Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books

Authors: Yukun Zhu ,Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, Sanja Fidler.
Publisher: Computing Research Repository
Year: 2015

Keywords: CRF, Visual-sentence embedding, Skip thought vectors, Microsoft CoCo.

1. **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 current datasets. 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."

2. **Introduction:**
   - It is required to upgrade the models to deal with the visual information and the language data at a time to understand the human behavior consistently.
   - For this purpose the authors had considered an approach to match the movie scenes and corresponding book sentences (or paragraphs) which could create an architecture to improve the model understanding to the virtual world.
   - As the books define very fine details of the character and circumstances, and movies depict them virtually, the authors thought it would be easier and practical to work with such approach.
   - Now, the authors needed to accomplish two goals:
     - The first is that the similar sentences has to be matched.
     - The scene with corresponding sentence has to be matched.
   - So, the authors adapted DVS and trained the model on that to achieve the purpose.

3. **Related Works:**

- Most of the similar works had been done in relating the scene with corresponding subtitles called as image captioning.
- These approaches had been done with the help of RNNs, with the help of combined image-text embedding.
- One of the closest prior research results in ordering the summary based on the story visual content retrieval.
- The parallel works also had been done to arrange the scenes based on the story of the corresponding books. However, they concentrated on coarser chapter level.
- On the other hand the authors concentrated on the finer sentence level matching and understanding.

4. Content and Approach:

- The initial requirement of the approach, which had not been before regarding this method is the dataset. So, the authors concentrated on creating the dataset based on their requirements at first.
- The authors were also needed two different datasets based on two goals through the procedure. One dataset is moviebook dataset in which 11 books were aligned with the corresponding movie scenes.
- The other one the Book corpus dataset in which only the books were collected from around 16 genres such as Romance, Sci-fi, Fantasy and teen etc.,
- For the former dataset, the assigning was done accordingly the scene timings and the sentences in the books describing that scene. This dataset had been very diverse varying in the number of sentences, length of sentences and styles etc.,
- Initially in the process of application, the subtitles of the scenes were taken and needed to match with the corresponding sentences of the book. This had been the first goal to be achieved.
- A simple Conditional Random Field is introduced by the authors to create a linear timeline and match between the movies and the books.
- This CRF also useful in matching the scenes with books, even when they are random.
- For the first aim of matching the subtitles with the sentences in the book, NLP technique skip though is implemented. In this method as analogous to skip gram, when the sentence is given the surrounding sentences were to be predicted.
- After the training, it is able to map any subtitle of the scene to the sentence in the book through encoder-decoder.
- Generally LSTMs were implemented, but the authors substituted it with GRUs, due to its simplicity.
- For the optimization Adam algorithm is adapted, followed by the visual-semantic embeddings.

- After achieving the first goal of matching sentences, the second goal of matching scenes with sentences was achieved by sentence-image ranking model.
- In this approach, the embedding of visual scenes is done by Descriptive Video Service(DVS). DVS is audio description for the visually impaired people.
- During the pre processing of DVS, the authors had replaced the names with the token called as *someone.*
- During training, they have used a dataset of 94 movies and 54000 clips, where each clip is represented by corresponding vector.
- The authors also focused on the semantics of the dialogues, as the sentence words may differ and corresponding BLEU score is calculated to compare the similarity.
- The matchings were done for the exact scene and the dialogues. A nearby threshold is given to match the clips with the help of Conditional Random Field(CRF).
- Inference is also possible dynamically with CRF as it is a chain. To speed up the process the authors even eliminated some states away from the uniform alignment.
- During the training, the model had watched around 1500 movies and reads 900 books per day.

5. Results:

- As finding the accurate sentence match would cost time and computations, the results had been evaluated on paragraph's scale with a threshold of 3 paragraphs around and scene with 5 subtitles away.
- To the fundamental results, the best increase had been due to the addition of CNN, which improved the results by 14%.
- Moreover, CRF, sentence embedding, video-text embedding had also improved the results.
- The highest score of the book given a movie is taken as 100% match, comparing the later books relatively. Thus, all the movies had matched to the correct books in the dataset.
- As calculating BLEU score for each subtitle-sentence match would result in computational expense, it is eliminated.
- Further the experiments on CoCoBook dataset created by Microsoft, had explained the given image with semantically significant stories.

6. Conclusion and Future works:

- Despite of creating own datasets, the authors had achieved in creating the better model than the prior for video-sentence matching.
- It is evident that the model had used the rich descriptions of the books to map the corresponding scenes as depicted by the authors.

- The future works can be concentrated on application of the model in autonomous driving and social robotics.
- Even more the applications can be extended to recognize the emotions based on the situations and voice modulations of human beings.