One Perceptron to Rule Them All: Language, Vision, Audio and Speech

Xavier Giro-i-Nieto
Universitat Politecnica de Catalunya
Barcelona, Catalonia
xavier.giro@upc.edu

ABSTRACT

Deep neural networks have boosted the convergence of multimedia data analytics in a unified framework shared by practitioners in natural language, vision and speech. Image captioning, lip reading or video sonorization are some of the first applications of a new and exciting field of research exploiting the generalization properties of deep neural representation. This tutorial will firstly review the basic neural architectures to encode and decode vision, text and audio, to later review the those models that have successfully translated information across modalities.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; *Natural language processing*; *Computer vision*; Multi-task learning; • **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

deep learning; multimodal; cross-modal; joint embeddings

ACM Reference Format:
Xavier Giro-i-Nieto. 2020. One Perceptron to Rule Them All: Language, Vision, Audio and Speech. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3372278.3390740

1 MOTIVATION

Research in multimedia has experienced major changes during the recent years thanks to the advances in applied deep learning. Deep neural networks have achieved outstanding performance in the field of feature learning by optimizing the parameters of their millions of basic units: perceptrons. While machine learning had already been broadly used before the adoption of deep neural networks, the general adoption of such machinery has also facilitated the interaction between multimedia researchers with diverse backgrounds. From one side, novel neural layers or optimization schemes proposed initially for a certain modality, are often ported to other modalities, boosting the exchange of ideas and interactions among the community.On the other side, the adoption of common neural representations and development frameworks has also facilitated the development of cross-modal applications at a very fast pace. The encoding and decoding of pixels, phonemes or characters with the same tools allows combining them in multiple and imaginative ways.

This tutorial presents the most common neural layers for multimedia encoding and decoding, and provides a review of how they have been combined to build cross-modal applications. It targets a technical audience who is already familiar with the learning mechanisms of deep neural networks, but whose expertise may be currently focused in a specific modality. At the end of the tutorial, attendees will have a broader view of the opportunities that deep learning offers to the multimedia community, by facilitating the interaction between both data and scientists.

2 COURSE DESCRIPTION

2.1 Multimedia Encoding and Decoding

The broad interest in deep learning is related in many cases by the unprecedented success of the AlexNet [17] in the task of image classification. The work showed how convolutional neural networks (CNN) [18] could be trained with backpropagation [24] to provide an end-to-end solution between raw pixels and one-hot encoding of the classes. On the other hand, another part of the multimedia community discovered the potential of deep learning when applied to machine translation. In its most basic set up, text encoded with recurrent neural networks (RNNs) [16] could be decoded into another language [9, 29], or even be used to synthesize speech [20, 21]. Attention mechanisms have been added on top of of both visual [32], textual [8] or spoken [7] representations learned by CNN and/or RNN, but also directly over the data [30].

2.2 Cross-Modal Architectures

The broad adoption of neural representations for both encoding and decoding multimedia data has boosted the research in cross-modal applications that translate data from one modality into another. In its most basic set up, a cross-modal architecture takes data from a source modality that must be converted into another modality. The task is addressed by encoding the source data into an intermediate representation, which is later decoded into the target modality.Image captioning [31] is one of the most representatives examples of such approach, in which pixels are encoded with a CNN and the words in the captions are decoded with an RNN, similarly how a basic neural machine translation pipeline works. Other well known applications from one modality into another are automatic speech recognition [14], speech synthesis [21], lip reading [27], image synthesis [23], speech reconstruction [12] or face hallucination [11].

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06.
https://doi.org/10.1145/3372278.3390740

More complex pipelines would consider multiple inputs or outputs. In the case of multiple inputs, a separate encoder learns single-modal features to be fused at a deeper layer of the network, before being decoded into the output modality. This would be the case of visual question answering [3], visual speech separation[2], speech recognition enhanced with video [1, 22], or visual re-dubbing [10]. In the case of multiple outputs, the multi-task learning paradigm is adopted, but normally the interest is in the primary task, while the secondary task is added to help into the training of the model. This would be the case of image captioning with visual grounding [19], or sign language translation predicting both the natural language and sign glosses transcriptions [6].

2.3 Joint Feature Learning

Features learned with deep neural networks are not always used as a proxy from one modality into another, but also as final and rich representations by themselves. Neural encoders for different modalities can be trained with pairs of data samples to learn joint multimodal embeddings. The first works combined language models with image labels to learn a feature space capable that may be exploited for zero-shot learning [13, 26]. This learning paradigm has been broadly exploited to learn features for multimodal retrieval, allowing search images to/from text [5, 25] or videos to/from their audio track [28]. Similarly, the alignment between the audio and visual tracks in video files has facilitated multiple self-supervised learning approaches that could tackle sound source localization [4] or the discovery of spoken words from pixels [15].

3 INSTRUCTOR BIOGRAPHY

Xavier Giro-i-Nieto is an associate professor at the Universitat Politecnica de Catalunya (UPC) in Barcelona, as member of the Intelligent Data Science and Artificial Intelligence Research Center (IDEAI-UPC) and Image Processing Group (GPI), and also a visiting researcher at Barcelona Supercomputing Center (BSC). He completed his undergraduate studies from UPC in 2000, after a master thesis at the Vrije Universiteit in Brussels (VUB) with Prof. Peter Schelkens. In 2012, he obtained his Phd on image retrieval from UPC under the supervision by Prof. Ferran Marqués (UPC) and Prof. Shih-Fu Chang (Columbia University). He serves as associate editor at IEEE Transactions in Multimedia and reviews for top tier conferences in machine learning, computer vision and multimedia.

ACKNOWLEDGMENTS

This tutorial was supported by the the Industrial Doctorate 2017-DI-064 from the Government of Catalonia, and the Spanish Ministry of Economy and Competitiveness through the European Regional Development Fund (TEC2016-75976-R).

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2018).
[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. The Conversation: Deep Audio-Visual Speech Enhancement. *Interspeech* (2018).
[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision.* 2425–2433.
[4] Relja Arandjelovic and Andrew Zisserman. 2018. Objects that sound. In *ECCV.*
[5] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2017. Cross-modal scene networks. *TPAMI* 40, 10 (2017), 2303–2314.
[6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. *CVPR* (2020).
[7] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*. IEEE.
[8] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *TMM* 17, 11 (2015).
[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP* (2014).
[10] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that? *arXiv preprint arXiv:1705.02966* (2017).
[11] Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i Nieto. 2019. Wav2Pix: speech-conditioned face generation using generative adversarial networks. In *ICASSP*.
[12] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. 2017. Improved Speech Reconstruction from Silent Video. *ICCV 2017 Workshops* (2017).
[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*. 2121–2129.
[14] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).
[15] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*.
[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.
[18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[19] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *CVPR*.
[20] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2017. SampleRNN: An unconditional end-to-end neural audio generation model. *ICLR* (2017).
[21] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).
[22] Shruti Palaskar, Ramon Sanabria, and Florian Metze. 2018. End-to-end multimodal speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5774–5778.
[23] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. In *ICML*.
[24] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.
[25] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *CVPR*.
[26] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*.
[27] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *CVPR*.
[28] Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto. 2018. Cross-modal embeddings for video and audio retrieval. In *ECCV Workshops*.
[29] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. 3104–3112.
[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
[31] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. 3156–3164.
[32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.