What You Say Is Not What You Do: Studying Visio-Linguistic Models for TV Series Summarization

Alison Reboud and Raphaël Troncy
EURECOM
Sophia Antipolis, France
alison.reboud,raphael.troncy@eurecom.fr

Abstract

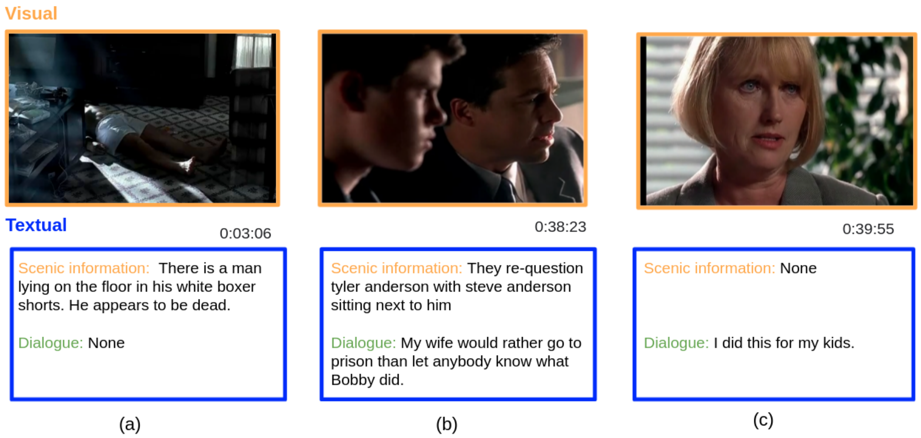
In this paper, we generate TV series summaries using both visual cues present in video frames and screenplay (dialogue and scenic textual descriptions). Recently, approaches relying on pre-trained vision and language representations have proven to be successful for several downstream tasks using paired text and images. For TV series summarization, we hypothesize that both scenic information and dialogues are useful to generate summaries. Visio-linguistic models being presented as task-agnostic, we explore if and how they can be used for TV series summarization by conducting experiments with varying text inputs and models fine-tuned on different datasets. We observe that such generic models, despite not being specifically designed for narrative understanding, achieve results closed to the state of the art. Our results suggest also that non aligned data also benefit from this type of visio-linguistics architecture.

1. Introduction

The need for automatic multimedia content understanding is exemplary for pushing the research in multimodal machine learning. In a position paper, [15] extends a previously machine-centered definition of multimodality focused on representations, to a broader definition that considers a task to be '*multimodal when inputs or outputs are represented differently* ***or are composed of distinct types of atomic units of information***'. While there has been a substantial amount of work addressing multimodal representations, it is typically not combined with the question of the unit of information. Visio-linguistics tasks are approached with general pre-trained models, said to be agnostic, but created for and tested on tasks where the information between language and vision is redundant: the text generally reflects what is going on visually, therefore, neglecting a vast amount of cases where text and images rather convey complementary aspects of meaning. For example, Figure 1-a could be used to evaluate the tasks of automatic image captioning ("A man is lying on the floor") [4] or of visual question answering (Q: "What is the man doing?", A: "Lying on the floor") [1]. In these tasks, the information present in the text is also contained in the image and the challenge consists in aligning the two modalities.

Summarizing TV series episodes, however, requires to go beyond alignment as dialogue information is not available in the visual scene, and vice versa. We hypothesize that both information are nonetheless essential to this task. In this paper, we want to summarize full-length crime TV series by producing shorter video summaries covering their most interesting parts. We use a dataset containing videos of the entire episodes of the CSI crimes TV series as well as their screenplays which are made of dialogues and scenic information. We expect interesting video segments to be characterized by the presence of elements such as remarkable dialogues and/or visual actions. Figure 1 presents three possible configurations of information spread for TV series episodes: in (a), we observe that the interesting part is contained in the scenic description, while in (b), it instead lies in the dialogue; Finally, in (c), it seems that none of the modalities is sufficient to grasp the scene content. The combination of the image description and the dialogue, however, is more interesting: the sentence ”I did this for my kids” becomes more dramatic when said in a police office. This case analysis suggests that visio-linguistics models are relevant candidates to push the frontiers of narrative summarization by adding visual information to a task that was previously only based on text [14].

When investigating the notion of complementarity for the task of TV series summarization, our work is also part of a wider reflection on multimodality and the role played by the original source of information. In an effort to assess the task-agnosticity of visio-linguistics models, we aim at shedding the light on assessing the performance of these generic models when used 'out of the box' and in particular in cases where the images and the text say different things.



Scenic information: There is a man lying on the floor in his white boxer shorts. He appears to be dead.

Dialogue: None

Scenic information: They re-question tyler anderson with steve anderson sitting next to him

Dialogue: My wife would rather go to prison than let anybody know what Bobby did.

Scenic information: None

Dialogue: I did this for my kids.

Figure 1. Examples of visual and textual segments from the CSI dataset [14]

Focusing on TV series, our experimental setup separates the dialogue and the scenic information text which are intertwined in the screenplays. We associate each text inputs to their corresponding video frames and we assess whether both types of texts benefit from the visio-linguistics models. Screenplays contain both redundant and complementary information with respect to the visual content, while enabling to easily separate them using the punctuation signs and are of equal high quality as both types are human produced. It has recently been pointed out that the pre-training choice of these models requires more attention [20]. Consequently, we also consider different pre-training strategies that make use of varying dataset size, domains and quality of image annotations. Our results show that non aligned data can benefit from pre-training too but that the pre-training dataset should be chosen carefully as it does not always help.

The remainder of the paper is structured as follows: we first present some related work (Section 2). In Section 3, we describe our approach and the design of our empirical study. In Section 4, we discuss the results before outlining some future work in Section 5.

2. Motivation and Related Work

Video summarization. Multimodal video summarization is the task of selecting representative video frames or segments using multimodal integration. Recent works have pointed out that multimodal video summarization is still approached by models developed prior to the 'deep learning era' [2, 11, 8]. Deep learning based models for video summarization [21, 10] generally focuses on the visual modality, using images or text captions [5] but neglecting the content of the speech.

Summarizing movies and TV series is often done using either visual or textual cues but not their combination. For example, [14] proposed a text based approach using latent narrative structures knowledge.

Visio-linguistics models and complementarity. As opposed to earlier works in vision and language which designed models with a task-specific architecture, many multimodal approaches use pre-trained generic visio-linguistics frameworks, which are fine-tuned on the downstream task of interest. Pre-training is typically done on image captioning datasets such as Conceptual Captions [18] or COCO Captions [6] and training rely on different self-supervised objectives, such as Image Caption Matching. There are two main types of visio-linguistics architectures: dual-stream models where the two modalities are fused at a later stage such as VilBERT [13] and single-stream models where visual and textual features are directly projected into one embedding space such as VisualBERT [12]. Our work is inspired by [16] who created a new task to push the research in complementarity modelling and who successfully used this type of visio-linguistics model. In terms of approach, our work is closest to [20] who recently created an experimental setup to question common pre-training choices for these models. Noticing that MM-IMDB [3], the out of domain task for which they found no pre-training to work better, also has unpaired data (movie synopsis and posters), we push the analysis further by making the distinction between redundant and complementary modalities.

3. Approach

3.1. TV Series Dataset

The Crime Scene Investigation (CSI) dataset [9, 14] contains 39 CSI video episodes together with their screenplays segmented into scenes, each one being associated to a binary label indicating whether the scene should be part of the summary or not. An episode contains in average 40 scenes from which 30% are labelled positively. To segment the videos into scenes, we used the word-level timestamps from the Perpetrator Identification corpus [9]. We split screenplays into dialogue and scenic information and we ultimately generate three types of text inputs: dialogue only (supplementary information), scenic information only (redundant information) and original screenplay (mixed information).

3.2. Pre-training Datasets

Following [20], we select three pre-training datasets which have different characteristics that potentially play a role in finding the most appropriate model for the task: size, domain, and quality of image annotations. **COCO captions** [6] contains 200K images from Flickr depicting everyday life situations containing common objects, each associated with five human-written captions in a fixed-style structure, yielding 1M image-caption pairs.

Conceptual Captions(CC) [18] is a collection of 3.3M image-caption pairs automatically scraped from the web using the alternate text of an image for captions. This process results in making CC a dataset with a very large diversity of visual content but suffering at times from noise in the captions. Due to broken links, the version used in this paper has 3.1M pairs.

Multimodal IMDb (MM-IMDb) [3] contains the plot (synopsis) and poster image of 26K movies. The task associated to this dataset is to classify each of these pairs according to 23 possible genres. With a total of 3113 movies in the training set, 'Thriller' (the closest genre to CSI) is the fourth most represented genre after Drama, Comedy and Romance. Although synopsis and screenplays do not share the exact same domain, they both tell the story of a movie (so do posters and the episode videos). We include this dataset to test whether sharing the movie domain could be relevant for our task. It is also a dataset where the text and the images are not aligned.

3.3. Models and Experiments

Models. Most large-scale pre-trained models have been created to handle static images. We therefore process videos as set of images (frames) and do not consider motion. We experiment with VisualBERT to account for the single-stream type of architecture and with ViLBERT for the dual-stream one. Both models treat images as region features extracted from pre-trained object detection models while text is represented as BERT global text features. In VisualBERT, these embeddings are concatenated and passed through transformer blocks (TRM). In ViLBERT, they go through two parallel transformer streams (a visual and a textual one) connected by co-attention TRM added for certain layers between the visual and textual TRM blocks. For both models, the final representation is contained in the [CLS] token and used for downstream tasks.

Experiments. We uniformly select 6 frames per video scene. We extract features for each of them and we average them afterwards. We use the MMF framework [19] for our experiments which contains, among others, the original implementation of VisualBERT [20, 12] and ViLBERT [13]. For fine-tuning via back-propagation on downstream tasks, we use binary cross entropy loss. The original CSI dataset is split in 10 folds that we re-use for our evaluation. For each fold the episodes used for training, validation and test are specified. We evaluate every 100 updates and report the model with the best loss on the validation set. We use the AdamW optimizer. The learning rate is 5e-5, a batch of size 2 and, due to computation time, we limit the training update steps to 3k (1h 16m for ViLBERT on one of the 10 folds on a NVIDIA TESLA K80 GPU). Due to class imbalance, we assigned respectively (1,3) weights to *not in* and *in* summary classes. These weights were obtained experimentally through a 10-fold cross validation with entire numbers candidates. The maximum length for textual inputs is set to 512. The default configuration as implemented in MMF is kept for the other hyper-parameters. We provide our implementation at https://github.com/alisonreboud/mmf.

4. Results analysis

		Dialogue	SI	All text
-	ViLBERT	48.36	**44.84**	48.92
COCO	ViLBERT	46.85	43.98	44.82
CC	ViLBERT	**51.19**	44.01	**50.16**
MM-IMDb	ViLBERT	47.04	44.51	48.73
-	VisualBERT	49.15	46.62	**51.07**
COCO	VisualBERT	46.80	45.91	47.71
CC	VisualBERT	**50.33**	**47.48**	49.66
MM-IMDb	VisualBERT	47.83	42.22	49.79
-	Best SUMMER	-	-	**52.00**

Table 1. Results for all text inputs and pre-training configurations in terms of F1 score (SI = Scenic Information). We also report on the state of the art performance on this dataset obtained by SUMMER [14]

Table 1 summarizes the results of our experiments using the F1 score as a metric. We also report on the performance of the SUMMER approach [14], the best performing on this dataset. The major observation we can make is that rather than a drop in performance when using complementary data (dialogue), this type of data systematically obtain better results than scenic information. More specifically, when using only dialogue text, we observe that for both ViLBERT and VisualBERT, the pre-trained CC dataset which is the most diverse and noisy dataset gives the best results and achieves near the state of the art performance without adopting a model specifically designed for narrative understanding like SUMMER does. The size of the pre-training dataset does not seem to influence the performance as MM-IMDB (the smallest) beats COCO (a dataset with a limited diversity) and no pre-training beats both. These results suggest that the diversity of the dataset is instead a decisive feature for an effective generalisation on the CSI dataset.

For dialogue, ViLBERT and VisualBERT obtain competitive results. Surprisingly enough, for scenic information, despite sharing the caption text domain with COCO and CC, both ViLBERT achieves better results without pre-training and for VisualBERT, only CC beats no pre-training. This suggests that the sensitivity to the domain of the data goes beyond the complementary vs redundant paradigm. The scenic information of this dataset is crime scene descriptions and therefore quite specific (probably more than dialogues). For scenic information, except for the non paired MM-IMDb dataset, VisualBERT outperforms ViLBERT, suggesting that the single-stream architecture is more powerful for aligned data.

For All text (the original screenplay combining both type of information), no pre-training and CC also obtain the best results. All text and dialogue achieve comparable results while scenic information systematically perform worse. Some possible explanations for the latter is that scenic information text is shorter and that TV series summarization benefits from complementary information. Finally, pre-training on MM-IMDb from the movie-domain dataset with non aligned data never achieved the best results. This could be due to the fact that despite sharing the movie domain with the downstream task, screenplays and TV episode videos are not similar to posters and IMDB plots. A major difference which could explain the results of this pre-training method is that posters and IMDB plots, while being indicative of a genre, avoid spoilers and therefore probably do not contain key-scenes type of content.

In summary, we observed that the dialogue text can benefit from visio-linguistics architectures and non-aligned pre-training while pre-training does however not systematically help. These observations are encouraging because they speak in favour of the possibility of relaxing the constraining requirement of having paired data for downstream tasks but also for pre-training datasets.

5. Conclusion

We conducted a study which isolates text elements of screenplays based on the nature of the information they convey (dialogue versus scenic information) and we tested different pre-training methods on two visio-linguistic models for the task of TV series summarization. We have shown that using a visio-linguistic architecture without paired data and without in-domain pre-training achieves near state of the art results. The fact that even with a small dataset, no pre-training beats some pre-training choices underlines the importance of in-domain and/or diverse pre-training datasets. In the future, our goal is to experiment pre-training with in-domain datasets such as movie captioning datasets [17] and video subtitles, to experiment with a very diverse pre-training dataset where the image-text alignment constraint is relaxed and to work with architectures handling videos and their temporal information [22]. In order to get more insights into the benefits of introducing images, we also plan to compare the performance of these visio-linguistic models with a text-only, general-purpose architecture such as BERT [7]. Finally, while our results suggest that the use of task-agnostic visual-linguistic models without paired data is a promising direction to look at, both for pre-training and downstream dataset, the conclusions about the possible use of complementary data need to be corroborated by more experimental results on other downstream tasks (than TV series summarization). Using visualisation techniques would also allow for a better understanding of the type of relation that the model learns between images and text, especially for complementary data.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. arXiv:2101.06072, 2021. 2

[3] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, Luis Enrique Erro No, Sta Ma Tonantzintla, and Fabio A González. Gated multimodal units for information fu. *Stat*, 1050:7, 2017. 2, 3

[4] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016. 1

[5] Yan-Ying Chen Bor-Chun Chen and Francine Chen. Video to text summary: Joint video summarization and captioning with recurrent neural networks. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 118.1–118.14. BMVA Press, September 2017. 2

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325, 2015. 2, 3

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186. Association for Computational Linguistics, 2019. 4

[8] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013. 2

[9] Lea Frermann, Shay B Cohen, and Mirella Lapata. Whodunnit? crime drama as a case for natural language understanding. *Transactions of the Association for Computational Linguistics*, 6:1–15, 2018. 2

[10] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European Conference on Computer Vision (ECCV)*, pages 505–520, 2014. 2

[11] Ijaz Ul Haq, Khan Muhammad, Tanveer Hussain, Javier Del Ser, Muhammad Sajjad, and Sung Wook Baik. Quicklook: Movie summarization using scene-based leading characters with psychological cues fusion. *Information Fusion*, 76:24–35, 2021. 2

[12] L. H. Li, M. Yatskar, D. Yin, C-J. Hsieh, and K-W. Chang. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019. 2, 3

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2019. 2, 3

[14] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay Summarization Using Latent Narrative Structure. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1920–1933, 2020. 1, 2, 3

[15] Letitia Parcalabescu, Nils Trost, and A. Frank. What is Multimodality? In *1st Workshop on Multimodal Semantic Representations (MMSR)*. ACL. 1

[16] Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 2751–2767. Association for Computational Linguistics, 2020. 2

[17] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 4

[18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2556–2565, 2018. 2, 3

[19] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. MMF: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`, 2020. 3

[20] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? Digging deeper into visio-linguistic pretraining. arXiv:2004.08744, 2020. 2, 3

[21] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *IEEE conference on computer vision and pattern recognition (CVPR*, pages 5179–5187, 2015. 2

[22] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 4