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:

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].
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. Neuronal 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 Schellens. In 2012, he obtained his Phd on image retrieval from UPC under the supervision by Prof. Ferran Marquès (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


