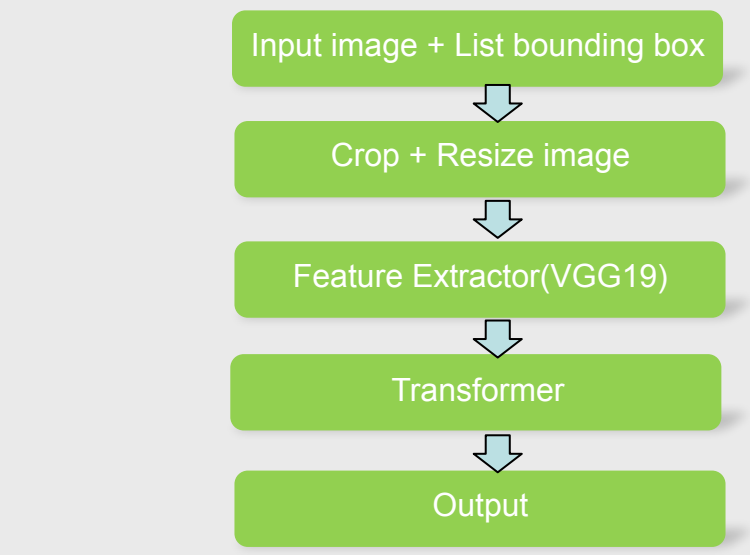
What ?

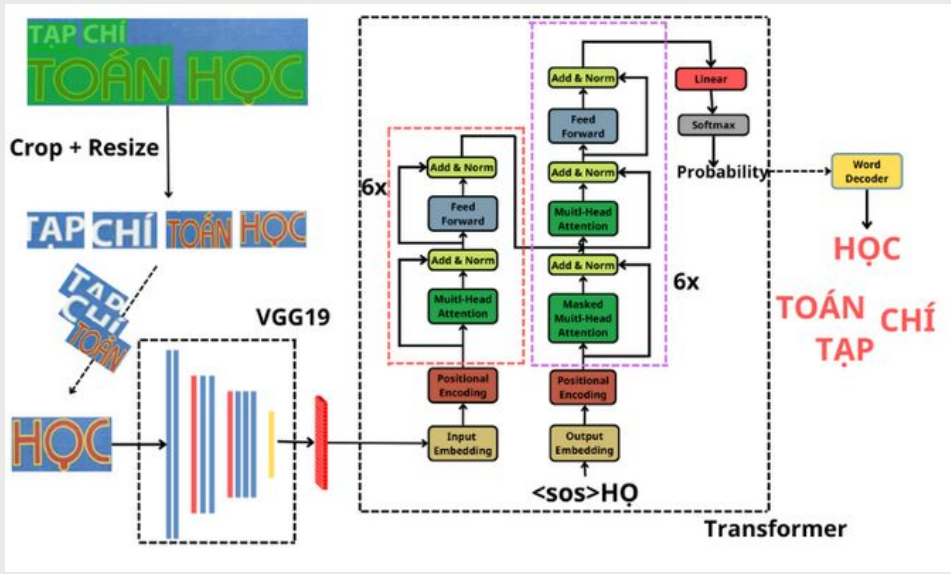
- This study proposes an advanced Scene Text Recognition (STR) framework that integrates the **VGG19** backbone for feature extraction with a Transformer architecture for sequence modeling.
- Propose a novel loss function named Cluster **Character Loss (CCL)**.

Why ?

- **Handling Visual Similarity**: To solve the problem where characters with similar structures are **easily confused** (e.g., '0', 'O', 'o'), requiring them to be clustered for better differentiation.
- **Training Awareness**: To ensure the model becomes **fully aware of confusion patterns and errors** during training. By explicitly penalizing these specific mistakes, the model is guided to avoid common visual misclassifications.

Overview





Description

1. Cluster Character Loss (CCL)

$$CCL = \frac{1}{N}\sum_{i=0}^{N}\sum_{j=1}^{min\{Wg_i, Wp_i\}} k_j$$

- $k_j = 0$ *if* $(Cp_j = Cg_j)$
- $k_j = 1$ *if* $\{(Cp_j \neq Cg_j)$ *and* $(Cp_j \cup Cg_j)$ *not in cluster*$\}$
- $k_j = \partial \in (0,1)$ *if* $\{(Cp_j \neq Cg_j)$ *and* $(Cp_j \cup Cg_j)$ *in cluster*$\}$
- $Cg_j \in Wg_i$ *and* $Cp_j \in Wp_i$
- Wg_i; Wp_i *is the word ground truth; prediction*
- Cg_j; Cp_j *is the word ground truth; prediction*

Figure 1. Formula to calculate Cluster Character Loss

Cluster Character Loss	Cg, p_1	Cg, p_2	Cg, p_3	Cg, p_4
Wg_i ='TOÁN'	T	O	Á	N
Wp_i ='TD4M'	T	D	4	M
k_j =	0	∂	∂	1

Figure 2. Example of calculating Cluster Character Loss

2. Total loss

$$TotalLoss = FocalLoss + ClusterCharacterLoss$$

$$TotalLoss = \frac{1}{N}\sum_{i=0}^{N}(1-p_i)^{\gamma} \times log(p_i) + \frac{1}{N}\sum_{i=0}^{N}\sum_{j=1}^{length(Wi)} k_j$$

Figure 1. Formula to calculate total loss

3. Experiment

- **Training**
  - We used VGG19 backbone to extract features and the Transformer model.
  - We training on 1500 images of VinText dataset with 20000 iterations.
  - Experiments include cross entropy loss and cluster character loss.
- **Other Problems**
  - **Out-of-Vocabulary:** we use IC13 dataset with 300 images has text instance is the English language not in the training vocabulary.
  - **Art-Text:** we use our dataset with 391 images has text instance is the style text and art text.

4. Results

	CCL	Accuracy	Levenshtein
VinText	✗	69.45%	20.10%
IC13	✗	38.72%	36.25%
Our dataset	✗	34.45%	43.28%
VinText	✓	**70.10%**	**20.08%**
IC13	✓	**39.82%**	**35.69%**
Our dataset	✓	**35.21%**	**43.58%**

Figure 4. Experimental results

NII
Nguyễn Khắc Thái – Trường Đại học Công nghệ thông tin
TEL : 086933**** Email: thaink.20@grad.uit.edu.vn