PreSTU: Pre-Training for Scene-Text Understanding

Jihyung Kil[1], Soravit Changpinyo[2], Xi Chen[2], Hexiang Hu[2], Sebastian Goodman[2], Wei-Lun Chao[1], Radu Soricut[2]
The Ohio State University[1], Google Research[2]

Highlights

- Propose **PreSTU**, a simple and effective pre-training recipe with OCR-aware objectives for scene-text understanding
- **PreSTU** encourages models to recognize text from an image and connect what is recognized to the rest of the image content
- **PreSTU** leads to improved STU on twelve diverse downstream VQA and image captioning tasks

Scene-text understanding (STU)

Understand the role of text in the context of a visual scene

Q: What does it say near the star on the tail of the plane?
A: Jet

Scene-text VQA

A tile wall with a red circle on it reading Mornington Crescent

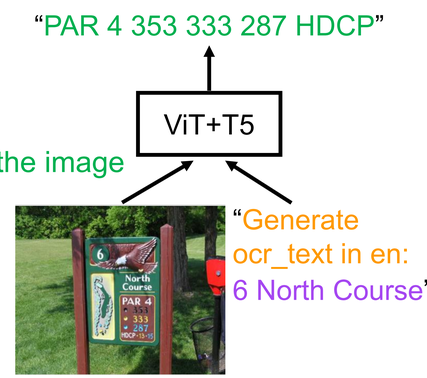
Scene-text Captioning

STU Challenges

- Models should learn two capabilities to solve STU tasks:
  i. Recognizing text in a visual scene
  ii. Connecting the text to its context in the scene
- Prior works have not explored pre-training objectives, targeting (i) and (ii). Simply use general V&L objectives (e.g., visual language modeling)
- Even in systems performing STU pre-training, they are not designed to learn both (i) and (ii)

PreSTU

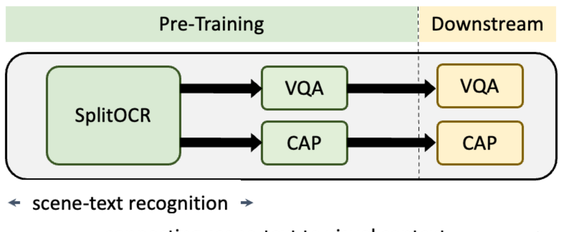
Pre-training recipe with OCR-aware objectives to learn two essential STU capabilities,
i. Recognizing text in a visual scene ii. Connecting the text to its context in the scene



Objectives

- **SplitOCR**: Given (image, prompt, first some OCR tokens), generate the rest of OCR tokens in the image
  - By predicting OCR tokens, models learn to recognize scene text
  - By giving OCR tokens and image as input, learn to connect scene text to its visual context
- **VQA**: Given (image, prompt, question, OCR tokens), generate an answer
- **CAP**: Given (image, prompt, OCR tokens), generate a scene-text caption
  - VQA/CAP further learn the connection & ease the knowledge transfer to downstream tasks with the same input/output format

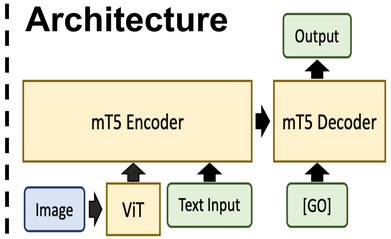
Pipeline



Input / Output

Objective	Text Input	Output
SplitOCR	Generate ocr_text in en: <OCR_1> <OCR_2>...<OCR_m>	<OCR_{m+1}>...<OCR_N>
VQA	Answer in en: <Question> <OCR_1> <OCR_2>...<OCR_N>	<Answer>
CAP	Generate alt_text in en: <OCR_1> <OCR_2>...<OCR_N>	<Caption>

PreSTU Architecture

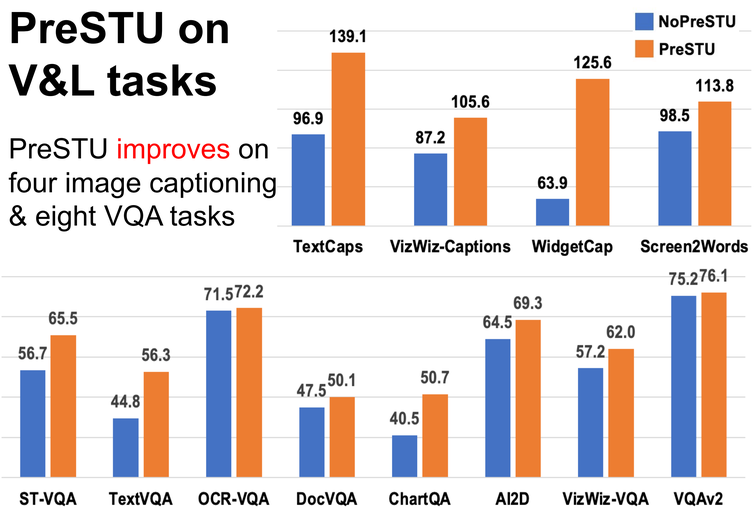


Pre-training Dataset

- SplitOCR/CAP: CC15M w/ OCRs
- VQA: ST-VQA w/ OCRs

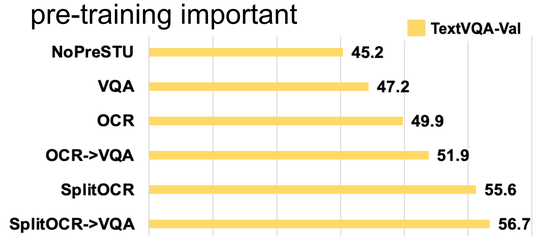
PreSTU on V&L tasks

PreSTU improves on four image captioning & eight VQA tasks



Detailed Ablation

- OCR: Predicting whole OCR tokens
- SplitOCR > OCR: SplitOCR balances two STU capabilities while OCR focuses too much on recognizing scene text
- SplitOCR->VQA best: Two-stage pre-training important



Applicability to SOTA

SplitOCR was used as one of pre-training objectives for PaLI-X, SOTA model on most V&L tasks