Future Vision of Interactive and Intelligent TV Systems using Edge AI

Álan L. V. Guedes
Antonio J. G. Busson,
João Paulo Navarro and
Sérgio Colcher

CITE THIS ARTICLE

Guedes, Álan L.V.; Busson, Antonio J.G.; Navarro, João Paulo; Colcher, Sérgio; 2020. Future Vision of Interactive and Intelligent TV Systems using Edge AI. SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING. ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2020.4. Web Link: http://dx.doi.org/10.18580/setijbe.2020.4

Future Vision of Interactive and Intelligent TV Systems using Edge AI

Álan L. V. Guedes, Antonio J. G. Busson, João Paulo Navarro, and Sérgio Colcher

alan@telemidia.puc-rio.br, busson@telemidia.puc-rio.br,
jpnavarro@nvidia.com, colcher@inf.puc-rio.br,

***Abstract*—Recently the Brazilian DTV system standards have been upgraded, called TV 2.5, in order to provide a better integration between broadcast and broadband services. The next Brazilian DTV system evolution, called TV 3.0, will address more deeply this convergence of TV systems not only at low-level network layers but also at the application layer. One of the new features to be addressed by this future application layer is the use of Artificial Intelligence technologies. Recently, there have been practical applications using Artificial Intelligence (AI) deployed to improve TV production efficiency and correlated cost reduction. The success in operationalize and evaluate these applications is a strong indication of the interest and relevance of AI in TV. This paper presents TeleMídia Lab's future vision on interactive and intelligent TV Systems, with particular focus on edge AI. Edge AI means use in-device capabilities to run AI applications instead of running them in cloud.**

***Index Terms*—Deep Learning; Video Analysis; TV; Edge AI.**

I. Introduction

Recently the Brazilian DTV system standards have been upgraded, called TV 2.5, in order to provide an integration between broadcast and broadband services. It is due to the widespread adoption of high-speed Internet services, TV devices also use both linear and nonlinear (on-demand) content. ITU calls this combination scenario as IBB (Integrated Broadcast-Broadband) service, which takes advantage of strong points from each one. The former has advantages regarding live events and high audiences, while the last benefits from better navigability and recommendations. The next Brazilian DTV system evolution, called TV, 3.0[1], will address more deeply this convergence of TV systems not only at low-level network layers but also at the application layer. One of the new features to be addressed by this future application layer is the use of Artificial Intelligence technologies.

According to ITU [1], there have been recently numerous practical applications of Machine Learning (ML) and Artificial Intelligence (AI) in TV They supported the broadcast program and production to improve production efficiency and correlated cost reduction. These applications include: automated programming; streaming Optimization; social media analysis; sign language synthesis; content creation from legacy archives; target advertisement; Personalized content; and others. The advances by industry to successfully operationalize and evaluate these applications is a strong indication of the interest and relevance.

This paper presents TeleMídia Lab's future vision on interactive and intelligent TV Systems with particular focus on edge AI. Edge AI means use in-device capabilities to run AI applications instead of running them in the cloud. In order to present this vision, we firstly present the state of art of ML/AI for TV (Section II). Then we discuss our future vision and present our final remarks (Section III and Section IV).

II. State of the art

A. Media Analysis

Methods based on Deep Learning (DL) became the state-of-the-art in various segments related to automatic media analysis. More specifically, Convolutional Neural Networks (CNN) architectures, or ConvNets, have become the primary method used for audio-visual pattern recognition. The classification task consists of mapping media content into one or more distinct categories. Deep Learning architectures based on CNN's (Convolutional neural network or ConvNets) have become the main method used for recognizing audiovisual patterns. Next, we present some method DL methods for image and video analysis.

Image Analysis. Since the victory of the AlexNet [2] in the ImageNet 2012 challenge, new and more accurate CNN-based architectures have emerged. The winner of ImageNet 2014, for example, was the InceptionNet [3], which proposed the use of the Inception block, a block that uses several filters of different sizes at the same level to solve the problem of finding information in images. One year later, the ResNet [4] network was the winner of ImageNet 2015, introducing the concept of residual connections, which increased performance and reduced the training time for CNNs. Later, the Inception-Resnet [5] architecture was proposed as a combination of the Inception blocks with the residual connections, producing one of the most popular models that form the basis for many other CNN architectures for extracting features. The SE-Net architecture (Squeeze-and-Excitation Network) [6] is the state-of-the-art in the image classification task,2 obtaining 2.25% error top-5 at ImageNet 2017. SE-Net proposes a new type of block called SE, which improves the network's power of representation by highlighting the inter-dependencies between the image channels and their features maps. For this, SE-Net uses a mechanism that allows the network to re-calibrate features, through which it uses global information to emphasize the most informative features and suppress the features less useful.

[1] https://forumsbtvd.org.br/tv3_0/

http://www.set.org.br/ijbe/ doi: 10.18580/setijbe.2020.4 Web Link: http://dx.doi.org/10.18580/setijbe.2020.4

Video Analysis. Unlike images, videos are not only visual but also have audible content. Current methods for video classification are generally divided into two stages: (1) the CNNs stage, called backbone, used to extract audio-visual features from the video content; (2) the after-extraction stage, comprised of sophisticated methods for aggregating features, such as NetVLAD [7] and NetFV [8], that can be applied to undermine audio-visual features and perform classification. These methods achieve state-of-theart performance on the YouTube-8M video classification task [9]. To extract the visual features from video, CNNs (e.g., Inception [3], ResNet) pre-trained in the dataset ImageNet are often used. For the extraction of features from the audio, CNNs adapted for the audio domain, such as AudioVGG [10] or WaveNet [11], pre-trained in dataset AudioSet are the most used models. Currently, most of the video analysis community is devoted to the task of classifying video at the segment level (i.e., temporally locating and classifying the segments in the videos). The Video Action Transformer Network [12] is a proposal to localize and classify actions in space and time.

B. AI for TV

Practical AI applications in broadcast program and production are used to improve production efficiency and correlated cost reduction [1]. It was made possible given the methods such as the image and video analysis discussed in the previous subsection. We discuss some of such applications in what follows.

Automated Video Previews. Previews of programs and digest videos give viewers brief introductions of their content. NHK has developed image analysis to recognize characters and performers in order to create previews and digests. Wimbledon Championships also have efforts, in partnership with IBM, to generate highlight clips. Automated Programming. BBC has projects that aim using ML/AI algorithms to mine through thousands of hours of legacy archived content dating back to 1953 to generate programming. In particular, they use information from past scheduling, past audience, content metadata, and other program attributes to learn and mine the archived content for optimal targeted demographic programming.

Automated Camera. Broadcasters traditionally use multiple ultra-high-definition cameras for capturing during live events. The capture from these cameras has been used as a feed to a highly reduced or even single-operator human-driven system. The BBC has implemented ML efforts to allow fewer cameras and camera operators to capture a richer scene and environment. Moreover, Endemol Shine Group used ML methods in the Spanish version of the reality show “Big Brother” to capture patterns of interactions happening in the relationships and dynamics of the house members. This way, they anticipate relationship dynamics of the group interactions and direct camera and recordings costs. Compliance tracking and content creation. Companies have also targeted news workflows to help facilitate and improve FCC mandated compliance in production and delivery. TVU Networks, for instance, provides a transcriber service that assures that all video content is US FCC compliant prior to on-air broadcast Additionally, they make use of ML/AI algorithms to mute audio during any profanity or excluded speech.

Social Media Analysis. Social media can influence and become a critical part of broadcasting, particularly in breaking news. ML systems can automatically search for trends and notable information for news production from a massive amount of Twitter data and subsequently judge the authenticity of such posts to determine the presence of any target terms typically associated with news-worthy breaking events. These terms (e.g., "fire" and "accident") are categorized information that is representative of types of information that can be featured in news programs Additionally, research is underway to improve image identification of relevant events and objects, from user image Twitter data.

Streaming Optimization. The industry is also using AI to improve video encoder efficiency. BitMovin[2] is successfully using AI to enable learning the video content complexity and other features from prior video encode to improve quality and efficiency during later stage encoding. This may result in optimized quality at maximum bandwidth efficiency

(a) Digital Announcer (b) Transcription

Figure 1: NHK Speech Applications [1]

Digital Announcer. NHK has successfully used an AI-driven announcer in one of the news programs. It is a CG-generated announcer (Fig. 1a) with a high-quality synthesized voice that reads out news manuscripts about topics being discussed on social networks. As input training data for ML voice synthesis, it uses text data collected from a large number of past news manuscripts and audio data as read by professional announcers This training facilitates the automated reading of news manuscripts in a natural cadence and prosody, reflecting the characteristics of professional newsreaders. NHK has also used the same system in an automatic voice generation scenario for weather information programs. The information is spoken by the system use a style similar to a professional announcer.

Transcription support. Live content such as conversations and interviews also require a way to effectively and3 accurately transcribe speech. NHK transcription system (Fig. 1b) greatly improve their transcription workflow. It uses ML speech recognition algorithms to not only generate text but also categorize each individual in the program cast.

Sign language CG synthesis. Sign Language accessibility is required for hearing-impaired individuals. There are some ML method efforts to translate the text to Sign language using a CG animation agent. This translation is enabled through the acquisition of training data from content with Sign Language. However, the corpus vocabulary available is small and still cause the translation to be inaccurate. Before achieving a larger corpus, NHK currently uses a speech recognition-

[2] https://bitmovin.com/

http://www.set.org.br/ijbe/ doi: 10.18580/setijbe.2020.4 Web Link: http://dx.doi.org/10.18580/setijbe.2020.4

based tool to support experts fixing Sign Language translation.

Content Creation from Legacy Archives. Colourisation of monochrome images in historical or documentary programs can enable more realistic, engaging, and immersive content production. NHK has developed a colourisation technology for monochrome video images that uses AI to learn the colors of various objects in advance, which learns the colors prevalent in the sky, mountains, or buildings.

Metadata Creation. Program producers often search for archived video footage and audio files for possible reuse during program production. To support this search, ML methods have become quite effective at successfully automating metadata generation in legacy and new content. Image and Video Analysis (Section II-A) technologies make it possible to automatically describe relevant information and features of the scenes. For instance, multiple TV affiliates use Prime Focus system to improve media asset management by recognizing elements within audio and video content and automatically generate associated metadata. Moreover, NHK has used text detection to add metadata about scenes depicted in TV programs, such as public signs, that can be used to identify the location or address of a building. In a similar way, Nippon TV uses ML methods to detect the player profiles and uniforms and then learns to recognize "individual players" in live event sports.

Content Personalization. Content personalization efforts can be deeply supported by ML/AI algorithms. The relevant efforts include demographically targeted and optimized content for different audiences. For instance, BBC used AI algorithms over archived content to create optimized programming to the user demographic. Its methods of learning include scene identification (e.g. landscapes, objects, people); text metadata, including subtitles; and motion activity of videos.

III. Future Vision

Today, most ML/AI algorithms run and use data stored in cloud services. This way, they take benefit from the scalability, redundancy, and better costs of those services. The cloud-based ecosystem has demonstrated itself as a practical platform to serve some AI applications. However, it has limitations that might prevent the adoption of other AI fields [13]. For instance, we cite autonomous driving that requires robustness and latency to enable vehicle safety and prevent collisions. In this direction, we recently see the rise of Artificial Intelligence at Edge (Edge AI). This approach proposes that the ML/AI algorithms be processed locally on a hardware device instead of the cloud. Such a device uses data, e.g. sensor data or signals, that are created/captured on the device. In particular, the device does not require to be always connected to the internet. This way, Edge AI may reduce costs for data communication, because fewer data will be transmitted because it is processing locally. This is also important regarding streaming and personal data application to prevent making the user vulnerable from a privacy perspective. The expectations from devices at the Edge are ever-growing especially the new in-device hardware specifically designed for AI tasks.

In our future vision for TV, we envisioned Edge-AI services running at the TV receiver. So, TV will also embed hardware specifically designed for AI tasks. To enable this vision, it is required to extend the current TV application layer to take advantage of such features. As previously mentioned, the current TV application layer focuses on support media presentation for both linear and nonlinear (on-demand) content. The extension for Edge-AI is required supporting programmers to describe learning data and recognize content semantics of the media/data at the receiver.

Some efforts have already tried to extend the NCL language, standardized in SBTVD application layer, with AI features. Moreno *et al.* [14] propose an NCL extension that supports the specification of relationships between knowledge concepts. Their approach is inspired by knowledge engineering standards, such as RDF and OWL. To recognize concepts during the presentation of media, they propose a new type of trigger event (called *InferenceRole*), which is an expected concept to be recognized in the media. Guedes *et al.* [15] extend the NCL to support the development of NCL applications with multimodal interactions. It defines expected user interactions through multimodal descriptions (e.g., SRGS for voice recognition). These descriptions are used in a new <*input*> element to represent input devices. Such elements may have a virtual anchor, called *RecognitionAnchor*, that triggers a recognition event when an expected interaction is recognized from the input device. Abreu and Santos [16] propose the *AbstractAnchor*, which is an anchor type that represents parts of a content node where concepts are detected. Then, during the document parsing, the processor analyses all the media and create the timestamps relative to the time interval where expected concepts are recognized in each media. Busson *et al.* [17] propose the *SemanticAnchor* to allow recognition events in run-time. This way, we also can perform recognition events even in applications of live streams. Additionally, they also offer access to properties of recognition events, such as: identifier of the recognized event, part of media where that concept appears, time of recognition event trigger, etc.

UC-01 In-device TV Program Analyses. ML methods for recognition, scene classification, and speech recognition, make it possible for TV to understand the semantics of the content of the transmitted video. By using a concept-aware application layer, TV applications can use the semantics of video content as an anchor to trigger events. For example, imagine an e-commerce TV application that notifies viewers of the availability of purchasing a product whenever they appear on the TV program.

UC-02 In-device Sensitive video filtering. TV applications may also present user-generated live streaming (e.g., music, lifestyle and gaming), which are available on social media. There are, however, many examples of uncomfortable situations during video conferences when a participant forgets to disable his/her microphone or camera and goes to the bathroom or change clothes. Moreover, some private video services leave room for the spread of sensitive and inappropriate content for certain ages, such as pornography, violence, or other potentially offensive content. This kind of exposure is especially concerning when considering the vulnerability of children spectators. This way, current AI methods for sensitive video analysis [18] can be used to

http://www.set.org.br/ijbe/ doi: 10.18580/setijbe.2020.4 Web Link: http://dx.doi.org/10.18580/setijbe.2020.4

automatically block this during their presentation on TV device.

UC-03 In-device Personalized Content Creation. With the TV empowered by AI, personalized content based on audio-visual content and subtitles can be generated automatically. Some useful applications involve content translation and auxiliary content for people with disabilities, such as sign language video [19] and audio-description [20]. Imagine, for example, instead of the broadcaster sending the video of sign languages superimposed on the main video, sign languages could be generated in a personalized way only by TV of persons with hearing disabilities.

UC-04 Spectator Understanding. TV application may gather user's identification (face recognition) or context information (sentimental analyses) to trigger pre-configured actions, such as: turn-off TV when the user gets slept; suggest new content when the user is not happy with the current one; suggest animation movies when the children are present. Moreover, the user may require to not store personal data in broadcasters' servers. This way, TV application must process ML task in user data (profile, channels choices) locally to provide recommendations.

UC-05 Virtual TV Assistant. A Bot is a conversational software that interacts with users. It uses natural language processing in users' textual input, or intents, to perform actions or responses with information. Today, they are used in different industry fields and especially in services regarding consumer support. A Bot may act as a Virtual TV Assistant running in-device and reacting to user actions or behavior. It may support meeting actions such as stop/start recording or even configure background music. In particular, the questions may be regarding the currently selected video, such as: "what is the name of this character"; and "what happens with him in the last episode".

IV. Final Remarks

In this paper, we presented an overview of the state of-the-art in Deep Learning for media content analysis (image, audio, and video) and described recent works that propose the integration of ML/AI and TV systems. Then we presented our future vision regarding the Edge-AI in the TV receiver. More precisely, we envisioned the ML/AI use on TV to enable in-device media and data analysis. To support such a vision, we describe a set of envisaged use cases to define new requirements and API design. As future work, we cite performing participatory design with TV programmers.

References

[1] ITU. (2019) Artificial intelligence systems for programme production and exchange. Available: https://www.itu.int/pub/R-REP-BT.2447-2019

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in AAAI, vol. 4, 2017.

[6] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2019.

[7] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.

[8] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," arXiv preprint arXiv:1706.06905, 2017.

[9] M. Skalic and D. Austin, "Building a size constrained predictive model for video classification," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[10] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017. Available: https://arxiv.org/abs/1609.09430

[11] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[12] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 244–253.

[13] Y.-L. Lee, P.-K. Tsung, and M. Wu, "Techology trend of edge ai," in 2018 International Symposium on VLSI Design, Automation and Test (VLSI-DAT). IEEE, 2018, pp. 1–2.

[14] M. F. Moreno, R. Brandao, and R. Cerqueira, "Extending hypermedia conceptual models to support hyperknowledge specifications," International Journal of Semantic Computing, vol. 11, no. 01, pp. 43–64, 2017.5

[15] Á. L. V. Guedes, R. G. de Albuquerque Azevedo, and S. D. J. Barbosa, "Extending multimedia languages to support multimodal user interactions," Multimedia Tools and Applications, vol. 76, no. 4, pp. 5691–5720, 2017.

[16] R. Abreu and J. A. dos Santos, "Using abstract anchors to aid the development of multimedia applications with sensory effects," in Proceedings of the 2017 ACM Symposium on Document Engineering, 2017, pp. 211–218.

[17] A. J. G. Busson, Á. L. V. Guedes, S. Colcher, R. L. Milidiú, and E. H. Haeusler, "Embedding deep learning models into hypermedia applications," in Special Topics in Multimedia, IoT and Web Technologies. Springer, 2020, pp. 91–111.

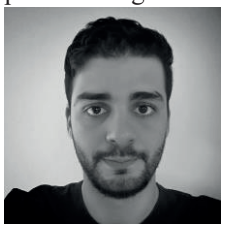
[18] P. Almeida, A. Busson, L. V. Guedes, and S. Colcher, "A deep learning approach to detect pornography videos in educational repositories," in Brazilian Symposium on Informatics in Education (Simpósio Brasileiro de Informática na Educação - SBIE), vol. 30, 2020, p. 912.

[19] V. Veríssimo, C. Silva, V. Hanael, C. Moraes, R. Costa, T. Maritan, M. Aschoff, and T. Gaudêncio, "A study on

the use of sequence-to-sequence neural networks for automatic translation of brazilian portuguese to libras," in Proceedings of the 25th Brazillian M. Hasegawa-Johnson, A. Black, L. Ondel, O. Scharenborg, and F. Ciannella, "Image2speech: Automatically generating audio descriptions of images," Proceedings of ICNLSSP, Casablanca, Morocco, 2017.

Álan L. V. Guedes is researcher at PUC-Rio TeleMídia laboratory and contributed Ginga and NCL standards to the SBTVD and ITU Forum. His research interests include Interactive Multimedia, Immersive Media, and Multimedia Analysis.

Antonio Busson is a PhD student in Informatics at PUC-Rio and a researcher at TeleMídia laboratory. Graduated (2012) and Master (2015) in Computing at UFMA. His research interests include multimedia/hypermedia systems and pattern recognition



João Paulo is a Solution Architect at NVIDIA with a focus on high-performance computing and Deep Learning. He has extensive experience in the development of algorithms and visualization techniques aimed at geophysical processing.

Sérgio Colcher is professor at the PUC-Rio Department of Informatics (DI) and coordinator of the TeleMídia laboratory. He was also professor of MBA courses in Telecommunications Management and MBA in e-Business at Fundação Getúlio Vargas. His areas of interest include computer networks, performance analysis of computer systems, multimedia systems and pattern recognition

Received in 2020-10-12 | Approved in 2020-11-17

http://www.set.org.br/ijbe/ doi: 10.18580/setijbe.2020.4 Web Link: http://dx.doi.org/10.18580/setijbe.2020.4