AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "BIOINSPIRED LEARNING"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Nicu Sebe** (University of Trento) **Wei Wang** (University of Trento) **Cigdem Beyan** (University of Trento) **Marco Formentini** (University of Trento) **Kasim Sinan Yildirim** (University of Trento) **Enver Sangineto** (University of Trento)

This report is part of the deliverable D2.3 - "*AI technologies and applications in media: State of Play, Foresight, and Research Directions*" of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

Bioinspired learning

Current status

During the last decade, the multimedia and computer vision research communities have witnessed the revolution brought by deep artificial networks, which are inspired by the biological visual system. However, there still exist discrepancies between deep networks and biological ones. To advance beyond the current deep learning scheme, one needs to re-think and re-model current deep network architectures by reverse engineering of the human visual system including the cognitive patterns, neuron connections, and the capability of continual learning.

Deep neural networks take inspiration from the human brain. The quick development of computing platforms (*e.g.*, Graphics Processing Unit[1] (GPU) and Tensor Processing Unit[2] (TPU)) provides strong computing power and paves the way to further development of AI. It has been shown for object recognition[3,4,5], tracking[6], image labelling[7] and other fields that features learned for a specific problem using deep convolutional neural networks (CNNs) show much better performance than traditional machine learning approaches. Instead of splitting the feature and classifier learning processes, CNNs supports *end-to-end* learning of the feature extractor and classifier simultaneously. The *end-to-end* learning mechanism enables CNNs to learn task-specific features automatically. The very first CNN model is LeNet[8] proposed in 1998. Eventually, after nearly 15 years, with the help of powerful computing platforms (GPU, TPU), ground-breaking models winning the ImageNet Large Scale Visual Recognition Challenge[9] were established, including AlexNet[3] in 2012, VGG19[7] & GoogleNet[4] in 2014 and ResNet[5] in 2015. Since then, no significant progress has been made and the new models are usually an ensemble of previous models.

[1] J. Sanders and E. Kandrot. CUDA by example: an introduction to general-purpose GPU programming. Addison-Wesley Professional, 2010.
[2] N. P. Jouppi, et al. In-datacenter performance analysis of a tensor processing unit. In International symposium on computer architecture, pages 1–12, 2017.
[3] A. Krizhevsky, I. Sutskever, and G. E Hinton. Imagenet classification with deep convolutional neural networks. NeurIPS, 25:1097–1105, 2012.
[4] C.n Szegedy, et al. Going deeper with convolutions. In CVPR, pages 1–9, 2015.
[5] Kaiming He, et al. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.
[6] N. Wang and Dit-Yan Yeung. Learning a deep compact image representation for visual tracking. In NeurIPS, 2013
[7] K. Simonyan and A.Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[8] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.
[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248–255, 2009.

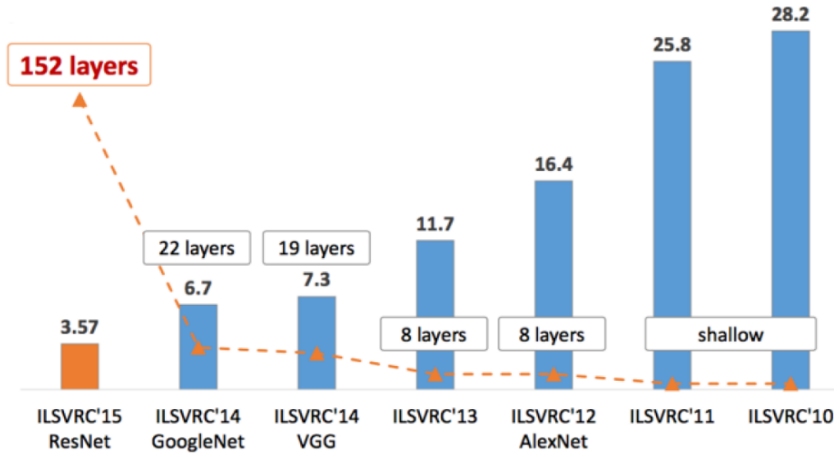

Figure 1: Evolution of CNN architectures[10]. The object recognition rate (%) of each network is depicted as a bar.

The **ImageNet competition** made a large contribution to the development of AI. ImageNet is a large visual database designed for visual object recognition research with more than 14 million images. ImageNet runs an annual contest, *i.e.*, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where software programs compete with each other to correctly classify and detect objects and scenes. Both the industry and research communities devote their energies in the ILSVRC competition by making their CNN models deeper and wider. With the popularity of CNNs, giant companies such as Google, Facebook, Microsoft, and Amazon also take active part in the development of CNNs. Figure 1 above shows the evolution of CNN architectures. The bar represents the object recognition error (smaller means better). We can observe that as the network goes deeper, the performance keeps improving. With the network depth gradually increasing from 8 to 152, the recognition error drops from 16.4% to 3.6%. Note that depth increase does not necessarily lead to the increase of the number of parameters. However, in the learning stage, the CNN model parameters and their intermediate features and gradients need to be stored. Therefore, more memory is required to train the very deep CNN models, and for example a 152 layer deep ResNet could drain out the memory and computing power of one individual GPU. Later on, more complicated CNN models such as DenseNet[11] needed to run on GPU clusters in order to have enough memory to host the model. In other words, the limited memory and computing power of the GPU have restricted the CNN models to go even deeper. Therefore, how to optimise CNN architecture to achieve better performance without increasing the memory and computing power consumption will be the main focus of the AI community in the future. The solution lies in the ***reverse engineering of the human vision system***.

To further advance the development of AI, computer science research scientists have tried to mitigate the gap between artificial and biological neural networks. For instance, *Geoffrey Hinton*, the pioneer of AI, has published two open access research papers[12,13] on the theme of capsule

[10] Image source: OpenGenus, Evolution of CNN Architectures: LeNet, AlexNet, ZFNet, GoogleNet, VGG and ResNet https://iq.opengenus.org/evolution-of-cnn-architectures/

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In CVPR, pages 4700–4708, 2017.

[12] S. Sabour, N. Frosst, and G. Hinton. Dynamic routing between capsules. arXiv preprint arXiv:1710.09829, 2017

[13] G. Hinton, S. Sabour, and N. Frosst. Matrix capsules with EM routing. In ICLR, 2018.

neural networks which can be used to better model hierarchical relationships. This approach is an attempt to mimic the organisation of biological neural structures more closely. In the meanwhile, from the neuroscience research community, many works on biologically inspired artificial neural networks have been developed, such as the spiking neural networks[14]. All these works share a similar vision, *i.e.,* **bringing more neural realism into deep networks and reducing the differences between the artificial and biological neural networks**. The human vision system perceives the outside world in a more efficient way which is quite different from CNN models. The main differences lie in **i) recognition patterns, ii) network topological structures, and iii) the memory mechanism.**

Research challenges

To reduce the gap between deep neural networks and biological ones, we need to take a closer look at their differences from three aspects: i) Is it possible to build an artificial neural network that has the same cognitive pattern and behaves in the same way as the biological human visual system? ii) Can we design a smarter artificial neural network by exploring diverse neural network topologies that exist in the human brain? iii) Is it possible to enable the artificial network to have a similar continual learning ability as a human?

To answer these three questions, the solution is to i) simulate the dual stream cognitive pattern of human vision; ii) model the diverse topological structures of biological neuronal circuits; and iii) explore the possibility of continual learning to reach a quite similar deep neural continual learning ability as a human.

One fundamental principle in human cognitive patterns is that the human visual cortex possesses two distinct streams i.e., ventral and dorsal as shown in Figure 2[15]. The ***ventral stream*** ('what' path- way) is involved in high-level perception[16] (e.g., object/scene recognition) while the ***dorsal stream*** ('where' pathway) is involved in spatial cognition. The two streams correspond to the classic definition of computer vision proposed by David Marr[17] which is to look at 'what' is 'where'. In the context of computer vision, 'what' denotes object recognition (object vision) and 'where' refers to 3D reconstruction and object localisation (spatial vision)[18]. This paradigm guides the research in computer vision, but the spatial and object vision tasks are usually studied independently. However, most deep networks, such as ResNet designed for either classification, segmentation or object detection have focused on designing one-shot methods, that is, algorithms that take an image as input, process it, and return an output without any feedback loop. This is in contrast with what we know about the human vision system where the two streams work collaboratively for the perception of the outside scene.

[14] W. Gerstner and W. Kistler. Spiking neuron models: Single neurons, populations, plasticity. Cambridge university press, 2002.
[15] D. Milner and M. Goodale. The visual brain in action, volume 27. OUP Oxford, 2006
[16] L. Cloutman. Interaction between dorsal and ventral processing streams: where, when and how? Brain and language, 127(2):251–263, 2013
[17] D. Marr. Vision: A computational investigation into the human representation and processing of visual information. 1982.
[18] M. Mishkin, L. Ungerleider, and K. Macko. Object vision and spatial vision: two cortical pathways. Trends in neurosciences, 6:414–417, 1983.

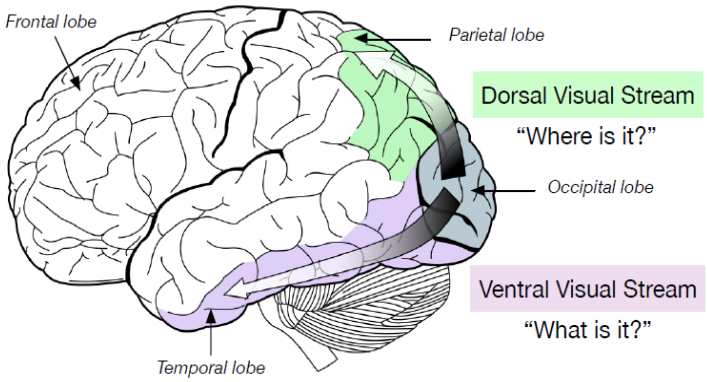

Figure 2: Human vision system. The dorsal stream determines 'where is it' while the ventral stream determines 'what is it'. The two streams originate from a common source in the visual cortex.[19]

This dichotomy shows that it is very necessary to design a **dual stream recognition pattern for recursive coarse-to-fine perception**: the ventral network is in charge of high-level perception while the dorsal network retrieves the memory associated with the ventral network and attends to salient objects to refine the perception repeatedly. In this way, the dual stream networks compose a cyclic loop and refine the perception progressively.

Second, the deep network neurons are activated non-linearly in the style of human neurons, but some important topological structures are ignored, such as the recursive connection. Besides, the biological neural network has very complex local topological structures as shown in Figure 3. Diverse combinations of these structures lead to various networks with diverse global structures. Many classical modules in artificial networks can be viewed as simplified duplicates of the biological neuronal circuits. For instance, the residual connection in ResNet can be viewed as a special case of a parallel neuronal circuit[20] (as shown in Figure 3(4)), in which there is only one branch running in parallel with the main stem. Besides, the inception module in GoogleNet can be viewed as the combination of diverging and converging circuits as shown in Figure 3 (1) & (2).

This limited amount of artificial modules with simplified local topological structures has helped the deep models to achieve remarkable performance in computer vision tasks. We can push it one step further through **modelling the neural circuits with rich topological structures**, and thus to design more powerful artificial modules which can model more complex functions. To bring this into reality, we need to rely on the strong computing power of servers nested with GPU clusters or TPUs. In the meanwhile, AI has defined its own advantages *w.r.t.* its electronic computing platform, such as the speed, reconfigurability, parallelisation, and scalability. It has the potential to surpass the cognitive intelligence it tries to mimic. For instance, our eyes can only orient our gaze to one salient object at a time, while the computer can process different regions in parallel.

[19] Image source: Wikipedia - https://en.wikipedia.org/wiki/File:Ventral-dorsal_streams.svg
[20] K. Saladin and R. McFarland. Human anatomy, volume 3. McGraw-Hill New York, 2008.

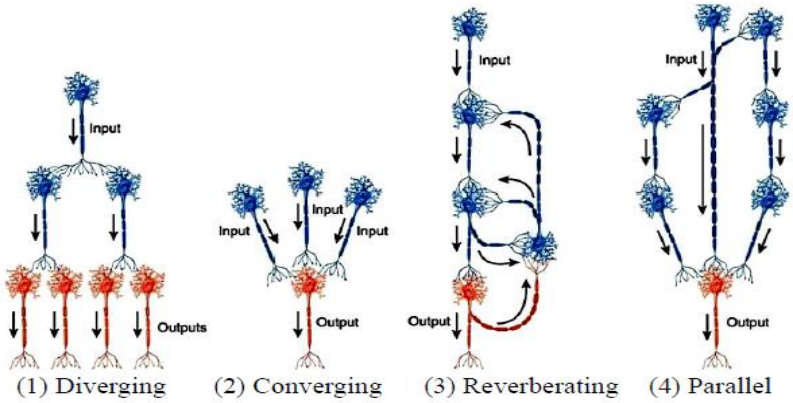

Figure 3: Biological neuron circuits.[21]

Moreover, in the real world, we are exposed to continuous streams of information. To adapt to the changing environment, we are able to learn multiple tasks from dynamic data distributions in a continuous manner. The ability to continually learn over time by accommodating new knowledge while retaining the previously learned one is referred to as ***continual or lifelong learning***. In the context of AI, it means being able to smoothly update the artificial network to perform more tasks but still being able to *re-use* and *retain* knowledge that has been previously learned without forgetting it.

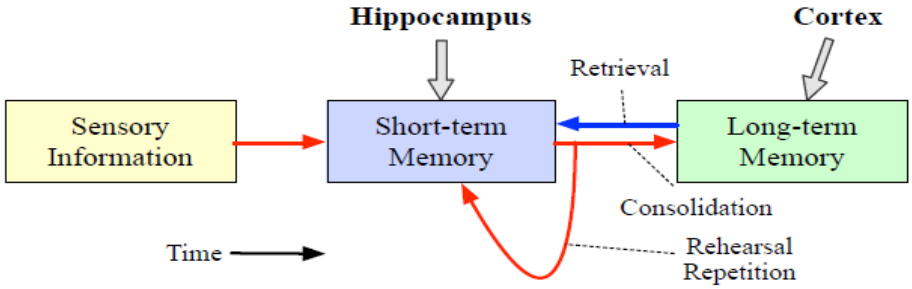

Figure 4: Dual memory system.

By studying *Henry Molaison*[22] who was unable to form new memories after removing his hippocampus to treat epilepsy, neuroscientists revealed the dual memory mechanism in our brain which may play a vital role in continual learning. It is believed that the hippocampus has a short-term memory and it is involved in rapid learning of new tasks. It encodes sparse representations to minimise interference. In contrast, the neocortex has a slow learning rate and is involved in learning generalities by building overlapping representations of the learned knowledge. As shown in Figure 4, knowledge is transferred from short-term memory to the long-term storage memory via knowledge consolidation[23]. The long-term memory can then be

[21] Image taken from: Curtis DeFriez , "Chapter 12 Nervous Tissue" at https://slideplayer.com/slide/5964323/ (slide 89)

[22] The Brain Observatory, Deconstructing Henry, https://www.thebrainobservatory.org/projecthm

[23] Y. Dudai. The neurobiology of consolidations, or, how stable is the engram? Annu. Rev. Psychol., 55:51–86, 2004.

recalled and reconsolidated[24] due to *neural plasticity*[25], meaning that it can be adapted by acquiring, refining, and transferring knowledge across multiple domains[26]. The cooperation between hippocampus and neocortex is key to learn high level regularised concepts and memory, but the exact mechanisms are still not yet completely understood. To address the catastrophic forgetting problem in continual learning, one solution is to **model the dual memory mechanism**. Specifically, the regularised concepts that are shared across tasks will be represented by an attribute dictionary, which will be stored in the long-term memory module. The short-term memory module is allowed for gradually forgetting and it can keep learning new attributes and transfer them to the attribute dictionary. By referring to the dictionary, novel objects can be easily represented. For instance, with the following attributes: 'shape like horse' and 'black-white' 'stripes, we are able to represent 'zebra'.

Societal and media industry drivers

Vignette 1: AI-enabled web-camera for highly realistic virtual interactions in the Metaverse

Marco is having a meeting with his friends. Because of the pandemic, they could not meet onsite. Therefore, they have to stay in their own room and join an online 3D virtual meeting room in which they project themselves as 3D-realistic avatars in the virtual 3D spaces. The webcam has the same cognitive pattern as a human and it can synchronise the avatar and Marco when he talks, moves, and interacts. When humans observe a scene, their eyes look into different directions and orient their gaze to the location where a visual object has appeared. For example, when observing a still face image, our eyes undergo a saccadic movement and re-target to salient regions (see Figure 5). With an AI-enabled web-camera, the moving trajectory of Marco's pupils can be tracked in real time to know his gaze orientation. Together with the sensors that detect his head movement, Marco's views in the virtual environment can be changed automatically. With the help of such a cognitive pattern, the webcam can capture all of Marco's movements and make the avatar come to life.

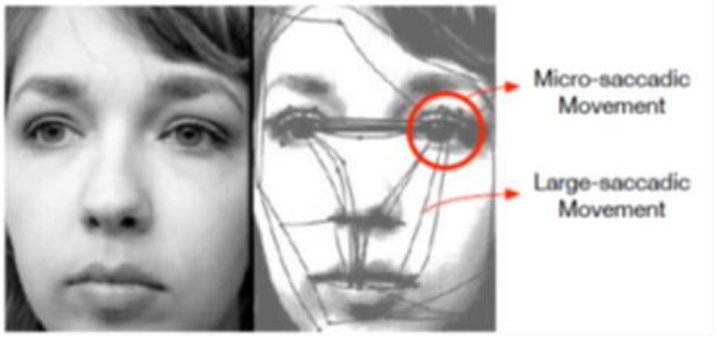

Figure 5: Saccadic eye movements.[27]

[24] K. Nader, G. Schafe, and J. Le Doux. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. Nature, 406(6797):722–726, 2000.
[25] E. Fuchs and G. Flugge. Adult neuroplasticity: more than 40 years of research. Neural plasticity, 2014, 2014.
[26] A. Bremner, D. Lewkowicz, and C. Spence. Multisensory development. Oxford University Press, 2012.
[27] Image source: Wikipedia - https://en.wikipedia.org/wiki/Saccade#/media/File:Szakkad.jpg

Vignette 2: Playing video games with smart AI companions

John is wearing a virtual reality camera and is playing a game with an AI agent. After being familiar with the rules, John can always win the game with the same strategy and he feels bored. He then selects a smarter AI agent with the ability of continual learning. The smarter agent can keep learning from mistakes without forgetting the experience learned previously. Similar to a human, the AI agent can adapt itself by acquiring, refining, and transferring knowledge. After each game, the AI agent will summarise the good strategies of good actions and bad strategies of bad actions. These strategies and experiences are all stored in the long term memory unit of the agent and it can behave more like a human. Therefore, John must come out with new strategies to win the game. Moreover, the AI agent can keep evolving automatically and become smarter by playing games with another smart AI agent. As such, John will not be bored anymore.

Future trends for the media sector

The progress of neuroscience and deep learning theory[28], together with the development of powerful computing platforms (from CPU to GPU and TPU), make up the basis for the development of AI. The next evolution of social networking is to help bring the Metaverse to life. One of its main characteristics is the use of 3D spaces that can let one socialise, learn, collaborate and play in new ways. To bring this to life, 3D modelling is the key. Therefore, it is very important for AI to understand the 3D world so that the 3D objects could be well reconstructed in the virtual 3D spaces.

As humans, we naturally have the ability to extract 3D information using our dual stream visual cognitive pattern. Figure 6 shows the dual stream visual cognitive pattern of a human. The ventral stream recognises that the image shows a pair of shoes (left). Next, the dorsal stream network will pay attention to the shoes and extract the 3D location (middle) and shape information which is represented by the segmentation mask (right). By mimicking human visual cognitive patterns, AI can better understand and reconstruct the 3D scenes and improve the user experience.

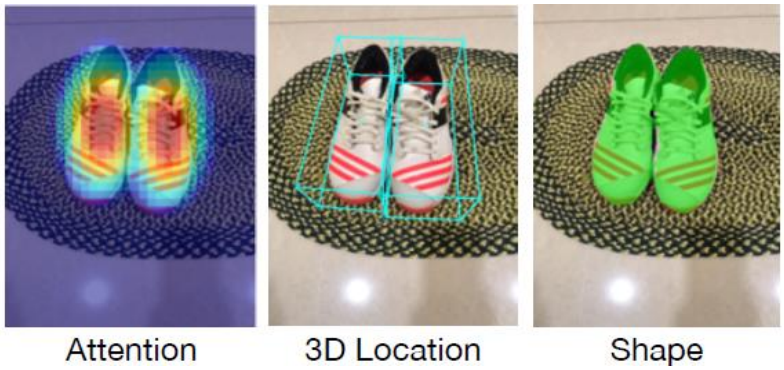

Figure 6: Dual stream recognition example.[29]

[28] Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. Frontiers in computational neuroscience, 10:94, 2016.

[29] Image source: MediaPipe - https://google.github.io/mediapipe/solutions/objectron.html

Besides, social media has a huge amount of content in the format of videos, images, texts, etc. It is very necessary to summarise these contents so that we can better categorise them and present them to users. Even though we already have many deep neural networks to do the job, they are usually only good at one media format. For instance, the network may be good at action recognition in videos, but not good at face recognition. Therefore, it is very necessary to design more powerful deep neural networks. One potential solution is to model the various topology structures of the neurons in the human brain. In this way, the deep neural networks may be good at multiple tasks across different media formats. Moreover, the world is changing dynamically, and new knowledge keeps coming out. It is also very important for the deep neural networks to learn continuously to adapt themselves by acquiring new knowledge, refining them, and transferring them across multiple domains.

Goals for next 10 or 20 years

Instead of focusing on pursuing high performance and fast speed relying on the target computing resources (*e.g.*, GPUs, embedded devices), in the future, AI will copy the processes that underlie the way a biological system thinks and remembers and will take them one step closer to a real living biological system. By rethinking the artificial networks from the view of the biological system and the reverse engineering of the human visual system based on the basic theories discovered by neuroscientists, AI will behave in a more similar way as humans.

The long-term vision of AI is to enable artificial networks to process and perceive visual information like the human visual system and to learn knowledge continually like the human brain (next 5-10 years). By pushing the artificial network closer to the biological one, it might be easier to integrate artificial networks into biological ones. This will also have great potential to replace damaged brain sectors or make our brain more powerful by planting artificial circuits into it (next 10-20 years).

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu