Multimodal Data Representations for Social Media Analysis

Supervisors: Ernest Valveny (UAB). Dimos Karatzas (UAB)

e-mail: ernest@cvc.uab.es, dimos@cvc.uab.es

Description

Deep Learning for Multimodal data representation is a fundamental technique to integrate data of different types into a common space in which both intermodality and cross-modality can be represented. These representations provide richer information than a single modality in data that is multimodal by nature -- such as social media content that usually includes images and text [1, 7] -- by taking advantage of the specific information of each modality. Cross-modal attention [2, 3], in which one modality is attended conditioned on the other is a popular choice for multi-modal feature fusion in Vision and Language tasks. Other multi-modal fusion techniques have also been explored such as bilinear models [4] or graph CNNs [5]. The current trend in more recent works is the fusion through multi-modal transformers such as the VilBERT (short for Vision-and-Language BERT) [6] model.

With this project we intend to bring this multimodal treatment of information to the analysis of social media, and more specifically, to the tasks of hate speech and fake news detection. The main goal of the project will be focused on the task of hate speech detection, by designing and training deep neural network models to extract rich representations from multimodal social media posts, based on previous works [1] combining both textual and visual information.

Tweet text: "What separates humans from animals"

Example of hate-speech tweet post where multimodal information is relevant

Two different kind of approaches can be explored. On one hand, a cross-modal approach, where a common semantic space is learnt where either modality can be projected separately and rich sematic representations are learnt. On the other hand, an intermodal approach, using deep network architectures that seamlessly combine textual and visual information into a single joint representation. Once trained these models will be tested and evaluated on existing public datasets [1, 7].

Additionally, these multimodal models can be fine-tuned to the task of fake news analysis by providing tools for multimodal retrieval that could be used to filter potential fake news candidates based on their similarity to previously labelled publications.

The project can be developed within the context of a paid internship as practices at the CVC. There is also the possibility to continue as a research assistant / PhD student on this topic once the TFM project finishes

References

[1] Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1470- 1478).

[2] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 21–29 (2016)

[3] Kazemi, V., Elqursh, A.: Show, ask, attend, and answer: A strong baseline for visual question answering. arXivpreprint arXiv:1704.03162 (2017)

[4] Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp. 2612–2620 (2017)

[5] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks.arXiv preprintarXiv:1609.02907 (2016)

[6] Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.arXiv preprintarXiv:1908.02265 (2019)

[7] Kiela, D, Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, NIPS 2020