Détection de scènes remarquables dans un contexte de séries TV

Aman Berhe, Camille Guinaudeau, Claude Barras

Détection de scènes remarquables dans un contexte de séries TV

Aman Berhe* — **Camille Guinaudeau*** — **Claude Barras****

* *Université Paris-Saclay, CNRS - LISN, 91400, Orsay, France*
** *Vocapia Research, 91400, Orsay, France*

RÉSUMÉ. *Pour faciliter l'accès à une large quantité de données multimédia, il est souvent utile d'en extraire un résumé ou l'élément le plus saillant. Dans le cadre des séries télévisées, une manière d'extraire le résumé d'un épisode consiste à détecter les scènes les plus remarquables, c'est-à-dire celles qui apportent un changement radical au récit d'un épisode, avant de les combiner pour produire un résumé de l'épisode, de la saison ou de la série entière. L'aspect remarquable d'une scène ou, plus largement d'un document multimédia, est porté par ses différentes modalités – texte, parole et image – de façon conjointe ou non. Par ailleurs, une scène ne peut se définir comme remarquable qu'en comparaison des scènes qui l'entourent. Nous présentons dans cet article les premiers résultats sur la combinaison des différentes modalités et de la prise en compte du contexte pour extraire les scènes remarquables des deux premières saisons de la série Game of Thrones. Nous montrons que l'utilisation du contexte et de la prise en compte de la multimodalité permettent d'améliorer la détection de scènes remarquables.*

ABSTRACT. *To access a large amount of multimedia data, it is often useful to extract a summary or the most salient element from the collection. In TV series, one way to extract the summary of an episode is to detect the most reportable scenes, i.e. those which make a radical change to the narrative of an episode, before combining them to produce a summary of the episode, the season, or the entire series. The reportable aspect of a scene or, more broadly of a multimedia document, is carried by its different modalities - text, speech and image - in a joint way or not. In addition, a scene can only be defined as reportable in comparison to its surrounding scenes. We present in this article the first results on the combination of the different modalities and the accounting for the context to extract the most reportable scenes of the first two seasons of the Game of Thrones TV series. We show that the use of context and multimodality can improve the detection of most reportable scene.*

MOTS-CLÉS : *Multimodalité, Détection d'éléments saillant, Scènes remarquables*

KEYWORDS: *Multimodality, Salient element detection, Most reportable scenes*

1. Introduction

La structuration de larges collections de documents multimédia peut prendre plusieurs formes en fonction des documents concernés et de l'application visée : clustering en fonction du thème abordé, résumés orientés sur les personnalités présentes dans le document (Bost, 2016), etc. Dans le cadre des séries télévisées, une manière d'organiser la collection d'épisodes consiste à en extraire la structure narrative, c'est-à-dire, extraire les différentes histoires qui s'entremêlent au sein d'un épisode et identifier les scènes les plus remarquables de chacune de ces histoires (qui pourront ensuite être utilisées à des fins de résumé ou d'indexation).

Les scènes remarquables d'un épisode de série télévisée, celles qui provoquent un changement radical d'un récit ou d'une histoire en perturbant la vie des personnages et des entités impliqués, sont les plus saillantes de l'épisode. Cette saillance peut se traduire à travers différentes modalités, de façon conjointe ou non. En effet, certaines scènes seront remarquables de part le contenu textuel prononcé, sans que les modalités audio ou visuelles ne soient spécifiques et d'autres ne seront remarquables qu'à travers les modalités audio-visuelles (scènes sans parole par exemple). Cette particularité rend difficile la combinaison des différentes modalités pour la détection de scènes remarquables. Par ailleurs, la saillance d'une scène remarquable s'exprime en comparaison des scènes qui l'entourent – dans notre cas, les scènes de la même histoire – il est donc nécessaire de prendre en compte le contexte de chaque scène, c'est-à-dire, les scènes précédentes et suivantes au sein de l'histoire considérée.

Dans cet article, nous présentons de premières expériences visant à évaluer l'impact des différentes modalités et du contexte de la scène pour la détection de scènes remarquables au sein de séries télévisées. Pour cela, nous proposons une architecture distribuée dans le temps de réseaux de neurones récurrents LSTM prenant en compte des plongements de scènes fondés sur des caractéristiques audio et des caractéristiques textuelles (transcriptions manuelles et résumés des scènes).

Notre travail est présenté de la façon suivante. Après un état de l'art sur les travaux connexes dans la section 2, nous décrivons l'approche proposée dans la section 3. Les données utilisées dans ce papier sont décrites dans la section 4. Finalement, nous discutons les résultats dans la section 5, avant de conclure et proposer des perspectives pour améliorer ces premières expériences dans la dernière section.

2. État de l'art

Notre travail sur la détection de scènes remarquables se rapproche de trois types de travaux : les études fondées sur la structure narrative des documents (Zhao et Ge, 2010), (Bost, 2016), (Chu et Roy, 2017), (Guha *et al.*, 2015), celles sur la recherche d'éléments saillants dans des documents textuels (Boguraev et Kennedy, 1999), (Gillenwater *et al.*, 2012) et les travaux sur l'analyse de sentiments dans des documents audio-visuels à partir de réseaux de neurones profonds (Hershey *et al.*, 2017), (Luo *et al.*, 2019).

Le premier type de travaux se base sur des annotations manuelles ou automatiques de la structure narrative des documents considérés. Dans (Zhao et Ge, 2010), les auteurs proposent un modèle pour l'extraction de la structure narrative de films. Leur modèle analyse les films sur trois niveaux sémantiques hiérarchiques: les actes, l'intrigue et les scènes. Ils introduisent également la notion de *plot point* qui constitue le tournant de l'histoire dans une nouvelle direction et qu'ils identifient à la fin de chaque partie de la structure narrative extraite.

Le travail présenté dans (Guha *et al.*, 2015) consiste à détecter automatiquement le point culminant d'un film dans la *courbe d'intensité de l'histoire* en utilisant une structure narrative en trois actes. Pour cela, les auteurs calculent une mesure dynamique d'intensité à partir de caractéristiques bas niveau (longueur de plans, mouvement visuel, harmonicité de la musique, vitesse de prononciation de la parole). Ils suggèrent également que l'utilisation d'informations textuelles pourrait être utile. (Bost, 2016), a quant à lui travaillé sur le résumé automatique de séries télévisées. Pour cela, il identifie les scènes à intégrer dans le résumé en se fondant sur les interactions entre les personnages. Il propose par ailleurs une approche permettant de classifier les scènes de séries télévisées en 3 groupes : les scènes silencieuses, celles contenant de la musique et celles contenant du texte.

(Gillenwater *et al.*, 2012 ; Boguraev et Kennedy, 1999) ont travaillé à la recherche de contenu fondée sur la saillance dans une collection de documents. Les premiers ont utilisé une méthode probabiliste pour trouver un chemin dans un graphe à travers une collection cohérente de documents tel que ce chemin couvre les parties les plus saillantes de la collection. (Boguraev et Kennedy, 1999) ont appliqué des techniques linguistiques pour détecter les "unités phrasales" comme empreintes thématiques dans une grande collection de documents textuels. Ils ont utilisé différentes granularités pour caractériser leurs empreintes thématiques comme représentatives du déroulement complet de la narration.

Finalement, les derniers travaux proches de notre approche de détection de scènes remarquables analysent des documents audio-visuels à l'aide de réseaux de neurones profonds. Les auteurs de (Chu et Roy, 2017), par exemple, s'intéressent à la structuration de films en utilisant des informations acoustiques pour extraire les arcs émotionnels des documents étudiés. Dans (Hershey *et al.*, 2017), les auteurs utilisent des modèles neuronaux CNN performants en classification d'image et montrent leur pertinence pour la classification audio de scènes. Enfin, (Luo *et al.*, 2019) proposent un vecteur de sentiment obtenu par des modèles neuronaux CNN et LSTM à partir d'un segment audio, capable de refléter les sentiments d'un segment audio de façon plus précise que des caractéristiques acoustiques traditionnelles.

Bien que beaucoup aient essayé de mesurer l'intensité d'un récit, leur travail s'est concentré sur une vidéo ou un texte très court. Dans ce travail, nous nous intéressons à une collection complexe de séries télévisées. La plupart des méthodes se concentrent sur les caractéristiques de bas niveau d'un film et sur certaines techniques utilisées par les cinéastes. Mais nous pensons que des fonctionnalités de haut niveau extraites par

réseaux de neurones profonds et modèles d'apprentissage profond peuvent obtenir de bonnes performances en matière de détection des scènes les plus remarquables.

3. Description de l'approche

Afin de détecter les scènes remarquables des épisodes de séries télévisées, nous utilisons des modèles de réseaux de neurones profonds de type LSTM ou CNN, prenant en compte ou non le contexte des scènes considérées, appliqués à des représentations audio et textuelles de chaque scène. Les scènes, obtenues automatiquement grâce à l'outil automatique de segmentation en scène présenté dans (Berhe *et al.*, 2019), sont dans un premier temps associées à leurs données textuelles : résumés et transcriptions manuelles. Ces transcriptions et résumés sont extraits de sites de fans et sont ensuite alignées à l'audio de la scène grâce à l'outil d'alignement forcé développé par (Gauvain *et al.*, 2002). Des caractéristiques calculées à partir des contenus textuels et audio de chaque scène sont ensuite fournis aux réseaux de neurones profonds pour classer les scènes de façon binaire en "scène remarquable" et "scène non remarquable" à l'aide de l'outil Keras[1]. L'extraction des caractéristiques, la prise en compte du contexte et les modèles utilisés sont discutés dans la suite de cette section.

3.1. *Extraction des caractéristiques des scènes*

Pour caractériser les scènes de notre corpus, nous avons extrait des informations à la fois audio et textuelles. Concernant la modalité audio, l'outil Librosa[2] a été utilisé pour extraire les caractéristiques acoustiques – le spectrogramme en échelle Mel (Mel) avec un taux d'échantillonnage de 22050 hertz – et musicales – pitch et tempo – de la scène. Ces caractéristiques sont extraites pour chaque trame de 10 ms. Par ailleurs, de récents travaux ayant montré l'intérêt des caractéristiques obtenues par réseaux de neurones profonds ((Zhao et Ge, 2010) et (Luo *et al.*, 2019) par exemple), nous utilisons également des caractéristiques extraites grâce au modèle pré-entraîné VGGish fourni par Google à partir du corpus AudioSet[3]. Ce modèle, entraîné sur un grand ensemble de données YouTube (2,1 millions de vidéos, soit 5 800 heures d'audio représentés par 527 classes), extrait des plongements en 128 dimensions à partir de signaux audio d'environ 1 seconde.

Le contenu textuel de chaque scène est quant à lui représenté par des plongements de documents calculés à partir de BERT[4] (Devlin *et al.*, 2019). Chaque phrase d'une scène est d'abord représentée par un plongement calculé par BERT, puis ces plongements sont ensuite concaténés pour représenter la scène. Deux types de plongements

1. Keras (`https://keras.io/`) est une librairie Python d'apprentissage profond, exécutée sur la plate-forme d'apprentissage automatique TensorFlow (`https://www.tensorflow.org/`)
2. Librosa (`https://librosa.org/`) est une librairie Python d'analyse audio.
3. `https://research.google.com/audioset/`
4. BERT est un modèle de langue à large échelle d'architecture Transformer.

sont calculés pour chaque scène : le premier type permet de représenter les transcriptions manuelles des paroles prononcées au sein de chaque scène et le second représente leur résumé manuel, c'est-à-dire, une description détaillée des scènes incluant le nom des personnages.

Les scènes variant en longueur, en nombre de mots prononcés et en taille du résumé, les caractéristiques extraites ont également des dimensions différentes. Lors de l'apprentissage de différents modèles, l'incohérence de longueur entre les scènes est résolue en re-modelant les données à la même longueur K. Si une scène a une longueur N supérieure à K, seules les K dernières images sont considérées ($[N-K, N]$)[5]. Au contraire, si N est inférieure à K, l'intervalle manquant est rempli de zéros, ce qui peut être interprété comme un ajout de silence à la fin de la scène.

3.2. *Génération du contexte*

Les scènes étudiées sont remarquables au regard de l'histoire dans laquelle elles apparaissent, il donc est important de considérer le contexte des scènes, pour identifier les changements typiques d'intensité qu'une scène peut avoir, par rapport à ses scènes adjacentes. Afin de générer le contexte d'une scène s, les n scènes précédentes et suivantes au sein de la même histoire sont également prises en compte, par concaténation, lors du calcul des caractéristiques de s. La dimension des caractéristiques d'une scène, lorsque son contexte est pris en compte, est ainsi égale à $D(s) = (2 \times n + 1, f, K)$ où f est la dimension de la caractéristique sélectionnée et K la longueur maximale.

3.3. *Modèles utilisés*

Afin de classifier les scènes de notre corpus en "scène remarquable" et "scène non remarquables", nous avons utilisé différentes architectures de réseaux de neurones profonds. Nous avons tout d'abord employé deux modèles de base : un modèle CNN et un modèle LSTM. Deux couches convolutives bi-dimensionnelles avec max-pooling ou de LSTM reçoivent les plongements avant une couche entièrement connectée puis une couche de sortie avec Softmax qui classe en scène remarquable ou non. Afin de prendre en compte les informations séquentielles des scènes, nous avons également proposé une architecture intégrant des couches distribuées de LTSM ou de CNN. Ces couches distribuées dans le temps sont utilisées pour obtenir les plongements des scènes et de leur contexte et ainsi extraire des caractéristiques importantes pour les prochaines couches denses, entièrement connectées. Des couches de *dropout* sont également incluses pour éviter un ajustement excessif ou insuffisant.

La figure 1 représente l'architecture générale de notre approche. L'entrée du modèle correspond aux plongements, fondés sur des caractéristiques audio ou textuelles, d'une scène et de son contexte. Les caractéristiques ayant des dimensions différentes, la dimension de la couche d'entrée sera différente dans la partie droite et la par-

5. L'intensité d'une scène apparaissant généralement à la fin des scènes (Zhao et Ge, 2010), nous privilégions la fin des scènes pour la détection de scènes remarquables.

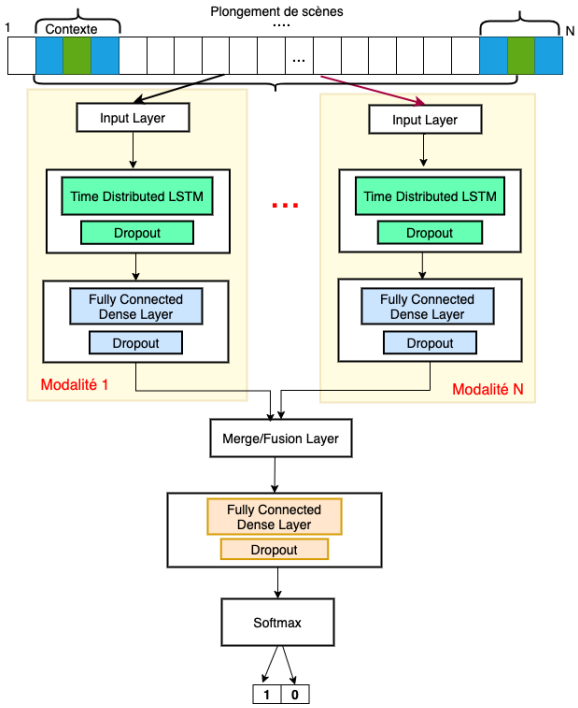

Figure 1: Architecture du modèle de détection de scènes remarquables proposé. Tous les éléments des blocs colorés sont optimisés.

tie gauche du système. Les deux blocs, couche distribuée dans le temps et couche dense, prennent en compte des caractéristiques de scènes différentes mais partagent les mêmes paramètres. Ensuite, la sortie apprise des deux blocs incorporant des scènes différentes est fusionnée sur le calque de fusion, en d'autres termes, nous effectuons une fusion tardive des plongements de scènes multimodales[6]. Enfin, la sortie de la couche de fusion est transmise à un bloc de couches denses entièrement connectées puis la fonction Softmax pour décider si une scène est remarquable (1) ou non (0).

Les blocs qui constituent les couches distribuées dans le temps et les couches denses sont optimisés à l'aide de l'algorithme TPE (Tree Pursing Estimator) avec Hyperas[7]. Les hyper-paramètres optimisés correspondent au nombre de couches, aux valeurs de *dropout*, aux fonctions d'activation et à l'optimiseur.

6. La différence de dimension entre les modalités rend difficile une fusion précoce. Une fusion plus précoce que celle proposée, avant deuxième réseau de neurones, est à l'étude mais augmente à la fois le temps de traitement et la complexité du modèle

7. Hyperas (`https://libraries.io/pypi/hyperas`) est une sur-couche de la bibliothèque Python Hyperopt afin d'optimiser des hyper-paramètres de modèles Keras.

Contexte	Exactitude	Rappel	Précision
1	0,81	0,06	0,20
3	0,85	0,44	0,53
5	0,77	0,22	0,25
7	0,80	0,17	0,30

Tableau 1: Impact du contexte sur la performance d'un modèle LSTM utilisant les paramètres VGGish. Le contexte est le nombre de scènes adjacentes, à gauche et à droite, à considérer.

4. Données

Dans ce travail, nous avons utilisé les deux premières saisons de la série télévisée "Game of Thrones", composée de 20 épisodes, divisés en 444 scènes[8]. Chaque scène est ensuite annotée en scène remarquable (1) ou non remarquable (0). L'ensemble de données est déséquilibré en termes de rapport entre les scènes remarquables et les scènes non remarquables. Sur les 444 scènes, seules 72 scènes sont étiquetées comme remarquables. Les données sont également déséquilibrées en termes de longueur de scène, la scène la plus courte ayant une durée de 1,4 seconde et la plus longue une durée de 472,8 secondes (avec une moyenne de 133,3 secondes).

Les transcriptions manuelles et les résumés des scènes sont extraits à partir du site https://gameofthrones.fandom.com/wiki/Game_of_Thrones_Wiki. La scène la plus longue est associée à une transcription de 859 mots, la plus courte à une transcription d'un seul mot, avec une moyenne de 209,4 mots. Les résumés sont composés au maximum de 763 mots, au minimum 8 mot et en moyenne de 188,5 mots. Il y a une scène sans parole qui n'est pas une scène remarquable. Finalement, les données sont divisées en données d'entraînement, de test et de validation, dans une proportion de 50%, 25%, 25% respectivement. La proportion des scènes remarquables est équivalente dans les 3 parties

5. Résultats et discussions

Les premiers résultats obtenus pour la détection de scènes remarquables sont présentés et discutés dans cette section. Les tableaux présentent les valeurs de précision et de rappel, calculés uniquement sur l'ensemble des scènes remarquables ainsi que le taux d'exactitude calculée sur l'ensemble des scènes[9]. Des expériences préliminaires nous ont montré que le modèle LSTM fournit de meilleurs résultats que le modèle CNN, lié sans doute à la mémoire intrinsèque au modèle. Le tableau 1 présente l'impact du contexte pour améliorer la qualité de la détection avec un modèle LSTM.

8. Ces scènes correspondent à une partie de l'ensemble de données annotées, pour la segmentation de scène, par (Bost, 2016)

9. Le taux d'exactitude lorsque le modèle ne prédit que des scènes non remarquables est égal à 0,83 sur les données de test.

Caractéristiques	Exactitude	Rappel	Précision
VGGish	0,77	0,22	0,25
VGGish+Pitch	0,0	0,0	0,0
VGGish+Tempo	0,79	0,22	0,31
VGGish+Résumé	0,74	0,17	0,18
VGGish+Transcription	0,77	0,17	0,23
VGGish+Résumé+Tempo	0,82	0,22	0,40
VGGish+Transcription+Tempo	0,82	0,11	0,33
VGGish+Résumé+Pitch	0,77	0,22	0,27
VGGish+Transcription+Pitch	0,77	0,22	0,27

Tableau 2: Performances de la fusion multimodale pour un modèle LSTM distribué sur un contexte de 5 scènes.

Nous nous sommes également intéressés à la combinaison des différentes modalités pour améliorer la qualité de la détection des scènes remarquables. Le tableau 2 montre les résultats de la fusion multimodale avec des modèles optimisés. Les paramètres textuels et musicaux (incrustations textuelles de scènes et pitchs) seuls ne fonctionnaient pas bien et ne prédisaient que des scènes non remarquables. La fusion tardive des modalités contribue à augmenter les performances des modèles et fournit les meilleurs résultats.

6. Conclusion

Détecter les scènes les plus remarquables est indispensable à la décomposition des épisodes des séries télévisées et à l'extraction de leur structure narrative. Dans cet article, nous avons étudié les caractéristiques les plus importantes pour identifier ces scènes remarquables.

Comme discuté dans la section 5, les modèles LSTM temporels basés sur le contexte fonctionnent mieux que les modèles CNN homologues. La fusion des caractéristiques multimodales aide également à améliorer la précision de détection de scènes remarquables bien que le rappel soit un peu inférieur à la précision. La fusion des paramètres audio VGGish avec des modalités textuelles, en particulier avec un résumé des scènes, fonctionne mieux que la fusion avec des paramètres de musique et de hauteur. La prise en compte du contexte d'une scène a également un grand impact sur les résultats, en capturant la différence d'intensité qu'une scène peut avoir par rapport à ses scènes adjacentes, bien qu'il soit difficile de conclure sur la taille optimale du contexte.

Nous pensons que les résultats peuvent être améliorés et que des modèles plus robustes peuvent être obtenus en utilisant l'augmentation des données, même si la prise en compte du contexte des scènes adjacentes complique la génération de données. Des paramètres visuels exprimant l'action et le mouvement pourraient également être utiles pour identifier les scènes qui manifestent visuellement un niveau élevé d'intérêt.

7. Bibliographie

Berhe A., Barras C., Guinaudeau C., "Video Scene Segmentation of TV Series Using Multimodal Neural Features", *Series-International Journal of TV Serial Narratives*, vol. 5, n° 1, p. 59-68, 2019.

Boguraev B., Kennedy C., "Salience-based Content Characterisation of Text Documents", *Advances in Automatic Text Summarization*, p. 99-110, 1999.

Bost X., A Storytelling Machine?: Automatic Video Summarization: the Case of TV Series, PhD thesis, Université d'Avignon, 2016.

Chu E., Roy D., "Audio-visual Sentiment Analysis for Learning Emotional Arcs in Movies", *2017 IEEE International Conference on Data Mining (ICDM)*, IEEE, p. 829-834, 2017.

Devlin J., Chang M.-W., Lee K., Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) on Human Language Technologies*, p. 4171-4186, 2019.

Gauvain J.-L., Lamel L., Adda G., "The LIMSI Broadcast News Transcription System", *Speech Communication*, vol. 37, n° 1-2, p. 89-108, 2002.

Gillenwater J., Kulesza A., Taskar B., "Discovering Diverse and Salient Threads in Document Collections", *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, p. 710-720, 2012.

Guha T., Kumar N., Narayanan S. S., Smith S. L., "Computationally Deconstructing Movie Narratives: an Informatics Approach", *2015 IEEE International Conference on Acoustics, Speech and Sprocessing (ICASSP)*, IEEE, p. 2264-2268, 2015.

Hershey S., Chaudhuri S., Ellis D. P., Gemmeke J. F., Jansen A., Moore R. C., Plakal M., Platt D., Saurous R. A., Seybold B. *et al.*, "CNN Architectures for Large-scale Audio Classification", *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 131-135, 2017.

Luo Z., Xu H., Chen F., "Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-based Parallel Neural Network", *Proceedings of the 2nd Workshop on Affective Content Analysis co-located with 33rd AAAI Conference on Artificial Intelligence, (AffCon@ AAAI)*, 2019.

Zhao Z., Ge X., "A Computable Structure Model for Hollywood Film", *2010 IEEE International Conference on Image Processing (ICIP)*, IEEE, p. 877-880, 2010.