Benchmarking Deep Learning based Automatic Scenes Description with User Study

Siddharth Sachan

Supervisors
Prof. Tom Gedeon
Dr. Sabrina Caldwell

The Australian National University

November 19, 2021

Introduction

Audio Descriptions (AD)

- AD explains the key audiovisual events in the video, making them more accessible for visually impaired people.
- Describe *visual details*, like *actions, characters, scene changes and on-screen text*.
- Usually added during *existing pauses in dialogue*.
- Positive Impacts:
- Amazon Prime Videos have AD for 3000/26000, Netflix has 1700/6000[1].
- Cost: 15-50 USD/minute[1].
- Issues with user uploaded data: 500 hours/minute on YouTube[2].

[1] https://adp.acb.org/

[2] https://expandedramblings.com/index.php/youtube-statistics/

Introduction

Audio Descriptions (AD)

- AD explains the key audiovisual events in the video, making them more accessible for visually impaired people.
- Describe *visual details*, like *actions, characters, scene changes and on-screen text*.
- Usually added during *existing pauses in dialogue*.
- Positive Impacts:
- Amazon Prime Videos have AD for 3000/26000, Netflix has 1700/6000[1].
- Cost: 15-50 USD/minute[1].
- Issues with user uploaded data: 500 hours/minute on YouTube[2].

[1] https://adp.acb.org/

[2] https://expandedramblings.com/index.php/youtube-statistics/

Background

Automatic AD

- A sequence to sequence translation with more than one ground truth
- Characterization in terms of What, Where and How
- Metadata based descriptions: Japan Olympics Broadcast[2]
- Descriptions extracted from script: Non-dialogue lines[3]
- Use of Deep Learning: Dense Video captioning dataset and model as backbone
- 3 sub-module each for What, Where and How[4]
- User Study: Helpful, but confusing and repetitive.

Background

Automatic AD

- A sequence to sequence translation with more than one ground truth
- Characterization in terms of What, Where and How
- Metadata based descriptions: Japan Olympics Broadcast[2]
- Descriptions extracted from script: Non-dialogue lines[3]
- Use of Deep Learning: Dense Video captioning dataset and model as backbone
- 3 sub-module each for What, Where and How[4]
- User Study: Helpful, but confusing and repetitive.

- A sequence to sequence translation with more than one ground truth
- Characterization in terms of What, Where and How
- Metadata based descriptions: Japan Olympics Broadcast[2]
- Descriptions extracted from script: Non-dialogue lines[3]
- Use of Deep Learning: Dense Video captioning dataset and model as backbone
- 3 sub-module each for What, Where and How[4]
- User Study: Helpful, but confusing and repetitive.

- Given a video predict the time stamps and respective captions
- Extension of Image captioning
- Popular datasets: You-CookII[5], ActivityNet Captions[6]
- Sequence to sequence structure: Encoder-Decoder models
- Examples: CNN-RNN, RNN-RNN, RNN-Hierarchical RNN, Transformers
- SOTA method: PDVC[7] uses deformable transformers on pre-extracted frame-wise video features (Two-stream Network[8] trained on action recognition).

Methodology

Pipeline

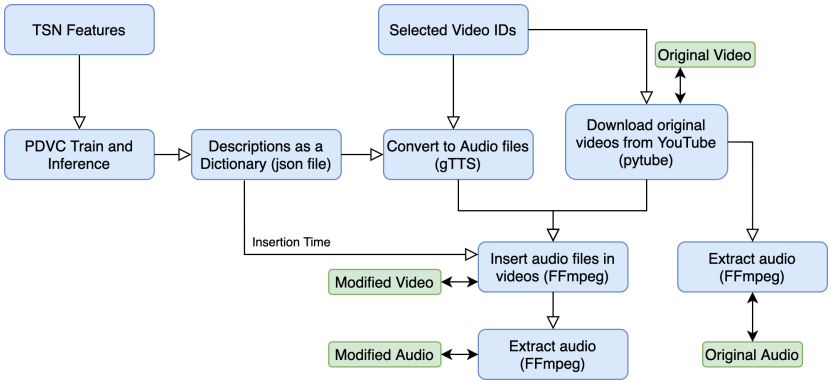

Figure 1: Pipeline followed to generate and insert AD in the videos.

Audio

1. **BEFORE:** Based on what you heard, how **confident** are you that you can describe the video content?
2. Did you think the descriptions are **confusing**?
3. Do the descriptions **match the information** given by existing audio?
4. How much do you agree with the following statement? The descriptions are **redundant or have grammar errors**.
5. How much do you agree with the following statement? The descriptions **help you picture the video content**.
6. What other information should the description provide?

Video

1. **BEFORE:** For what reason did you think the video currently needs to be added with audio descriptions?
2. Did you think the descriptions are **confusing**?
3. Do the descriptions **match the information present in the video?**
4. How much do you agree with the following statement? The descriptions are **redundant or have grammar errors**.
5. How much do you agree with the following statement? The descriptions help you **understand the video content**.
6. What other information should the description provide?

Methodology

User Study

Audio

1. **BEFORE:** Based on what you heard, how **confident** are you that you can describe the video content?
2. Did you think the descriptions are **confusing**?
3. Do the descriptions **match the information** given by existing audio?
4. How much do you agree with the following statement? The descriptions are **redundant or have grammar errors**.
5. How much do you agree with the following statement? The descriptions **help you picture the video content**.
6. What other information should the description provide?

Video

1. **BEFORE:** For what reason did you think the video currently needs to be added with audio descriptions?
2. Did you think the descriptions are **confusing**?
3. Do the descriptions **match the information** present in the video?
4. How much do you agree with the following statement? The descriptions are **redundant or have grammar errors**.
5. How much do you agree with the following statement? The descriptions help you **understand the video content**.
6. What other information should the description provide?

Results

Need of AD

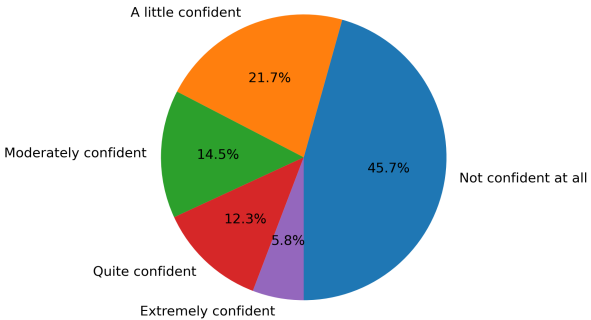

Figure 2: Percentage of responses about confidence in video content after listening to the original audio.

Results

Impact of AD

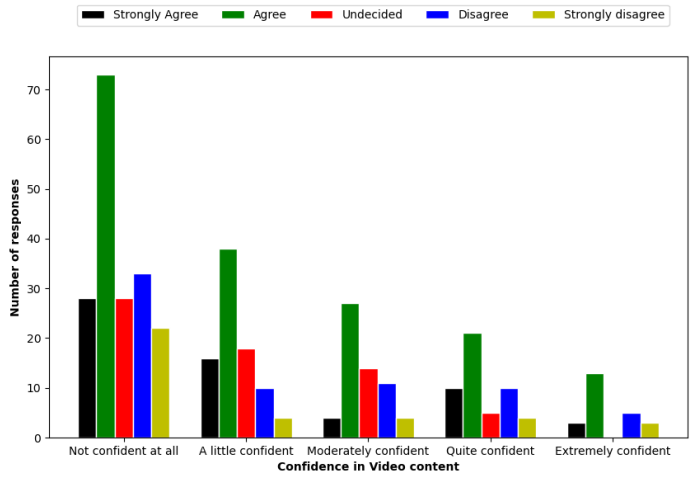

Figure 3: Relation between the people's confidence after listening to the original audio and finding modified audio helpful in picturing the video content.

Results

Confusing Content?

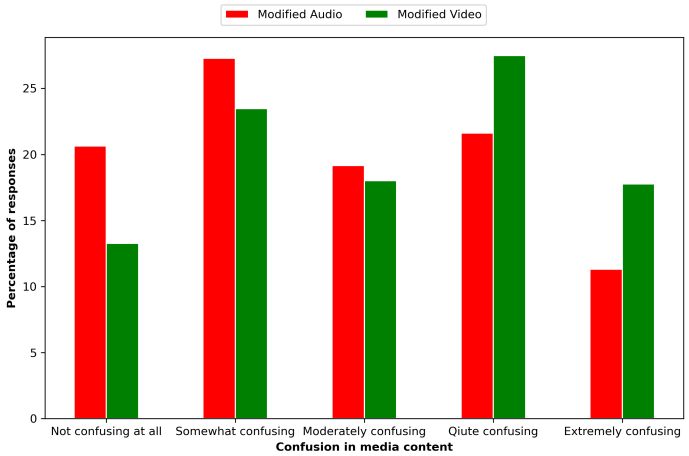

Figure 4:
Percentage of people who found the descriptions confusing.

Results

Information mismatch?

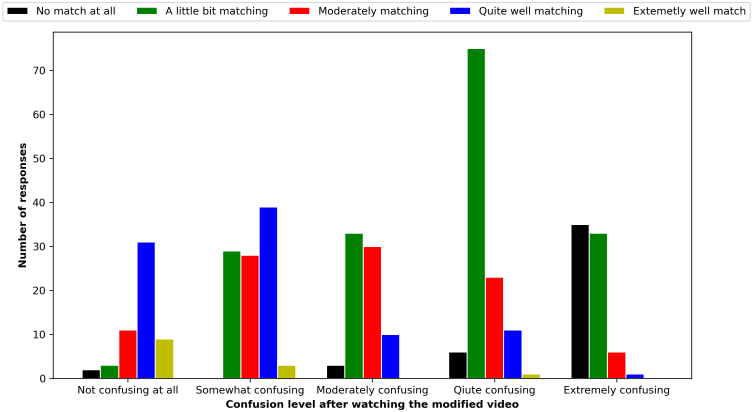

Figure 5: Relation between the confusion level of the video and information mismatch between the description and the visual content.

Results

Grammatical errors and redundancy

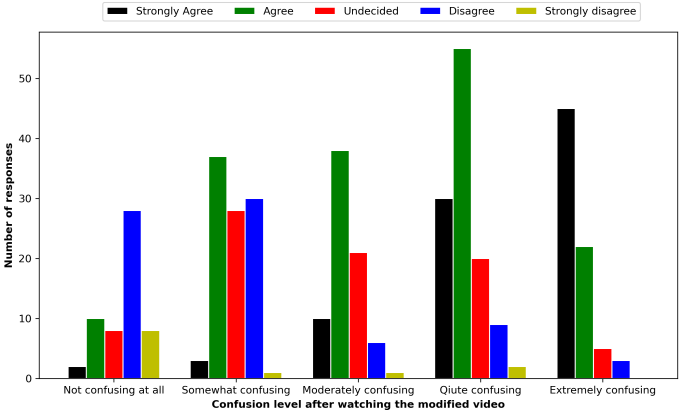

Figure 6: Relation between the confusion level of the video and grammatical errors and redundancy of descriptions.

Results

What is lacking/wrong in the descriptions?

- Little background details: Where?
- Gender misidentification in the pronouns used
- No impact of the text appearing on the screen
- Background noise/Loud music: Need varying level of loudness in AD
- Inadequacy when multiple people are present: Re-identification?

Conclusion and Future Opportunities

- Helpful in visualising the video content from descriptions.
- Information mismatch and grammatical errors make them confusing.
- Descriptions lack background details and can misidentify people.
- Fusing with other models like text extraction.
- Identifying overlay music: change the loudness of AD.
- Person re-identification for person centric descriptions: New data?

Conclusion and Future Opportunities

- Helpful in visualising the video content from descriptions.
- Information mismatch and grammatical errors make them confusing.
- Descriptions lack background details and can misidentify people.
- Fusing with other models like text extraction.
- Identifying overlay music: change the loudness of AD.
- Person re-identification for person centric descriptions: New data?

References

M. Plaza, "Cost-effectiveness of audio description production process: comparative analysis of outsourcing and 'in-house' methods," *International Journal of Production Research*, vol. 55, no. 12, pp. 3480–3496, 2017.

K. Kurihara, A. Imai, N. Seiyama, T. Shimizu, S. Sato, I. Yamada, T. Kumano, R. Tako, T. Miyazaki, M. Ichiki, T. Takagi, and H. Sumiyoshi, "Automatic generation of audio descriptions for sports programs," *SMPTE Motion Imaging Journal*, vol. 128, no. 1, pp. 41–47, 2019.

V. P. Campos, T. M. de Araújo, G. L. de Souza Filho, and L. M. Gonçalves, "Cinead: a system for automated audio description script generation for the visually impaired," *Universal Access in the Information Society*, vol. 19, no. 1, pp. 99–111, 2020.

Y. Wang, W. Liang, H. Huang, Y. Zhang, D. Li, and L.-F. Yu, "Toward automatic audio description generation for accessible videos," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2021.

L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-end dense video captioning with parallel decoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6847–6857, 2021.

L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks for action recognition in videos," *CoRR*, vol. abs/1705.02953, 2017.

Thank you for listening!

Questions?

Siddharth Sachan

siddharth.sachan@anu.edu.au
https://github.com/sidsachan