The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Towards Knowledge Oriented Intelligent Audio Analytics

Alessandro Oltramari,[1] **Joseph Szurley,**[1]
Samarjit Das,[1] **Jonathan Francis,**[1,2] **Juncheng Li**[1,2]
[1]Bosch Research & Technology Center North America; Pittsburgh, Pennsylvania, United States
[2]School of Computer Science, Carnegie Mellon University; Pittsburgh, Pennsylvania, United States
{alessandro.oltramari, jon.francis,joseph.szurley, samarjit.das, billy.li}@us.bosch.com

Abstract

In this position paper we discuss the benefits of combining knowledge technologies and deep learning (DL) for audio analytics: knowledge can enable high-level reasoning, helping to scale up intelligent systems from sound recognition to event analysis. We will also argue that a knowledge-integrated DL framework is key to enable smart environments.

1 Introduction

Of declarative or procedural form (Newell, Shaw, and Simon 1959), knowledge plays a crucial part in the architecture of the human mind (Anderson and Lebiere 2014; Laird, Newell, and Rosenbloom 1987): we first interact with our surroundings by reacting to perceptual stimuli, but we learn how to interpret our experiences only by reflecting on cumulated knowledge. Knowledge is also considered a fundamental part of artificial minds, or at least it used to until increasingly complex 'deep' neural networks started to perform close to human-level – and in many instances outperforming humans – in all sort of perception-based tasks: thanks to the use of high performance GPUs for machine learning, groundbreaking improvements have been recently made across a variety of applications, including image classification, video analytics, speech and sound recognition, etc. (Krizhevsky, Sutskever, and Hinton 2012).

Despite the astounding results that Deep Learning (DL) has been achieving in the last years, perception only accounts for knowledge-agnostic forms of intelligence, which common sense, logical reasoning, and semantic abstraction are not reducible to. DL frameworks can be trained to effectively recognize and reliably distinguish between sounds like *door unlocking*, *door opening*, and *door closing* (figure 3b), but are not suited to perform high-level inferences, e.g. to *understand* that a sequence of *door unlocking/opening/closing* may entail that somebody (most likely, a family member) entered the house from the front door (e.g., the only one with a lock), or to set a rule according to which, if any event-sequence of that type is recognized, the house alarm can be disarmed and the temperature increased in all rooms. In this regard, knowledge can still serve as 'propellant' of advanced machine intelligence (Sheth et al. 2017), and it's actually required to enable high-level reasoning mechanisms, improve learning algorithms, make real-time data analysis more efficient and robust. For instance, a knowledge-integrated framework would be particularly useful to design smart environments: regardless of the level of granularity, what can make cars, houses, and whole cities 'smart' is certainly learning from a variety of sensor-based data streams[1], but also the capability of aggregating and processing heterogeneous data sources according to context (Francis et al. 2017): we call the latter *sense-making*.

2 Approach

Historically, computer vision (CV) algorithms tasked with classification relied largely on domain knowledge (hand-crafted features) and often exhibited poor performance when compared to human accuracy. However, over the last decade due to advancements in DL frameworks, CV algorithms have seen large improvements in performance which often surpass that of human accuracy in classification tasks (figure 1). Advances in automatic speech recognition (ASR) systems have mirrored those in CV (figure 2), for many of the same reasons. Indeed, over the past few years, many voice assistants and chatbots have been able to exploit these large gains in ASR systems to provide a richer user experience in terms of understanding and interaction.

Although DL for audio applications has primarily been on ASR, recent attention has focused more on audio event detection and classification (AEDC). AEDC is typically more challenging than ASR, even though both are acoustic based, as audio events are more random in nature, may require a greater understanding over longer time intervals, and must simultaneously perform detection and classification on continuous streams of data. However, even with these challenges AEDC has seen similar increases in performance much like CV and ASR systems. The increase in AEDC can be attributed but not limited to several key factors:

First, the adaption of similar CV and ASR DL frameworks, which are centered around convolutional neural net-

[1]Audio-video signals from surveillance cameras and ambient microphones, occupancy and motion-based information, real-time traffic monitoring, gunshot detection systems, etc.

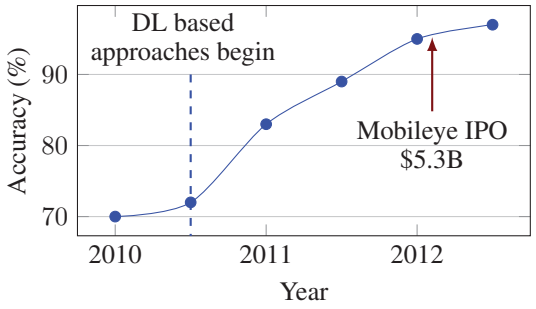


Figure 1: Comparison of the accuracy of computer vision algorithms over the years. Notice the inflection point once DL algorithms were applied. Source: (Goodfellow, Bengio, and Courville 2016)

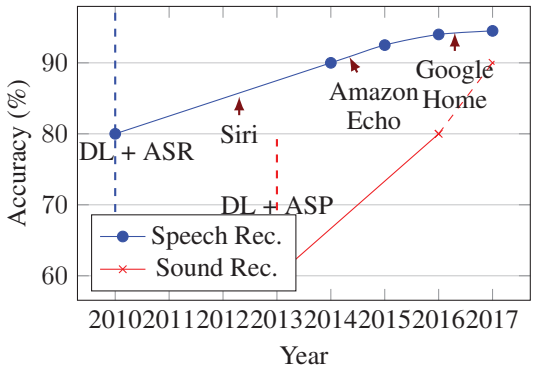


Figure 2: Comparison of the accuracy of speech and sound recognition algorithms over the years. Notice a rapid increase in accuracy once DL algorithms were introduced similarly to Figure 1. Source: (Goodfellow, Bengio, and Courville 2016)

works (CNN). To exploit the power of CNNs, the audio signal is first segmented into small, sometimes overlapping, temporal frames and converted to a time-frequency representation, or spectrogram. This spectrogram can then be viewed in a similar fashion to a natural image, where each event has a specific geometric structure. These geometric structures can then be exploited by the CNNs and used to classify trained events.

Second, recent advances in MEMS (Micro-Electro Mechanical Systems) technology have drastically decreased the size of digital microphones while at the same time improving their performance, e.g., low noise floor, near flat frequency responses, etc. This allows for the MEMS microphones to be deployed unobtrusively in essentially any environment. By exploiting these type of microphone deployments, we can slowly begin to understand what exactly the devices are *hearing*.

Third, the recent availability of large audio scene datasets (Stowell et al. 2015; Gemmeke et al. 2017) has aided in the rapid increase in AEDC accuracy. Indeed, much of the increases in CV and ASR can be directly attributed to the availability of large high quality image datasets and language corpora respectively.

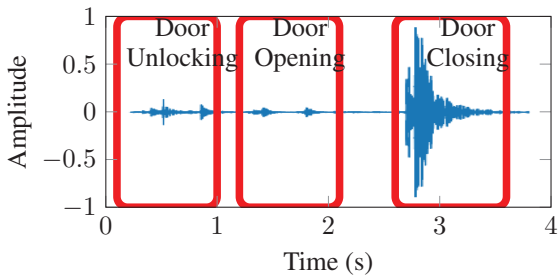
However, much like how a single classified frame, image from a video, may not be represent an entire clip, a single classified audio event does not necessarily capture the overall semantics of an audio scene. For example, in figure 3b, three different events are classified but a larger body of knowledge is needed to understand that the audio scene corresponds to someone with a key unlocking the door with a possible entry or exit event.

An 'audio scene' can be defined as a meaningful sequence of atomic 'audio events', where *meaningful* implies that individual events are aggregated according to semantic criteria, such as spatio-temporal relations (e.g., precedence) or conceptual properties (e.g., a door opening event is symmetrical to a door closing event), and *atomic* entails that individual sounds are not further decomposable (i.e., they denote minimal semantic units). Studies in Psychology show that human cognitive processing adopts high-level abstractions, also known as *schemas*, to carve perceptual contents according to principles of mental organization, optimizing the information processing effort and minimizing the cognitive overload (Albertazzi, Van Tonder, and Vishwanath 2010). For instance, human vision performs segmentation of complex scenes into action-object couplets (Tversky, Zachs, and Martin 2008): "to reduce the amount of input information into manageable chunks" (p. 457). In this regards 'visual intelligence' can be conceived as the human capability to understand a scene by means of recognizing the core interactions holding between the most salient entities detected from the environment. In this sense, perceptual data, conceptual representations and reasoning are combined together by humans to make sense of a scene: for instance, when we see a dog chasing a flying stick thrown by a person, we identify the type of entities into play (dog, person, stick) and then we break the complex event into smaller components (e.g., the person extending the arm from the back, the dog jumping and running, the stick falling on the ground, etc.), inferring its teleological features (make the dog play and bring back the stick) and causal nexus (when the persons hand releases the stick, it starts moving on air with a curved trajectory whose range depends on the exerted force). Reproducing this capability at the machine level requires a comprehensive infrastructure where low-level visual detectors and algorithms are coupled with high-level knowledge representations and processing: this was the main topic of the DARPA Minds Eye program[2], whose goal was to design artificial systems capable of analyzing the content of a video footage in real time, focusing on identification of human action types. In our previous work in the Mind's Eye program, we applied knowledge representation and reasoning to improve machine vision algorithms (Oltramari and Lebiere 2012): **analogously, we claim that human-level audio scene understanding requires DL-based sound classification to be complemented by knowledge representation and reasoning** (KRR) methods.

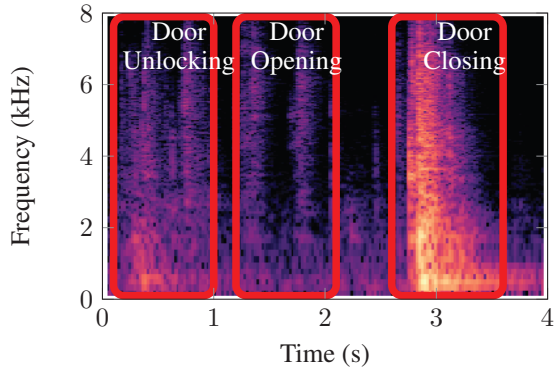
Consider the following scenario. Standard video analysis

[2]https://phys.org/news/2012-10-surveillance-tech-carnegie-mellon.html

can only detect severe failures of industrial equipment, e.g., associated with fire events or explosions, but is generally inadequate to recognize fine-grained anomalies: for this reason, sound monitoring systems are increasingly deployed in power plants to assess the status of operating machines and detect anomalies in real-time. But sound recognition systems are typically bound to individual machines, and thus can commonly target only individual audio events. By using DL-based frameworks, audio *signatures* of multiple anomalous events could be learned at scale, and a knowledge-based representation system could be adopted to represent multiple audio events *in context* (e.g., recognizing event sequences as specific patterns of malfunction), eliciting implicit knowledge (e.g., which stages of production are impacted by the anomalies, which recovery actions can be executed to repair the impacted machines, etc.). Moreover, a KRR system could infer the cause(s) of anomalies if previously encountered and documented in a suitable machine processable format, or even support the discovery of a new class of malfunctions, by generalizing from correlated properties learned through DL algorithms. Similar approaches, which combine DL and KRR frameworks, have been successfully tested for improving cancer detection, gene identification, prediction of proteins function (Danaee, Ghaeini, and Hendrix 2017; Cohen et al. 2017; Hong et al. 2017; Rifaioglu et al. 2017).
By aggregating video and audio signals[3], alongside with other sensor-based information, the scenario outlined above can be further expanded: more generally, *making sense* of a variety of data patterns using DL and KRR methods can be considered a key solution for any knowledge-intensive IoT application.



(a) Raw samples of an audio event consisting of a door unlocking, opening, and closing.



(b) Spectrogram of an audio event consisting of a door unlocking, opening, and closing.

Figure 3

3 Roadmap and Conclusions

To the best of our knowledge, we are the first to propose a combination of DL and KRR methods to enhance audio analytics with artificial intelligence. In future work we plan to test our approach and implement prototype systems accordingly, relying on distributed computing solutions, which are computationally more sustainable than monolithic software architectures, and typical of IoT deployments in smart environments. We plan to adopt distributed DL algorithms, where individual devices only have partial observations, and therefore knowledge, of an environment: by exchanging information at a local level (device-to-device), thereby conserving both communication power and computational resources, these devices can still learn a global representation of the general context. With the ubiquity of microphone equipped devices, e.g. mobile phones, this type of distributing sensing is already a possibility. Furthermore, with the introduction of such frameworks as Core ML and TensorFlow Lite, many DL algorithms are readily executable on millions of mobile platforms. Aggregating this information – while ensuring user privacy – will allow for audio scene knowledge representations and high-level inferences to span an immense expanse, and will ultimately be a key enabler for smart environments.

[3]In this context, it's worth mentioning SoundNet, the MIT's project of exploiting massive audio data available from online videos to train sound recognition algorithms: http://soundnet.csail.mit.edu/

4 Acknowledgements

This work was performed within *Bosch Corporate Sector Research & Advanced Engineering*, and results from a collaboration between two projects, Ubiquitous Personal Assistant - Connected Life (UPA-CL) and *Smart Ears Intelligent Audio Analytics Activity*.

References

Albertazzi, L.; Van Tonder, L.; and Vishwanath, D., eds. 2010. *Perception Beyond Inference. The Information Content of Visual Processes*. The MIT Press.

Anderson, J. R., and Lebiere, C. J. 2014. *The atomic components of thought*. Psychology Press.

Cohen, I.; David, E. O.; Netanyahu, N. S.; Liscovitch, N.; and Chechik, G. 2017. Deepbrain: Functional representation of neural in-situ hybridization images for gene ontology classification using deep convolutional autoencoders. In *International Conference on Artificial Neural Networks*, 287–296. Springer.

Danaee, P.; Ghaeini, R.; and Hendrix, D. A. 2017. A deep learning approach for cancer detection and relevant gene identification. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, 219–229. World Scientific.

Francis, J.; Oltramari, A.; Munir, S.; Shelton, C.; and Rowe, A. 2017. Context intelligence in pervasive environments: Poster abstract. In *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation*, IoTDI '17, 315–316. New York, NY, USA: ACM.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Hong, H.; Yin, X.; Li, F.; Guan, N.; Bo, X.; and Luo, Z. 2017. Predicting potential gene ontology from cellular response data. In *Proceedings of the 5th International Conference on Bioinformatics and Computational Biology*, 5–10. ACM.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Laird, J. E.; Newell, A.; and Rosenbloom, P. S. 1987. Soar: An architecture for general intelligence. *Artificial intelligence* 33(1):1–64.

Newell, A.; Shaw, J. C.; and Simon, H. A. 1959. Report on a general problem solving program. In *IFIP congress*, volume 256, 64. Pittsburgh, PA.

Oltramari, A., and Lebiere, C. 2012. Using ontologies in a cognitive-grounded system: Automatic action recognition in video-surveillance. In *STIDS*, 20–27.

Rifaioglu, A. S.; Doğan, T.; Martin, M. J.; Cetin-Atalay, R.; and Atalay, M. V. 2017. Multi-task deep neural networks in automated protein function prediction. *arXiv preprint arXiv:1705.04802*.

Sheth, A.; Perera, S.; Wijeratne, S.; and Thirunarayan, K. 2017. Knowledge will propel machine understanding of content: Extrapolating from current examples. In *Proceedings of the International Conference on Web Intelligence*, WI '17, 1–9. New York, NY, USA: ACM.

Stowell, D.; Giannoulis, D.; Benetos, E.; Lagrange, M.; and Plumbley, M. D. 2015. Detection and classification of acoustic scenes and events. *IEEE Trans. on Multimedia* 17(10):1733–1746.

Tversky, B.; Zachs, J.; and Martin, B. 2008. The structure of experience. In Shipley, T., and Zacks, T., eds., *Understanding events: From Perception to Action*. 436–464.