Aligning books and movies: Towards story-like visual explanations by watching movies and reading books

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, Sanja Fidler

The IEEE International Conference on Computer Vision (ICCV), 2015, pp. 19-27

Abstract

Books are a rich source of both fine-grained information, how a character, an object or a scene looks like, as well as high-level semantics, what someone is thinking, feeling and how these states evolve through a story. This paper aims to align books to their movie releases in order to provide rich descriptive explanations for visual content that go semantically far beyond the captions available in the current datasets. To align movies and books we propose a neural sentence embedding that is trained in an unsupervised way from a large corpus of books, as well as a video-text neural embedding for computing similarities between movie clips and sentences in the book. We propose a context-aware CNN to combine information from multiple sources. We demonstrate good quantitative performance for movie/book alignment and show several qualitative examples that showcase the diversity of tasks our model can be used for.

Related Material

[pdf]
[bibtex]
@InProceedings{Zhu_2015_ICCV,  
author = {Zhu, Yukun and Kiros, Ryan and Zemel, Rich and Salakhutdinov, Ruslan and Urtasun, Raquel and Torralba, Antonio and Fidler, Sanja},  
title = {Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books},  
booktitle = {The IEEE International Conference on Computer Vision (ICCV)},  
month = {December},  
year = {2015}  
}  

To align movies and books we exploit a neural sentence embedding that is trained in an unsupervised way from a large corpus of books, as well as a video-text neural embedding for computing similarities between movie clips and sentences in the book. We propose a context-aware CNN to combine information from multiple sources. We demonstrate good quantitative performance for movie/book alignment and show several qualitative examples that showcase the diversity of tasks our model can be used for. Subjects: Computer Vision and Pattern Recognition (cs.CV); Computation and Language (cs.CL). Cite as: