Vision and Language Integration Meets Multimedia Fusion

Marie-Francine Moens
KU Leuven

Katerina Pastra
Cognitive Systems Research Institute and Institute for Language and Speech Processing

Kate Saenko
Boston University

Tinne Tuytelaars
KU Leuven

Multimodal information fusion at both the signal and semantics level is a core part of most multimedia applications, including indexing, retrieval, and summarization. Prototype systems have implemented early or late fusion of modality-specific processing results through various methodologies including rule-based approaches, information-theoretic models, and machine learning.[1] Vision and language are two of the predominant modalities that are fused, with a long history of results in TRECVid, ImageClef, and other international challenges. During the last decade, vision–language semantic integration has attracted attention from traditionally non-interdisciplinary research communities such as computer vision and natural language processing. This is due to the fact that one modality can greatly assist in processing another by providing cues for disambiguation, complementary information, and noise/error filtering. Recent advances in deep learning have opened up new opportunities in joint modeling of visual and co-occurring verbal information in multimedia.

VISION AND LANGUAGE INTEGRATION

Emerging research on the joint processing of visual and language data is stimulated by the active development of deep learning algorithms. For instance, deep neural networks (DNNs) offer numerous ways to learn mappings between visual and language media and multimodal representations of content.

Deep learning has become a standard approach for tasks like automated image and video captioning[2–6] and visual question answering (VQA).[7–9] Captioning refers to the automated description of images or video with natural language sentences. In VQA, the goal is to automatically formulate an answer to a question posed about an image, where both the question and answer are in natural language.

Vision and language integration also facilitates the processing of language grounded in perception and/or actions in the world.[10] Grounded language processing contributes to automated language understanding and machine translation of language. Recently, it has been shown that visual data provide world and common-sense knowledge that is needed in automated language understanding.[11]

Published by the IEEE Computer Society
1070-986X/18/$33.00 ©2018 IEEE

In addition, jointly processing visual and language data makes it possible for researchers to explore theories on the complementarity of language and visual data[12] that could lead to better understanding of human and human–machine communication and improve statistical learning of knowledge representations informed by visual and language data as well as inferences about these representations.

MULTIMEDIA FUSION

From its early days, multimedia processing has been concerned with the fusion of information from different modalities. Fusion models are often classified according to the level of fusion.[13] Early-fusion models accomplished fusion at the level of input features, for instance by concatenating the vector representation of text with the vector representation of accompanying images. Late-fusion models act at the decision level—for instance, classification decisions are computed per modality and these are combined to obtain a best or approximately best result, which is often accomplished via an inference layer on top of the output layers of each modality model. The most popular models impose a linear interpolation of the outputs for each modality, with interpolation weights learned from the training data. Deep learning blurs the distinctions between early- and late-fusion models. In principle, features can be combined at each level in a DNN, leading to many hybrid forms of fusion.[14]

IN THIS ISSUE

The three articles in this special issue address several key aspects of vision and language integration. These include identifying the right fusion architecture and applications such as video hyperlinking and recommendation, multimodal classification, and image and video captioning, in which vision and language provide complementary information to train the systems.

In "A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking," Vedran Vukotić, Christian Raymond, and Guillaume Gravier show the importance of continuous representation spaces that capture visual and language content and learning such spaces with deep learning methods. The authors propose a new method to perform high-level multimodal fusion using crossmodal translation. This goal is realized by symmetrical encoders that are cast into a bi-directional DNN (BiDNN). The article especially contributes to representation learning in multimodal spaces. The presented models are evaluated on the tasks of video hyperlinking and video segment recommendation. Results obtained within the 2016 TRECVID video hyperlinking benchmarking initiative show that the proposed method obtains the best score for this task, thus defining the state of the art.

In "Integrating Vision and Language for First-Impression Personality Analysis," Jelena Gorbova, Egils Avots, Iiris Lüsi, Mark Fishel, Sergio Escalera, and Gholamreza Anbarjafari present a novel method of vision and speech integration in the context of personality analysis. Co-occurring image and audio frames are sampled to form a time-series, which is processed using a layer of long short-term memory cells to recognize the personality of a job candidate in a brief video conversation. The model is evaluated on a database of 10,000 clips extracted from YouTube videos created for the ChaLearn Looking at People Job Candidate Screening Competition. The article demonstrates the value of fusing video, audio, image frames, and transcribed speech using a deep learning approach.

Finally, in "Image and Video Captioning with Augmented Neural Architectures," Rakshith Shetty, Hamed R. Tavakoli and Jorma Laaksonen show that convolutional neural network (CNN)-based image and video captioning can be substantially improved with architectures that leverage special features from the scene context, objects, and locations. The language model for caption generation is first initialized with contextual image features or dense trajectory video features, and then has access to the CNN features during the entire caption-generation process. A discriminatively trained evaluator network for choosing the best caption among those generated by an ensemble of caption generator networks further improves accuracy. The proposed framework outperforms the state of the art on the MS-COCO image-captioning leaderboard and tops the MSR-VTT challenge leaderboard.

CONCLUSION

The emerging field of visual and language integration offers many research opportunities. Given the success of current approaches based on deep learning methods, we foresee great success in automated understanding of multimodal data. Deep learning models enable content to be represented in a continuous format that is shared across the modalities, which raises many questions. For instance, we know very little about how to better leverage the learned representations that integrate visual and language content in multimedia tasks for which we have few annotated training data, how to incrementally learn with one or a few examples, and how to reason with the learned representations to detect additional information, which is not made explicit in multimedia but could be inferred.

ACKNOWLEDGMENT

We thank all the reviewers who contributed to the selection of articles for this special issue.

REFERENCES

1. K. Pastra and Y. Wilks, "Vision-Language Integration in AI: A Reality Check," *Proc. 16th European Conf. Artificial Intelligence* (ECAI 04), 2004, pp. 937–941.
2. J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, 2017, pp. 677–691.
3. S. Venugopalan et al., "Sequence to Sequence—Video to Text," *Proc. 2015 IEEE Int'l Conf. Computer Vision* (ICCV 15), 2015, pp. 4534–4542.
4. S. Venugopalan et al., "Translating Videos to Natural Language Using Deep Recurrent Neural Networks," *Proc. 2015 Ann. Conf. North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies*, 2015, pp. 1494–1504.
5. O. Vinyals et al., "Show and Tell: A Neural Image Caption Generator," *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR 15), 2015, pp. 3156–3164.
6. K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *Proc. 32nd Int'l Conf. Machine Learning* (ICML 15), 2015, pp. 2048–2057.
7. M. Malinowski, M. Rohrbach, and M. Fritz, "Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images," *Proc. 2015 IEEE Int'l Conf. Computer Vision* (ICCV 15), 2015, pp. 1–9.
8. S. Antol et al., "VQA: Visual Question Answering," *Proc. 2015 IEEE Int'l Conf. Computer Vision* (ICCV 15), 2015, pp. 2425–2433.
9. H. Xu and K. Saenko, "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering," *Proc. 2016 European Conf. Computer Vision* (ECCV 16), 2016, pp. 451–466.
10. K. Pastra et al., "Embodied Language Processing: A New Generation of Language Technology," *Proc. 14th AAAI Conf. Language-Action Tools for Cognitive Artificial Agents*, 2011, pp. 23–29.
11. G. Collell, L. Van Gool, and M.-F. Moens, "Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates," *Proc. 32nd AAAI Conf. Artificial Intelligence* (AAAI-18), 2018, pp. 6765–6772.
12. K. Pastra, "COSMOROE: A Cross-Media Relations Framework for Modeling Multimedia Dialectics," *Multimedia Systems*, vol. 14, no. 5, 2008, pp. 299–323.
13. P.K. Atrey et al., "Multimodal Fusion for Multimedia Analysis: A Survey," *Multimedia Systems*, vol. 16, no. 6, 2010, pp. 345–379.
14. G. Farnadi et al., "User Profiling through Deep Multimodal Fusion," *Proc. 11th ACM Int'l Conf. Web Search and Data Mining* (WSDM 18), 2018, pp. 171–179.

ABOUT THE GUEST EDITORS

Marie-Francine Moens is a professor in the Department of Computer Science at KU Leuven, where she is director of the Language Intelligence and Information Retrieval (LIIR) research lab, a member of the Human Computer Interaction group, and head of the Informatics section. Her main research lies in the development of novel methods for automated content recognition in text and multimedia using statistical machine learning, and exploiting insights from linguistic and cognitive theories. Moens received a PhD in computer science from KU Leuven. She is a member of IEEE and editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Contact her at sien.moens@cs.kuleuven.be.

Katerina Pastra is director of the Cognitive Systems Research Institute and a senior researcher at the Institute for Language and Speech Processing, ATHENA Research Center. Her research focuses on the computational integration of language, perception, and action based on experimental findings from neuroscience. Pastra received a PhD in artificial intelligence from the University of Sheffield. She is a Senior Member of IEEE, the vice-chair of the European Network on Vision and Language Integration, and co-editor of *Language and Cognition Research Topic in Frontiers in Robotics and Artificial Intelligence*. Contact her at kpastra@csri.gr.

Kate Saenko is an associate professor of Computer Science at Boston University, where she is the director of the Computer Vision and Learning Group and co-director of the AI Research initiative. Her research interests are in the broad area of artificial intelligence with a focus on adaptive machine learning, learning for vision and language understanding, and deep learning. Saenko received a PhD degree in electrical engineering and computer science from MIT. She is a member of IEEE. Contact her at saenko@cs.uml.edu.

Tinne Tuytelaars is a research professor in the Department of Electrical Engineering at KU Leuven. Her research interests include robust image and video representations, object and action recognition, multimodal analysis, and image and video understanding. Tuytelaars received a PhD in computer science from KU Leuven. She is a member of IEEE. Contact her at tinne.tuytelaars@esat.kuleuven.be.