



ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
UIT-HCM

VIETNAMESE HANDWRITTEN RECOGNITION

Thành viên:

Trần Vĩ Hào

19521482

Lê Đăng Đăng Huy

19521612

Tô Thanh Hiền

19521490



Nội dung

- I. Mô tả bài toán
- II. Phương pháp tiếp cận
- III. Dataset
- IV. Evaluate

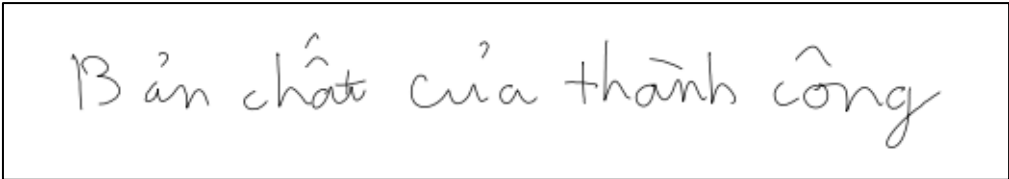
I. Mô tả bài toán

Input: Một bức ảnh chứa **một** dòng văn bản bất kỳ.

Output: Dòng văn bản tương ứng của bức ảnh ở dạng text.

VD:

Input:



Output:

Bản chất của thành công

II. Phương pháp tiếp cận

1. CRNN + CTC (1)

2. VietOCR (2)

3. TrOCR (3)

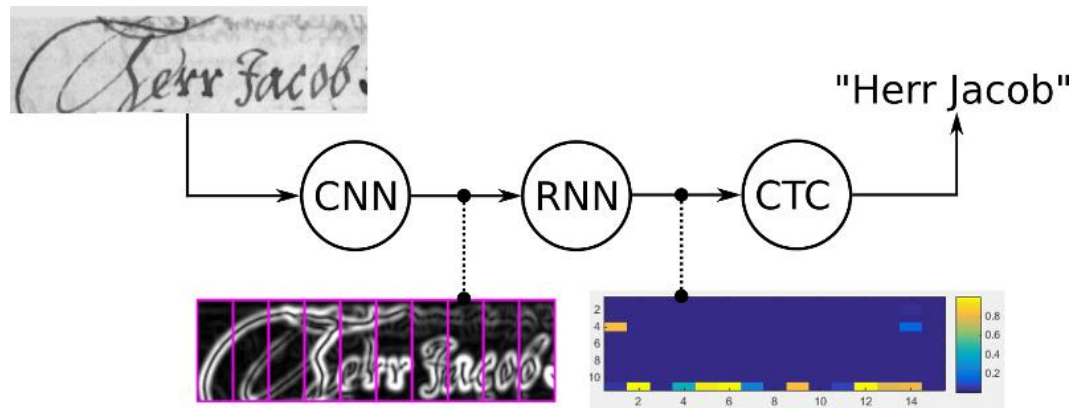
(1) [Baoguang Shi](#), [Xiang Bai](#), [Cong Yao](#): An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

(2) <https://github.com/pbcquoc/vietocr>

(3) [Minghao Li](#), [Tengchao Lv](#), [Lei Cui](#), [Yijuan Lu](#), [Dinei Florencio](#), [Cha Zhang](#), [Zhoujun Li](#), [Furu Wei](#) TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models

1. CRNN + CTC

- Nhóm sử dụng CRNN là sự kết hợp của CNN, RNN và CTC (**Connectionist Temporal Classification**) cho các tác vụ nhận dạng chuỗi dựa trên hình ảnh, như nhận dạng văn bản.
- Mô Hình gồm 3 thành phần chính:



[Baoguang Shi](#), [Xiang Bai](#), [Cong Yao](#): An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

1. CRNN + CTC

1.1. Feature Sequence Extraction với Convolution layers

- Mục tiêu: Trích chọn các đặc trưng của ảnh.
- Chúng ta sẽ sử dụng các mạng. Ta có thể sử dụng một số model CNN chuẩn như VGG, ResNet làm backbone.
- Output sẽ là các feature maps. Từ feature maps, ta tạo ra một chuỗi các features vector bằng cách reshape matrix thành feature vector để cho bước tiếp theo.

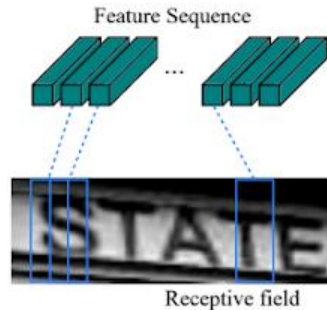
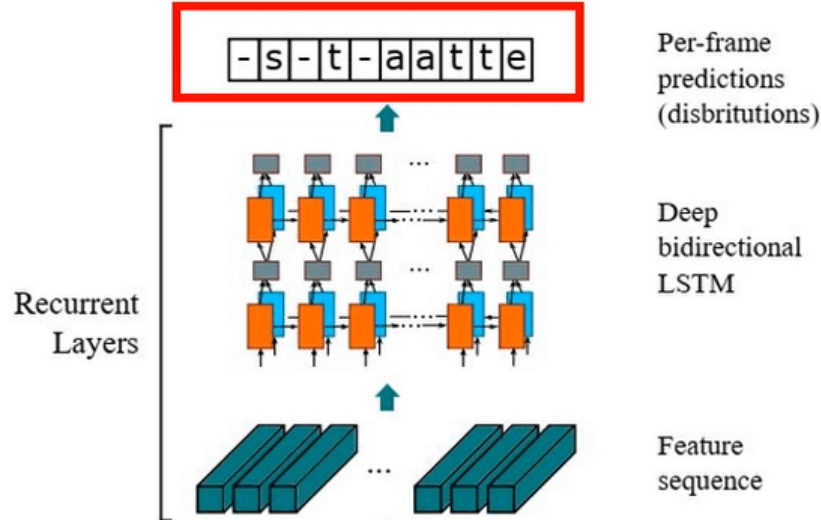


Figure 2. The receptive field. Each vector in the extracted feature sequence is associated with a receptive field on the input image, and can be considered as the feature vector of that field.

1. CRNN + CTC

1.2. Sequence Labeling với Recurrent layers

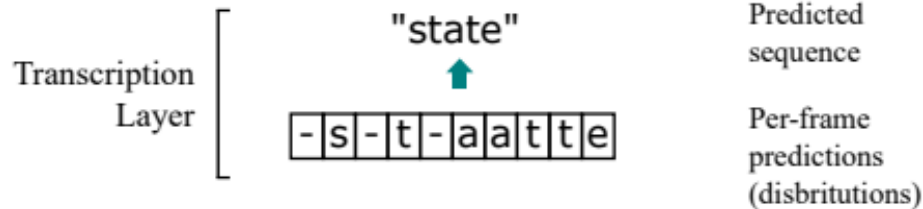
- Mục tiêu: đưa ra dự đoán phân bố nhãn cho từng frame một.
- Trong phần này, từ các feature vector x_1, x_2, \dots, x_T ở bên trước, ta sẽ output ra một phân phối nhãn y_t cho từng frame x_t .



1. CRNN + CTC

1.3. Transcription layers

- Mục tiêu là chuyển per-frame prediction của RNN thành final predicted sequence.
- Ta sẽ sử dụng CTC, CTC loss tức Connectionist Temporal Classification để giải quyết bài toán.



1. CRNN + CTC

1.4. Connectionist Temporal Classification

CTC giải quyết bằng cách đề xuất 1 loại ký tự là ký tự khoảng trắng, kí hiệu “-” hoặc “ ϵ ”, để tạo ra các alignment. Khi encoding text, chúng ta sẽ thêm rất nhiều ký tự trắng tùy ý vào các vị trí bất kỳ trong câu. Đồng thời, giữa 2 ký tự liền nhau và giống nhau, bắt buộc phải thêm khoảng trắng.

h h e ϵ ϵ l l l ϵ l l o

h e ϵ l ϵ l o

h e l l o

h e l l o

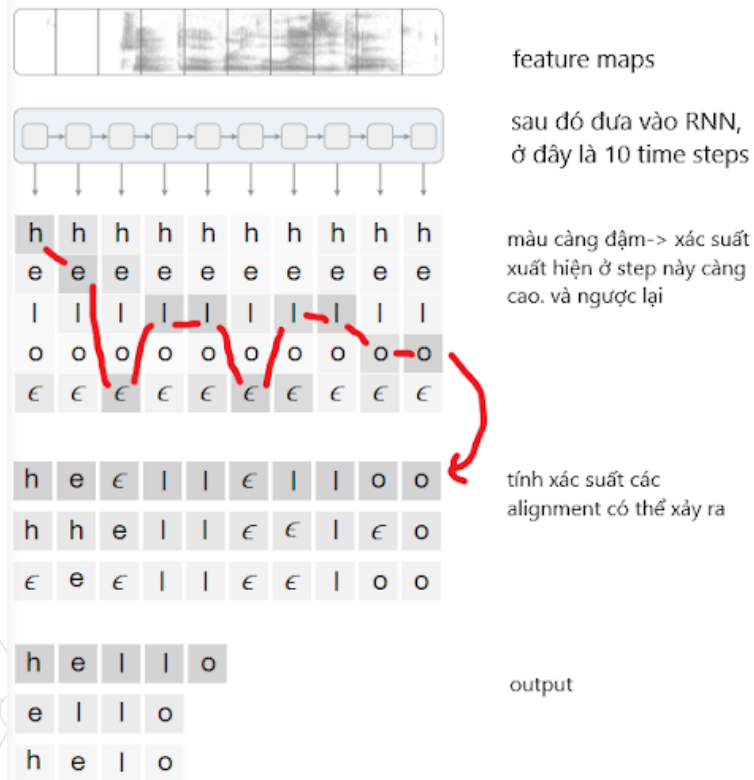
First, merge repeat characters.

Then, remove any ϵ tokens.

The remaining characters are the output.

1. CRNN + CTC

1.4. Connectionist Temporal Classification



- Lúc này, score của 1 đường đi qua tất cả các từ (hay còn gọi là 1 alignment) bằng tích score các điểm trên đường. Ở giai đoạn encode, ta sẽ tính toán tất cả các alignment có thể xảy ra, sau đó cộng chúng lại. Cuối cùng, chúng ta có được hàm loss.
- Sau khi có được hàm loss, chúng ta có thể tính toán gradient như thông thường. Tham số sẽ được điều chỉnh để minimize hàm negative log-likelihood.

1. CRNN + CTC

1.4. Connectionist Temporal Classification

Quá trình Decoder khá đơn giản với 2 steps:

1. Tìm alignment nào đi qua các ký tự có xác suất cao nhất trong từng time step.
2. Bỏ đi ký tự giống nhau liên tiếp, rồi sau đó mới bỏ đi các ký tự trắng.

⇒ Best path decoding.

Ngoài ra, còn có nhiều bộ decoder nâng cao hơn như beam search decoding, prefix-search decoding hay token passing,...



2. VietOCR

Như trên

Như trên;

Viêm da khác

Viêm da khác

LÊ VĂN THỨC

LÊ VĂN THỨC

Repurchases

Repurchases

Nguyễn Thị Hồng Hiến

Nguyễn Thị Hồng Huệ

1.21

1.21

khởi nghĩa lam sơn gồm ba giai đoạn lớn:
hoạt động ở vùng núi thanh hoá
(1418-1423)

MONIXER

MONIXER

038071004740

038071004740

ngoài ra rất phổ biến các loại rượu ngâm
hồn hợp nhiều loại động

ngoài ra rất phổ biến các loại rượu ngâm hồn hợp nhiều loại động

Silliest

Silliest

NGUYỄN THỊ NGOÃN

NGUYỄN THỊ NGOÃN

REAL MADRID BẤT NGỜ ĐƯỢC TRONG TÀI CHO
HƯỞNG MỘT QUẢ PENALTY GÂY TRANH CÃI

REAL MADRID BẤT NGỜ ĐƯỢC TRONG TÀI CHO HƯỞNG MỘT QUẢ PENALTY GÂY TRANH CÃI

she more than doubled the party's vote
in the constituency

she more than doubled the party's vote in the constituency

Thanh Xuân - Hà Nội - Việt Nam

Thanh Xuân - Hà Nội - Việt Nam

these have come at a cost of selling
land

001085019081

001085019081

030010001197

030010001197
Quốc gia: Việt Nam

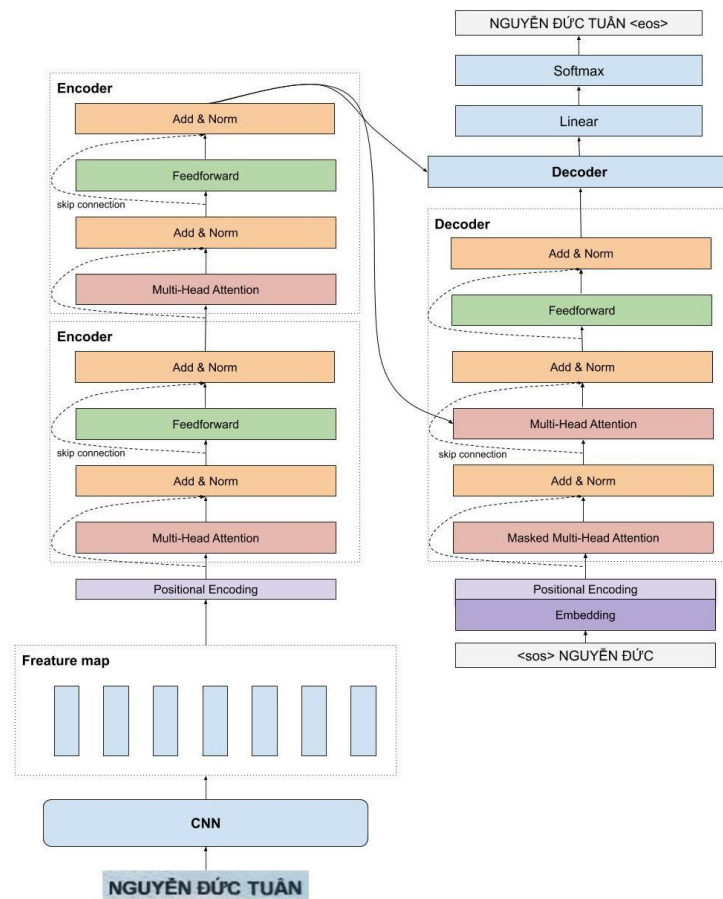
2006 And Was Featured On Tech News Blog
Techcrunch [1]

2006 And Was Featured On Tech News Blog Techcrunch [1]

Trọng Đó Có Tổng Bình

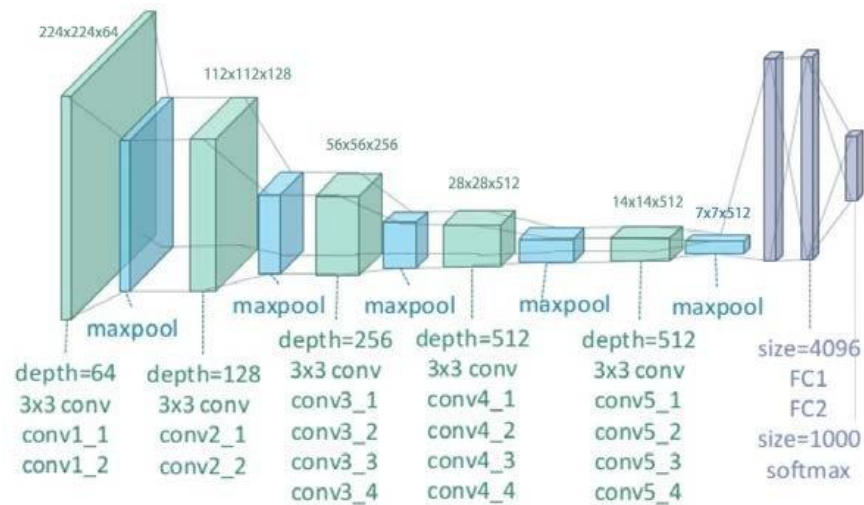
TRỌNG ĐÓ CÓ TỔNG BÌNH

2. VietOCR



2. VietOCR

Backbone



3. TrOCR

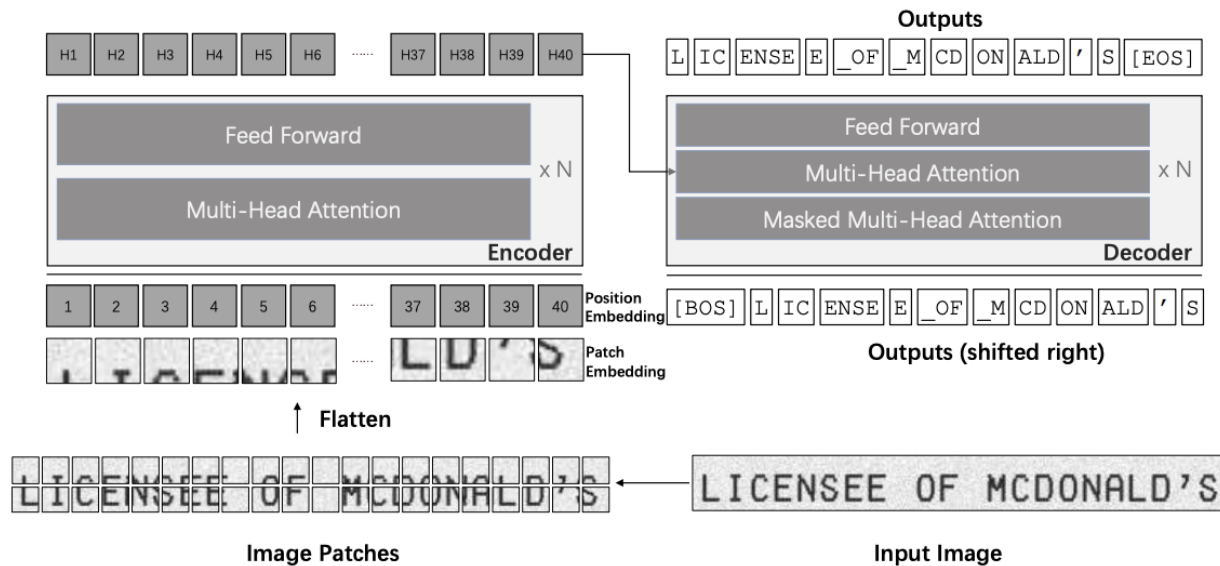


Figure 1: Model Architecture of TrOCR, where an encoder-decoder model is designed with a pre-trained image Transformer as the encoder and a pre-trained text Transformer as the decoder.

3. TrOCR

3.1. Encoder Initialization

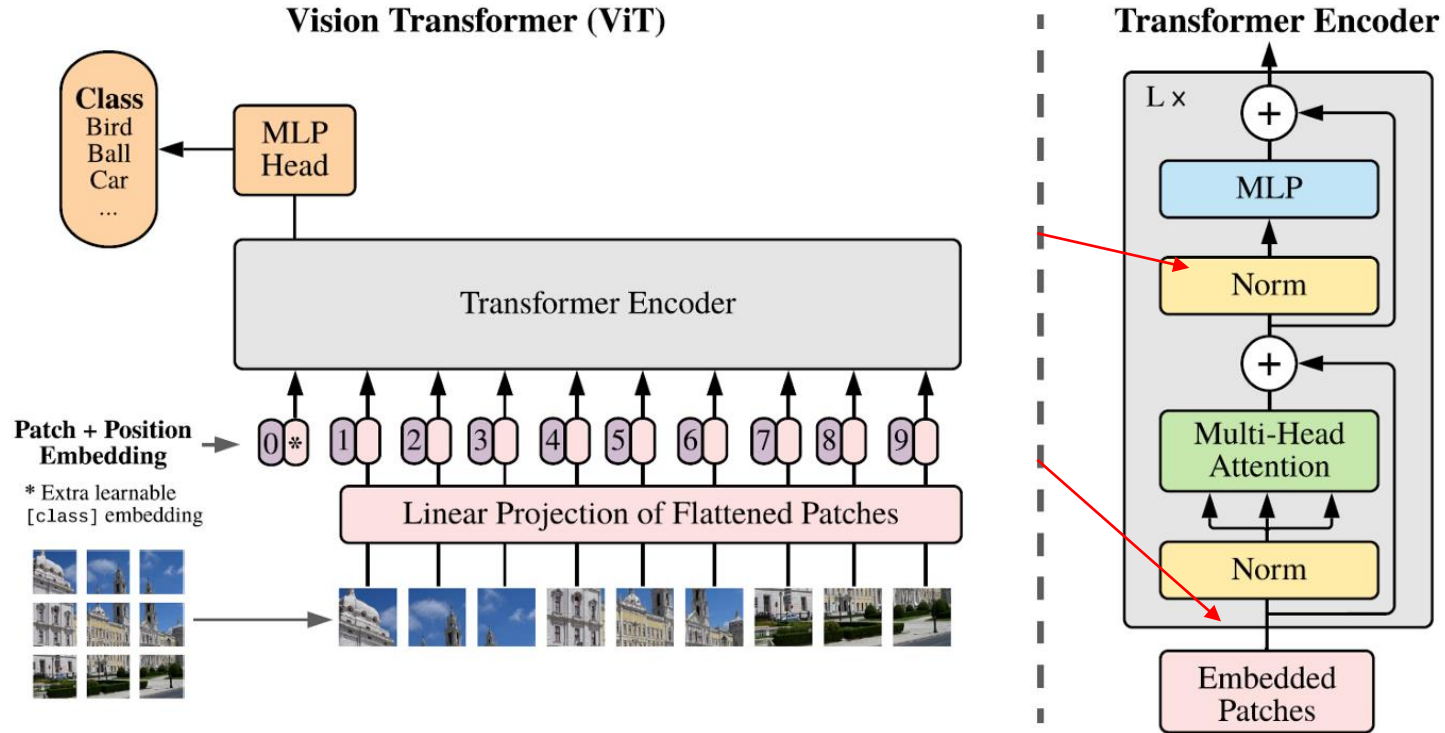
- The DeiT (Touvron et al., 2021a) models are used for the encoder initialization in the TrOCR models

3.2. Decoder Initialization

- RoBERTa models are used for the decoder initialization in the TrOCR models. The encoder-decoder attention layers are absent in the RoBERTa models. To address this, the decoders are initialized with the RoBERTa models and the absent layers are randomly initialized.

3. TrOCR

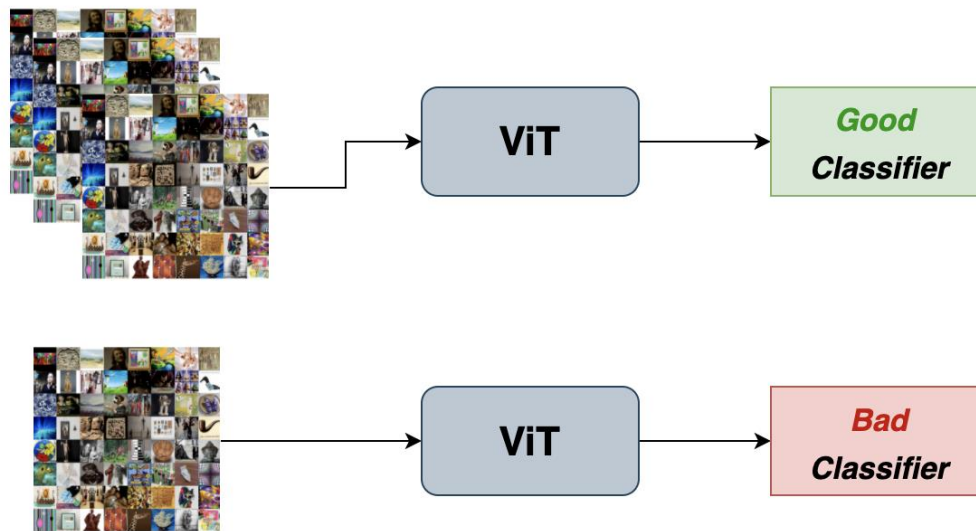
3.3. Vision transformer (ViT)



3. TrOCR

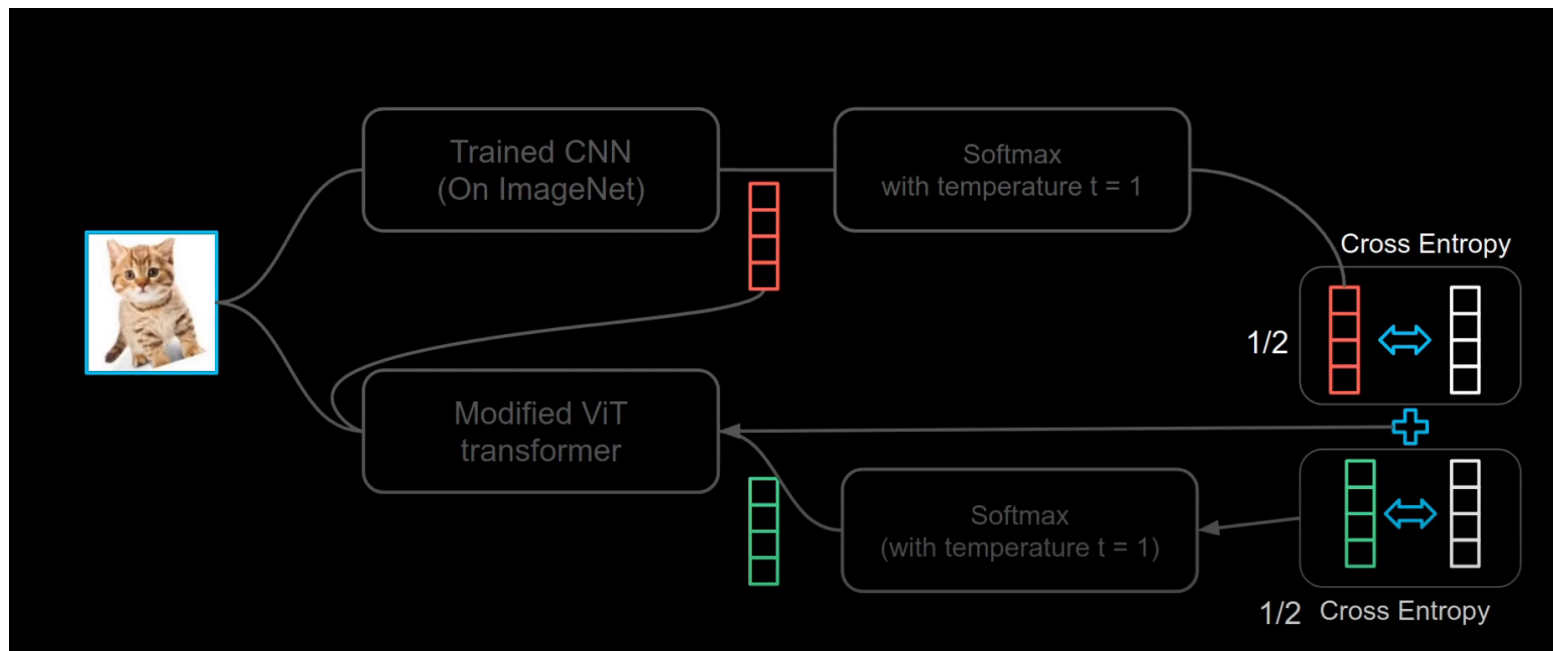
3.3. Vision transformer (ViT)

❖ Vấn đề với Vision Transformer.



3. TrOCR

3.4. Data-efficient image transformer (DeiT)



[Hugo Touvron](#), [Matthieu Cord](#), [Matthijs Douze](#), [Francisco Massa](#), [Alexandre Sablayrolles](#), [Hervé Jégou](#) Training data-efficient image transformers & distillation through attention

3. TrOCR

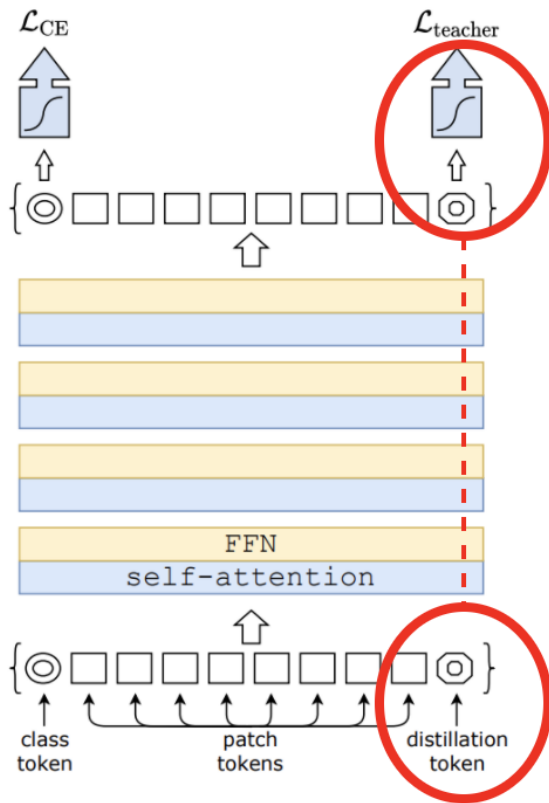
3.4. Data-efficient image transformer (DeiT)

Hard Distillation

$$\mathcal{L}_{\text{global}}^{\text{hardDistill}} = \frac{1}{2} \mathcal{L}_{\text{CE}}(\psi(Z_s), y) + \frac{1}{2} \mathcal{L}_{\text{CE}}(\psi(Z_s), y_t).$$

3. TrOCR

3.4. Data-efficient image transformer (DeiT)



3. TrOCR

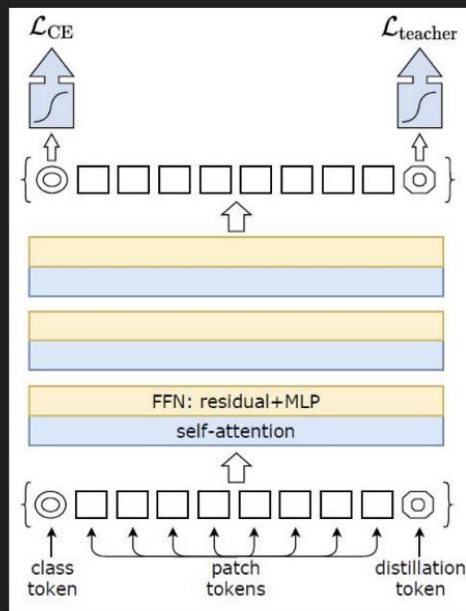
3.4. Data-efficient image transformer (DeiT)

DeiT Architecture¹

True Label →



Size 16 x 16 patches



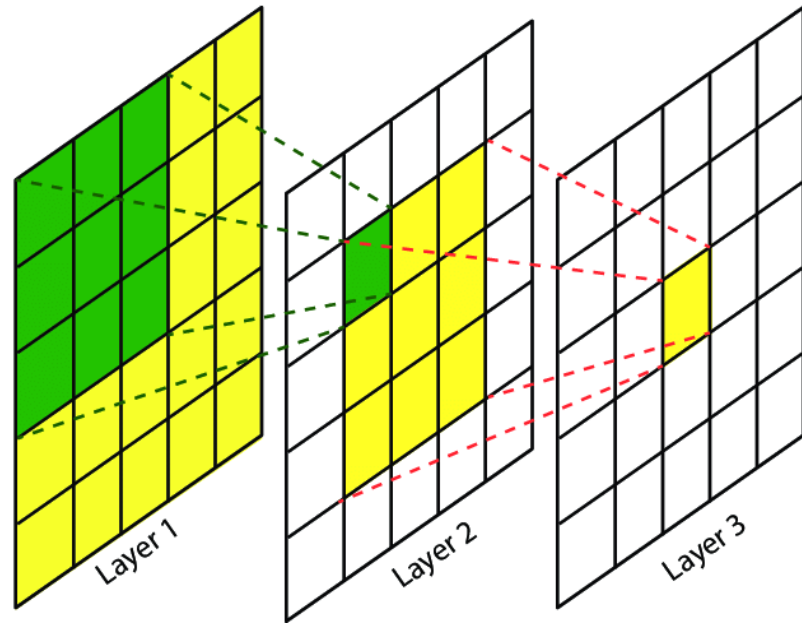
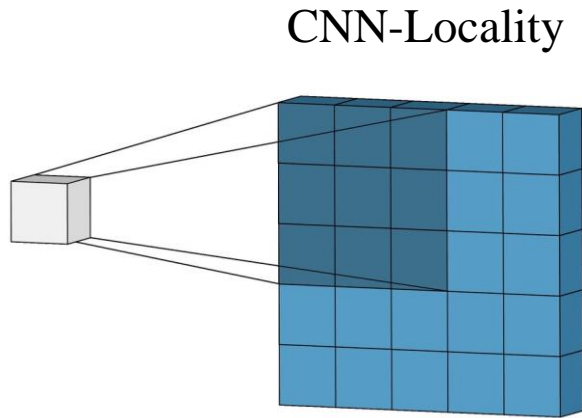
Vision Transformer

RegNet Y-16GF Prediction
(84M Params)⁴

Class and distillation tokens
are learned by backprop.

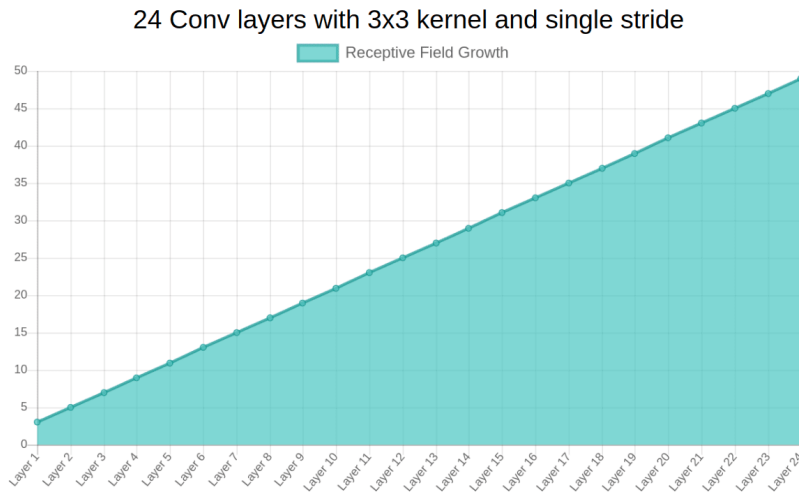
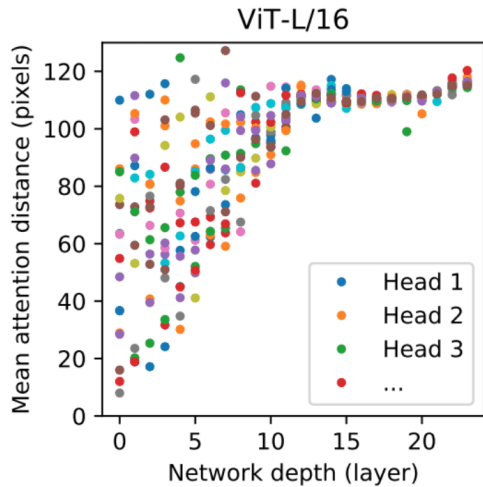
3. TrOCR

Tại sao lại cân nhắc sử dụng Transformer làm backbone?



3. TrOCR

Tại sao lại cân nhắc sử dụng Transformer làm backbone?



3. TrOCR

Tại sao lại cân nhắc sử dụng Transformer làm backbone?

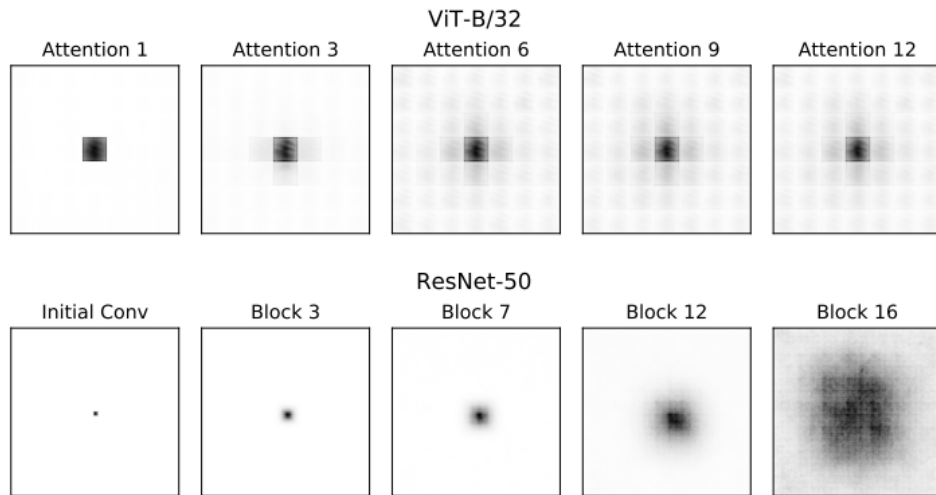
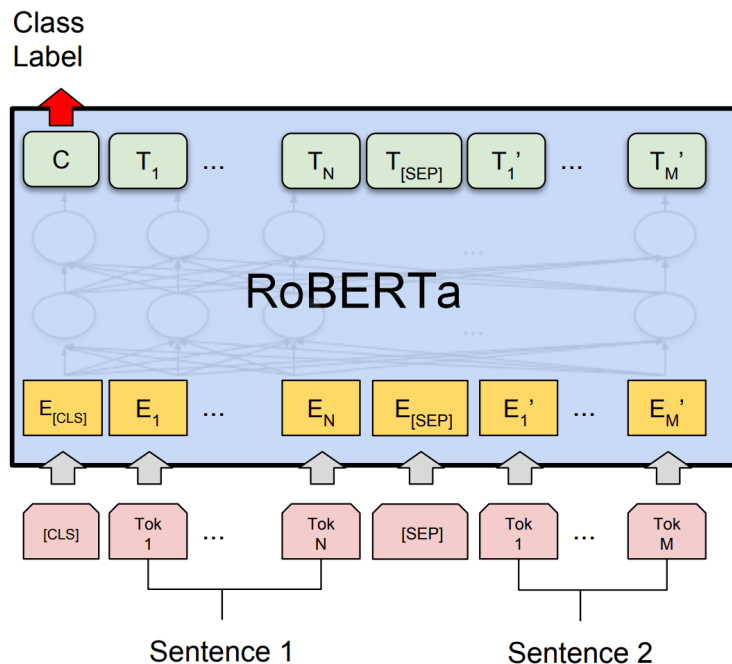


Figure 6: ResNet effective receptive fields are highly local and grow gradually; ViT effective receptive fields shift from local to global. We measure the effective receptive field of different layers as the absolute value of the gradient of the center location of the feature map (taken after residual connections) with respect to the input. Results are averaged across all channels in each map for 32 randomly-selected images.

3. TrOCR

3.6. Robustly Optimized BERT Pretraining Approach (RoBERTa)



[Yinhan Liu](#), [Myle Ott](#), [Naman Goyal](#), [Jingfei Du](#), [Mandar Joshi](#), [Danqi Chen](#), [Omer Levy](#), [Mike Lewis](#), [Luke Zettlemoyer](#), [Veselin Stoyanov](#)
RoBERTa: A Robustly Optimized BERT Pretraining Approach

III. Dataset

Dataset được lấy từ ICFHR2018 Competition on Vietnamese Online Handwritten Text Recognition Database bao gồm 7282 bức ảnh.

VNOnDB

giữ được chính thức ra sân. Nhưng đó không phải là thất bại. Trái lại, thành công

thời. Cuộc sống vẫn chào đón họ với NV2, NV3. Quan trọng là họ đã nỗ lực hết sức để

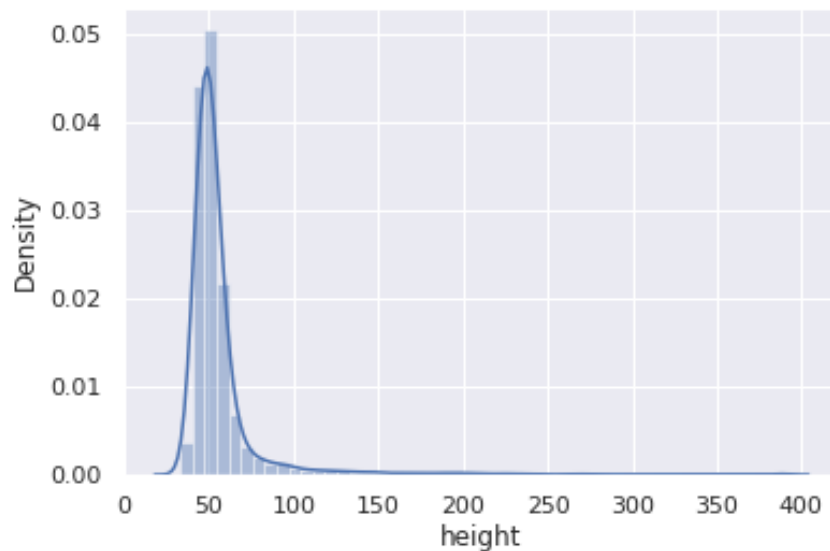
Bản chất của thành công

III. Dataset

Width

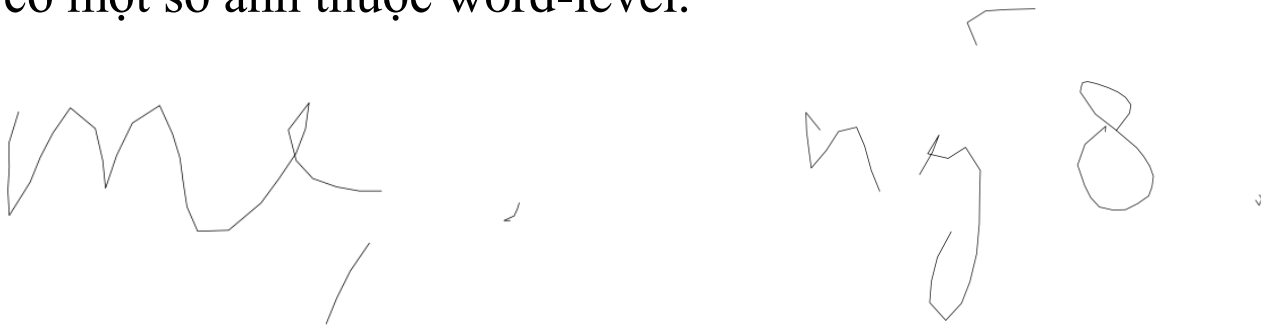
99% ảnh của dataset có width là 515px

Height



III. Dataset

Dataset có một số ảnh thuộc word-level.



Dataset có một số ảnh không crop vào giữa ảnh.

vali

mũi thuốc thông tiêu

Họ lặn cả ngày, có khi cả đêm, dưới biển lạnh trên dưới 30C.

IV. Evaluate

Character Error Rate (CER)

$$CER = \frac{S + D + I}{N}$$

- S = Number of Substitutions.
- D = Number of Deletions.
- I = Number of Insertions.
- N = Number of characters in reference text (a.k.a ground truth).

IV. Evaluate

Word Error Rate (WER)

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

- S = Number of Substitutions.
- D = Number of Deletions.
- I = Number of Insertions.
- N = Number of characters in reference text (a.k.a ground truth).

IV. Evaluate

CRNN + CTC

▶ prediction[309]

↳ 't àn tàn tàn tàn.'

	CER	WER
VietOCR	0.119	0.294
TrOCR	0.092	0.213

IV. Evaluate

Một số trường hợp model nhận diện tặc:

tra kiểm soát giao + hàng đưa bàn Dầu Giây, Trảng Bom, Thống Nhất, Đồng Nai - đồng

vietOCR: tra thành báo giao xrong đạo hồi đến chẳng, tăng tin, những thiếu đồng tin đi gia
trOCR: tra biển rản giao - trong đưa bàn Dầu Giây, Trảng Bom, Thống Nhất, Đồng khu thành

giám lịch. Theo đó, muốn tôi đồng bỏ trực những cá chình nghiệp, cơ quan

vietOCR: giám lịch. " Theo bị, mùa thời bỏ sang là trực dân với chả manghiện, có giao Yêu

trOCR: giám được. Theo đó, muốn tôi đồng bỏ trực những cá chình nghiệp, cơ quan



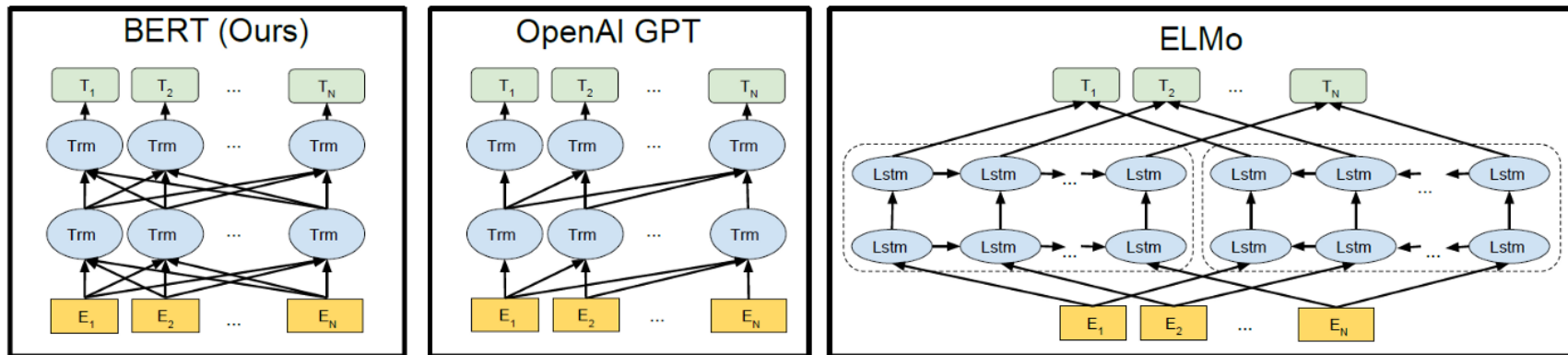
Thank you!

Any questions?

3. TrOCR

3.5. Bidirectional Encoder Representations from Transformers (Bert)

Architecture



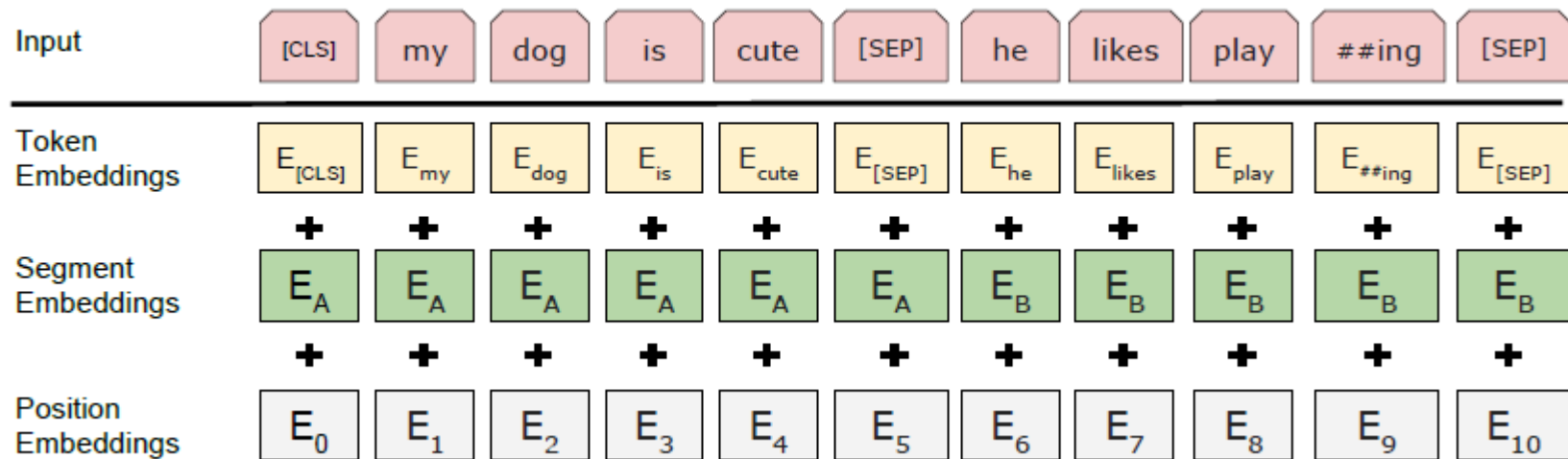
[Jacob Devlin](#), [Ming-Wei Chang](#), [Kenton Lee](#), [Kristina Toutanova](#)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

3. TrOCR

3.5. Bidirectional Encoder Representations from Transformers (Bert)

Input Representation

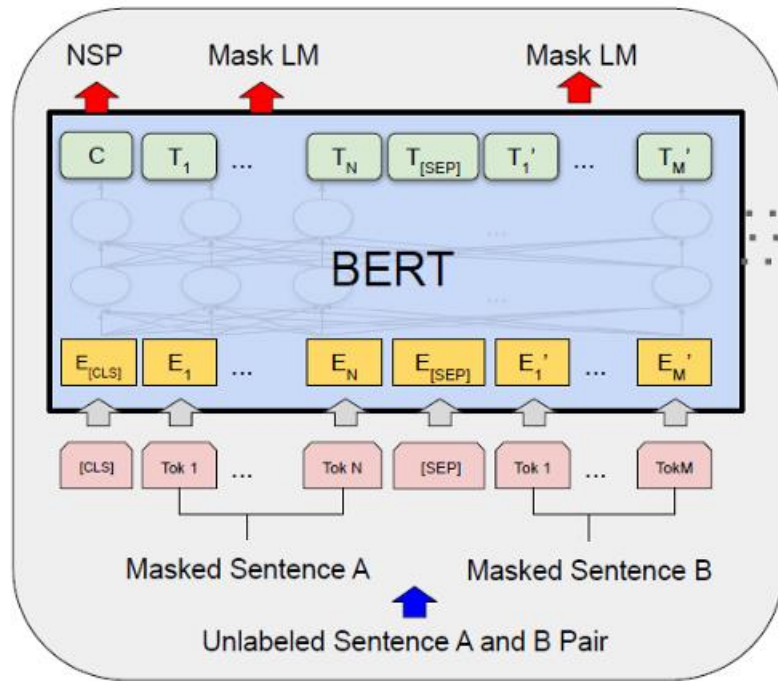


3. TrOCR

3.5. Bidirectional Encoder Representations from Transformers (Bert)

Pretraining BERT

1. Masked LM (MLM)
2. Next Sentence Prediction (NSP)

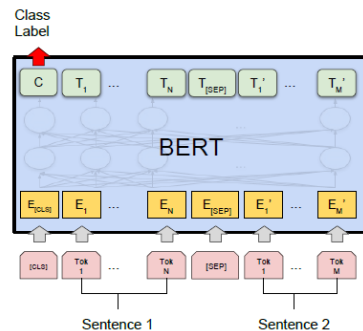


Pre-training

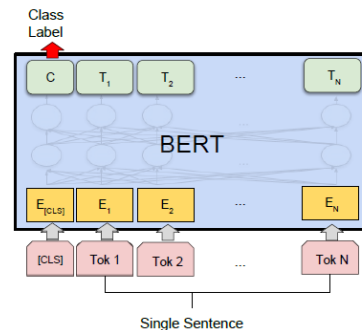
3. TrOCR

3.5. Bidirectional Encoder Representations from Transformers (Bert)

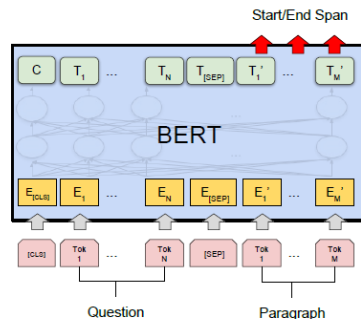
Fine-tune Bert



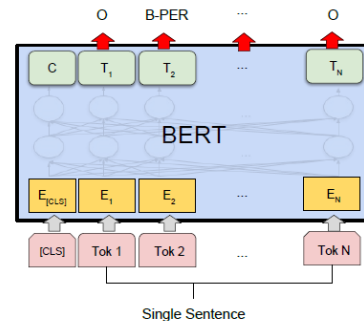
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER