Scene Text Spotting

Scene Text Spotting

END-TO-END SCENE TEXT DETECTION & RECOGNITION

End-to-End Scene Text Detection & Recognition

Text Detection

Bounding box
[[192.5, 339.0],
[567.4, 344.7],
[566.7, 398.6],
[191.7, 393.0]]

Bounding box

RoI
Rotate

Text Recognition

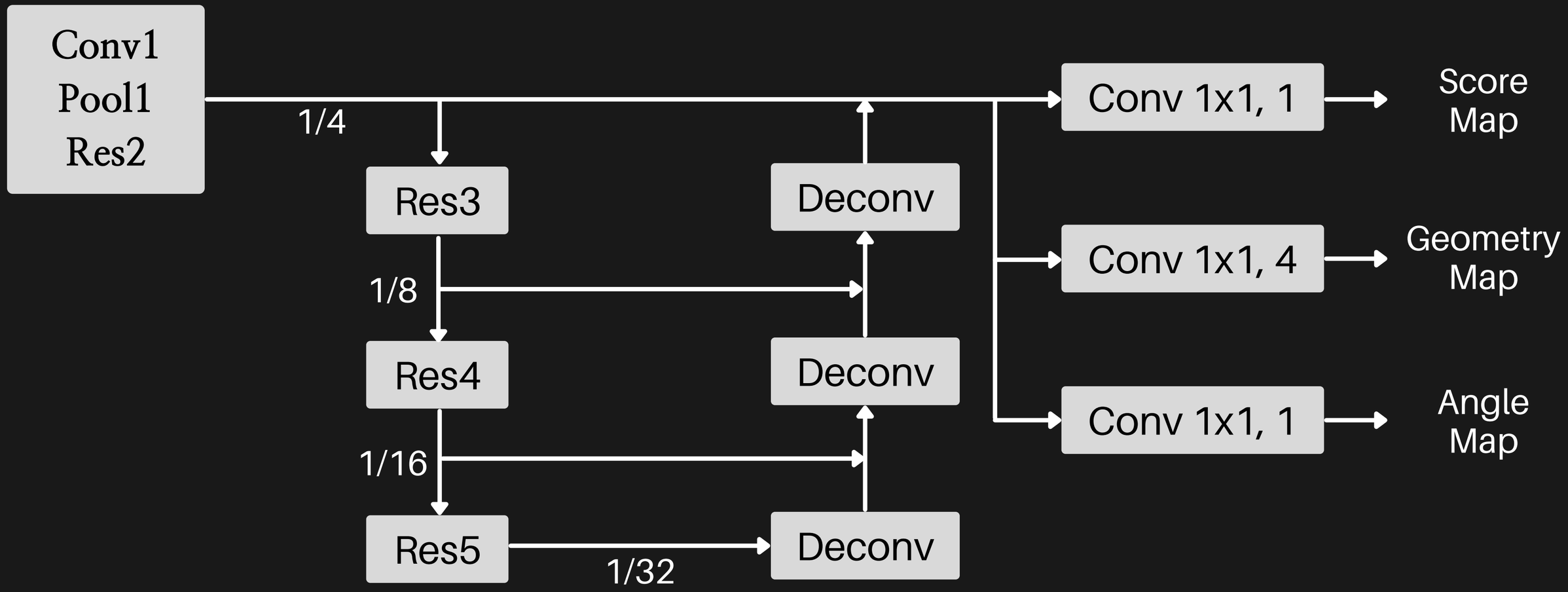
TEXT
Mc Donald's

- **affine transformation**
- **bilinear interpolation**

Text Detection

#FCN

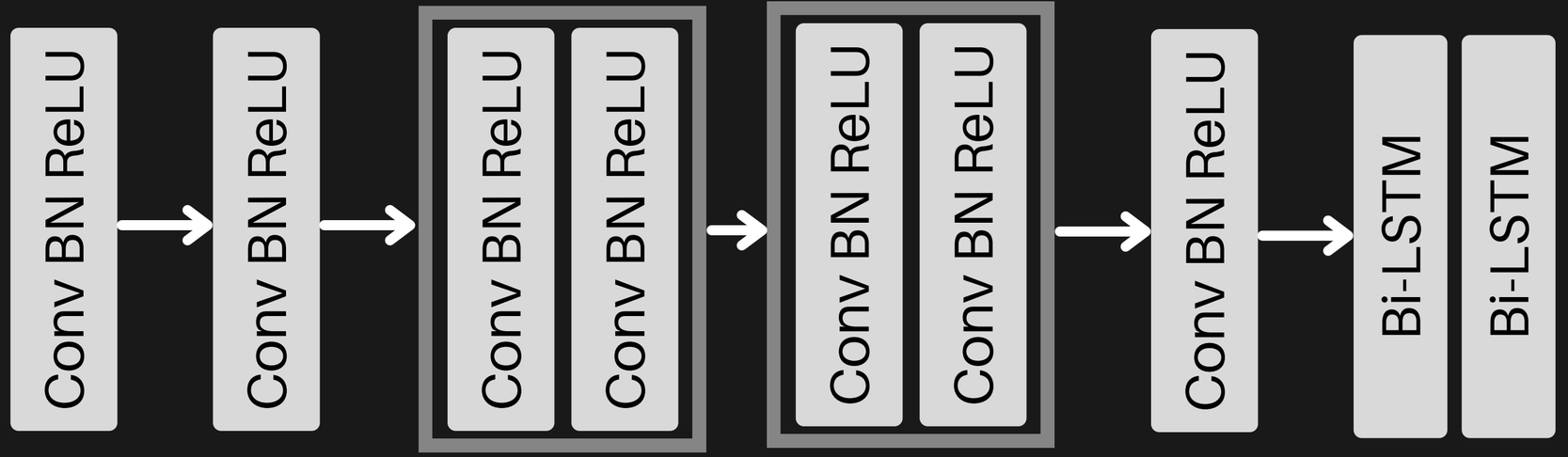
Text Detection ResNet FCN + EAST



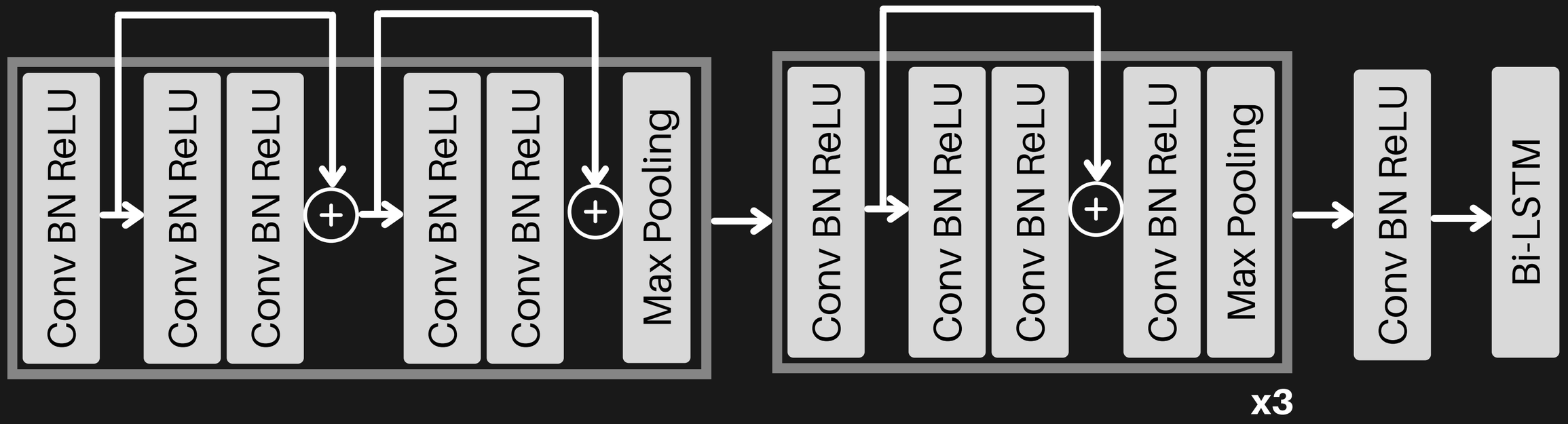
Text Recognition

#CRNN #R-NET

CRNN


Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Bi-LSTM
Bi-LSTM

R-Net


Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Max Pooling
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Conv BN ReLU
Max Pooling
x3
Conv BN ReLU
Bi-LSTM

OE (OCR in Editor)

#VIDEO-EDITOR #R-NET

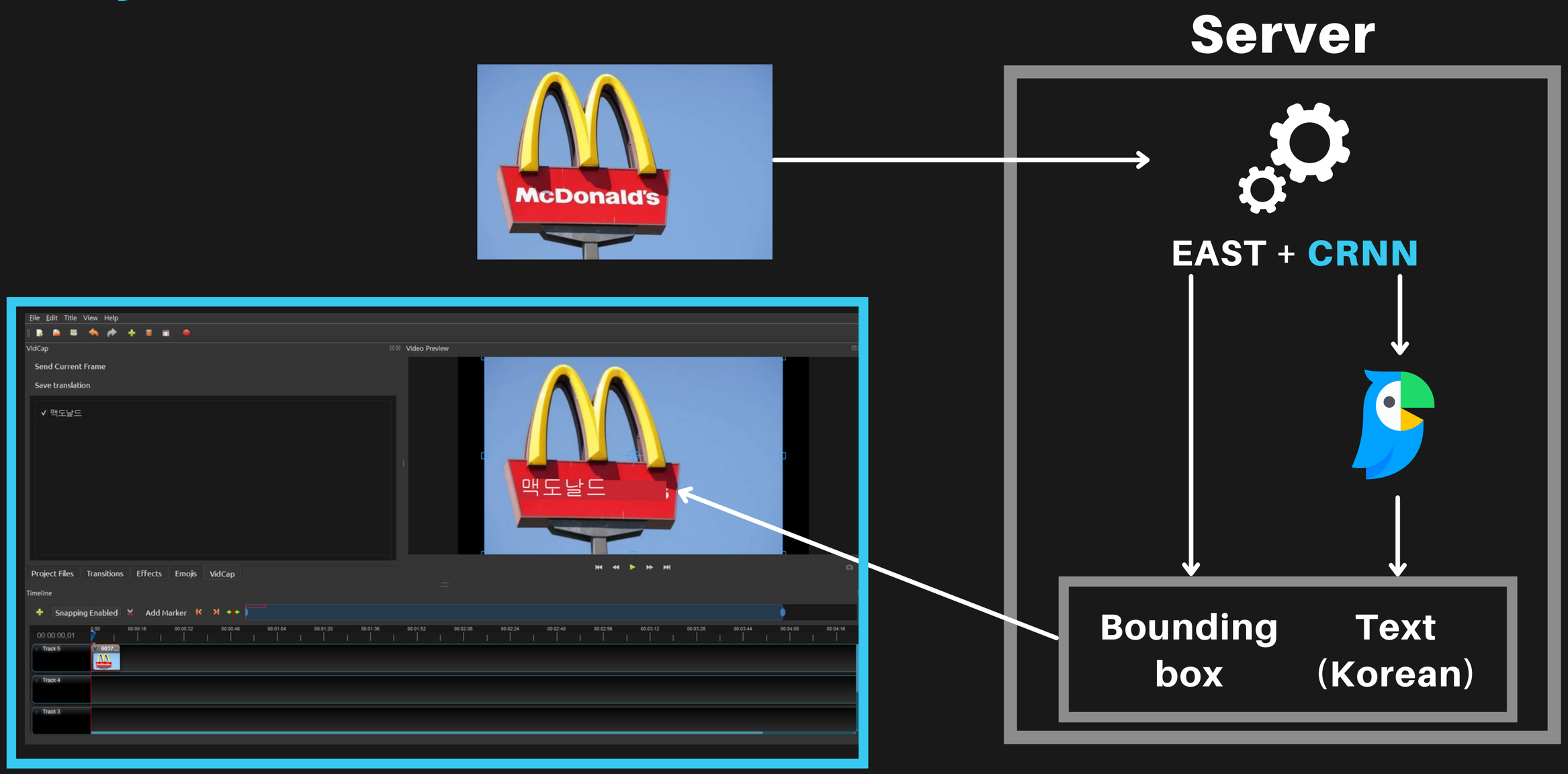
Goals

- Test F1 score >= 0.6
- Inference time <= 10s

Achievement

- Test F1 score
  - Input image size < 800기준
  - 0.62
- Inference time
  - Bounding box 3개 기준
  - 1~2초 소요

Project Overview



CRNN vs R-Net

	CRNN	R-Net
Depth	1x	2.5x
Residual Connection	X	O
Pretained weight	O	X

- 같은 조건에서는 CRNN(0.37)에 비해 R-Net(0.43)의 성능이 확실히 좋았음
  - 깊은 depth + residual connection(Gradient vanishing 문제 예방) 존재로 성능이 더 높을 것으로 예상해서 적용
- Pretrained weight를 적용 후, CRNN과 R-Net의 성능은 비슷해짐
- 두 모델에 제일 잘 되는 hyperparameter 적용 후의 성능으로 최종 모델로 결정

R-Net Hyperparameter Experiments

Dataset	optimizer	Learning Rate	Recog Input	Valid F1
ICDAR 15	Adam	0.0001	(16, 128)	0.4507
ICDAR 15	Adam	0.0001	(32, 180)	0.5064
ICDAR 15	AdamW	0.0001	(32, 180)	0.5988
ICDAR 15+17	AdamW	0.0001	(32, 180)	0.6721

=> Test F1 Score : 0.631