Implicit Supervision in Vision and Language Tasks

VL Corpus

“ group of people protesting in the street … ”

“ A man swing the baseball stick and a man in red topping standing …

“ several people in black suits are hosting a press … “

“ athletes are jumping into the pool … “

Data: Image + Caption or Image Level Tag

VL Representation Learning from image-text pairs

VL Corpus → VL Representation

Fine-tuning and transferring to downstream tasks

VQA

Q: “ What are they doing?
A: “ playing basketball. ”

Image captioning

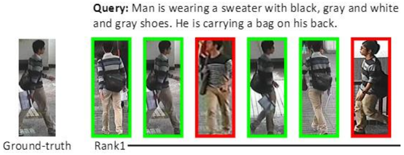
“ The is a man playing basketball … ”

Image/text retrieval

“ man playing basketball”
“ people are enjoying food”
“ few guys are talking”

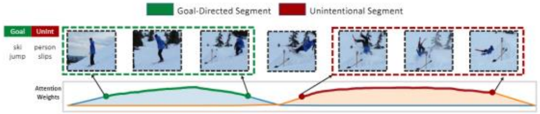
Task-specific Semantic Correspondence Learning

Person Search by Language

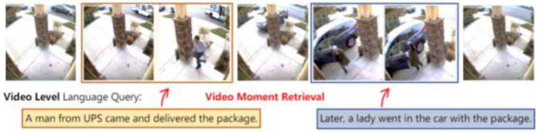


Query: Man is wearing a sweater with black, gray and white and gray shoes. He is carrying a bag on his back.

Human Activity Grounding



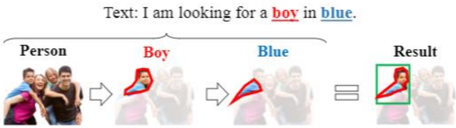
Video Moment Localization



A man from UPS came and delivered the package.

Later, a lady went in the car with the package.

Textual Grounding Task



Text: I am looking for a boy in blue.