Weakly Supervised Attention Learning for Textual Phrases Grounding

Zhiyuan Fang\textsuperscript{1}, Shu Kong\textsuperscript{2}, Tianshu Yu\textsuperscript{1}, and Yezhou Yang\textsuperscript{1}

Ira A. Fulton Schools of Engineering, Arizona State University\textsuperscript{1}
University of California, Irvine\textsuperscript{2}
Two Branches:
- Visual feature extraction
- Textual embedding stream

Attention Generation & Shifting:
- Visual and textual features are concatenated through bilinear pooling
- Attention maps are applied back on visual stream for weakly supervised learning

Weakly Supervised Feature Learning

Weakly Supervised Attention Learning for Textual Phrases Grounding, Zhiyuan Fang, Shu Kong, Tianshu Yu, and Yezhou Yang