Deep Learning Architectures for Image and Speech Recognition: A Comprehensive Review

Dr. Sofia M. Kovacs
Department of Electrical Engineering and Intelligent Systems,
Budapest University of Technology and Economics, Hungary

Received: 10/07/2025 ; Accepted:14/02/2026 ; Published: 28/04/2026

Abstract
Machines can now learn hierarchical feature representations from raw data thanks to deep learning, which has transformed picture and speech recognition. Recognition systems have been much better and more reliable over the last decade, thanks to sophisticated neural network topologies that outperform the old-school machine learning methods. Due to its superior performance in picture categorization, object detection, and feature extraction, Convolutional Neural Networks (CNNs) have emerged as the backbone of image recognition tasks. Similarly, designs based on Transformers, Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have all done an excellent job at modeling sequential data for voice recognition. thorough examination of deep learning architectures employed in voice and picture recognition, showcasing their development, essential traits, and comparative efficacy. For picture processing, it looks at ResNet, VGG, and EfficientNet; for voice processing, it looks at DeepSpeech, WaveNet, and Transformer-based models. combine convolutional and sequential layers into a hybrid model to improve performance in multimodal applications.
Keywords: Deep Learning, Image Recognition, Speech Recognition, Convolutional Neural Networks (CNNs)

Introduction

Image and speech recognition systems have been greatly enhanced by the advent of deep learning, a revolutionary approach to artificial intelligence. Deep learning models automate the learning of hierarchical representations from raw data, allowing for more accurate and stable performance than classic machine learning approaches that depend on handcrafted features. Many tasks have been greatly enhanced as a result of this shift, including speaker identification, object detection, facial recognition, and speech-to-text translation. Due to their effectiveness in capturing spatial hierarchies in visual data, Convolutional Neural Networks (CNNs) have become the dominating architecture in the domain of image recognition. On massive picture classification and object recognition tasks, models like EfficientNet, ResNet, and VGG have shown outstanding performance. Convolutional layers, pooling mechanisms, and deep feature extraction techniques are used by these architectures to detect intricate patterns in images. As a result, they are well-suited for use in autonomous driving, healthcare imaging, and surveillance systems. Similarly, developments in deep learning have been very beneficial to speech recognition. In the beginning, people used statistical models like Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) together. Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) improved the ability to describe temporal

© The author. Published by ***Revista Infoacceso***, ISSN 2311-7605.

dependencies in sequential data, though. Modern automatic speech recognition (ASR) systems have achieved state-of-the-art performance because to architectures based on Transformers, which capture long-range relationships and enable parallel processing. The development of multimodal systems that integrate visual and auditory data for improved performance has also been spurred by the integration of deep learning architectures across both the image and speech domains. For instance, when picture and speech recognition technologies work together, it improves applications like video captioning, lip-reading, and human-computer interaction. A number of obstacles remain, even with these improvements. When training, deep learning models frequently necessitate copious quantities of labelled data and powerful computing resources. Overfitting, poor interpretability, and data quality sensitivity are some of the issues that might impact the reliability and deployment of models. Furthermore, scalable and efficient architectures that can function with limited resources are required of real-time applications.

Fundamentals of Deep Learning Architectures

Machines can learn complicated patterns and representations from massive amounts of data thanks to deep learning architectures, which are the foundation of contemporary AI systems. Artificial neural networks, which take their cues from how the human brain operates, are the basis of these designs. Deep learning models are great at picture and speech recognition because, by arranging numerous layers of linked neurons, they can gradually extract higher-level information from raw inputs.

An input layer, a hidden layer (or layers), and an output layer are the fundamental building blocks of a neural network, which is crucial to deep learning. Neurons in each layer use activation functions and weighted connections to analyze data. By increasing the number of hidden layers in the network, or "depth," the model is able to learn hierarchical representations, with lower levels capturing more concrete information and higher layers capturing more abstract patterns.

Rectified Linear Unit (ReLU), sigmoid, and tanh activation functions, among others, bring non-linearity to deep learning models and are essential to their architectures. Neural networks are able to learn intricate relationships that linear models are unable to grasp because of this. The computational efficiency and capacity to minimize difficulties such vanishing gradients have led to ReLU's widespread adoption among them.

Backpropagation, the algorithm utilized to train neural networks, is another essential notion. In order to update the weights, backpropagation first determines the discrepancy between the expected and actual outputs, and then propagates this error backwards through the network. Optimization techniques like RMSprop, Adam, or stochastic gradient descent (SGD) iteratively tweak model parameters to minimize the loss function; they are often used in conjunction with this process.

The structural design and data type that deep learning architectures process allow for broad categorization:

- **Feedforward Neural Networks (FNNs):** A basic model, devoid of loops and with data flowing unidirectionally from input to output.
- **Convolutional Neural Networks (CNNs):** Developed for use with geographical data, including photos, and employing convolutional layers to detect hierarchical structures and local patterns.

© The author. Published by ***Revista Infoacceso***, ISSN 2311-7605.

- **Recurrent Neural Networks (RNNs):** Designed to remember past inputs and process sequential data, making it ideal for use with speech and text. Problems with vanishing gradients are addressed by variants such as LSTM and GRU.
- **Transformer Architectures:** Improving performance on vision and language tasks and enabling parallel processing can be achieved by using self-attention methods to represent data's long-range dependencies.

When it comes to enhancing model generalization and avoiding overfitting, regularization approaches are just as important as architecture types. To make sure models work well with unknown data, methods like data augmentation, batch normalization, and dropout are used.

Large datasets and improvements in computer capacity, especially with GPUs and distributed computing, are also crucial to the success of deep learning architectures. Due to these considerations, deep and complicated models can be trained, which would have been computationally impossible before.

Convolutional Neural Networks for Image Recognition

One type of deep learning architecture that was developed with visual data processing and analysis in mind is the Convolutional Neural Network (CNN). Their capacity to automatically learn feature hierarchies from raw images has made them the backbone of contemporary image recognition systems. Complex visual tasks are well-suited to convolutional neural networks (CNNs) because, unlike conventional machine learning approaches, they learn features directly from data.

The convolution operation is fundamental to convolutional neural networks (CNNs). It searches the input image for local patterns like edges, textures, and forms by applying filters, or kernels. By gliding over the picture, these filters create feature maps that draw attention to key details. From basic edges in the first layers to more complicated objects in the later ones, convolutional neural networks (CNNs) acquire progressively abstract representations as the network depth grows.

A typical CNN architecture consists of several key layers:

- **Convolutional Layers:** To get characteristics out of the source picture, these layers use a cascade of filters. Certain patterns, like edges or color gradients, are captured by each filter.
- **Activation Functions:** In order for the model to acquire intricate patterns, non-linear functions such as ReLU are used to introduce non-linearity.
- **Pooling Layers:** Incorporating these layers into the model makes it more resistant to tiny input perturbations and decreases computational complexity by reducing the spatial dimensions of feature maps. Two popular kinds are average pooling and maximum pooling.
- **Fully Connected Layers:** Layers that perform categorization or prediction by combining retrieved features are found towards the network's end.

Parameter sharing, in which the same filter is applied across many regions of the picture, is one of the main benefits of convolutional neural networks (CNNs). Because of this, CNNs are more efficient and scalable than fully linked networks, and the number of parameters is reduced dramatically. Furthermore, convolutional neural networks (CNNs) display translation

© The author. Published by ***Revista Infoacceso***, ISSN 2311-7605.

invariance, which implies that they are able to identify objects independent of their location in the picture.

Over the years, several advanced CNN architectures have been developed to improve performance and efficiency:

- **VGGNet:** Known for its simplicity and use of deep stacked convolutional layers.
- **ResNet (Residual Networks):** Introduces skip connections to address the vanishing gradient problem, enabling the training of very deep networks.
- **EfficientNet:** Focuses on scaling network dimensions (depth, width, and resolution) in a balanced manner to achieve high accuracy with fewer parameters.

CNNs have proven to be highly effective in numerous image recognition tasks, such as object identification, facial recognition, image segmentation, and classification. Autonomous vehicles (AVs) rely on them for tasks like lane and object detection, medical imaging (for tumor detection, for example) and security systems (for biometrics and surveillance).

Despite their achievements, CNNs encounter specific obstacles. They can be computationally demanding, have a high training dataset requirement for labelled data, and might have trouble generalizing to situations with great data variation. Furthermore, essential applications may be concerned due to their "black-box" nature, which might make interpretation difficult.

Thanks to their ability to learn features automatically and in a hierarchical fashion, Convolutional Neural Networks have completely changed the face of image recognition. Their performance, efficiency, and interpretability are being continuously improved through study, and their ability to efficiently record spatial patterns makes them indispensable in modern computer vision systems.

Recurrent Neural Networks and LSTM for Speech Recognition

A subset of deep learning architectures known as Recurrent Neural Networks (RNNs) excels at speech recognition because of its ability to process sequential data. One key difference between RNNs and feedforward neural networks is the presence of feedback connections in RNNs, which enable data to be preserved between time steps. Since the significance of a sound is frequently dependent on sounds that came before it, this allows the model to capture temporal relationships in sequential inputs like audio signals.

Audio streams are usually transformed into acoustic feature sequences (such as spectrograms or Mel-frequency cepstral coefficients) for use in speech recognition. Step-by-step analysis is performed by RNNs using a hidden state that functions as memory. This allows the model to retain context from earlier inputs. To grasp phonology, phonetics, and language structure, this skill is crucial.

Unfortunately, the vanishing gradient problem severely hampers the capacity of conventional RNNs to learn long-term relationships, as gradients shrink to negligible sizes during training. This is an especially big concern in speech recognition since it requires context across longer time periods.

Long Short-Term Memory (LSTM) networks were developed as a more sophisticated version of RNNs to overcome this drawback. Long short-term memories (LSTMs) use a more complex memory architecture with specific parts called gates:

- **Forget Gate:** Determines which information from the previous state should be discarded.

© The author. Published by ***Revista Infoacceso***, ISSN 2311-7605.

- **Input Gate:** Decides which new information should be stored in the memory.
- **Output Gate:** Controls how much of the stored information is used to produce the output.

By using these gating mechanisms, LSTMs are able to overcome the vanishing gradient problem and preserve relevant information across longer sequences while eliminating irrelevant features. Consequently, LSTMs are now the de facto architecture for most sequence modeling and speech recognition applications.

In order to simplify the LSTM architecture and reduce computational complexity while keeping equivalent performance, another often used alternative is the Gated Recurrent Unit (GRU).

The usage of RNNs and LSTMs is common in contemporary speech recognition systems, however they are not the only methods employed. One example is:

- **Connectionist Temporal Classification (CTC):** Enables end-to-end training without requiring precise alignment between input audio and output text.
- **Encoder-Decoder Architectures:** Convert input speech sequences into textual output, often enhanced with attention mechanisms.

Applications like speech-to-text systems, real-time transcription, and voice assistants have all made good use of these models.

Although RNN-based designs have many advantages, they also have a few drawbacks. While modern models like Transformers handle extremely lengthy sequences with ease, these older models are computationally costly, hard to parallelize, and could still fail. Therefore, in order to achieve better efficiency and scalability, attention-based and Transformer designs have recently become the focus of research.

Speech recognition has come a long way thanks to RNNs and LSTMs, which allow for excellent modeling of sequential and temporal data. Although more recent designs are getting all the attention, these models are nevertheless necessary to grasp the background and fundamentals of contemporary voice recognition systems.

Conclusion

Thanks to deep learning architectures, picture and speech recognition have undergone a sea change, with systems now capable of achieving efficiency and accuracy never before seen. These architectures have revolutionized the way machines understand and handle complicated data. For example, Convolutional Neural Networks (CNNs) are great at capturing spatial features in pictures, while RNNs and LSTM networks are great at modeling temporal dependencies in speech. Recognition systems are now better able to manage diverse and large-scale datasets, thanks to advancements in deep learning models such as hybrid techniques and Transformer-based architectures. These developments have proven the practical significance of deep learning technologies with their extensive use in domains including healthcare diagnostics, autonomous systems, voice assistants, and multimedia processing. Several obstacles still need to be overcome, even with these victories. To get the most out of deep learning models, you usually need a lot of processing power, a big labeled dataset, and a lot of tuning. Overfitting, interpretability, and deployment in situations with limited resources are still problems. Furthermore, issues with scalability and energy efficiency are brought up by the increasing complexity of models. By enhancing data economy and enabling real-time

© The author. Published by ***Revista Infoacceso***, ISSN 2311-7605.

processing on constrained hardware, emerging technologies including edge AI, self-supervised learning, and transfer learning are anticipated to tackle numerous of these difficulties in the future. Intelligent systems are expected to undergo additional innovation as multimodal learning becomes more prevalent. This sort of learning combines several data types, including picture, speech, and others. Modern image and speech recognition relies heavily on deep learning architectures, which provide robust resources for deriving insights from intricate datasets. Building more efficient, scalable, and interpretable systems will advance the capabilities of artificial intelligence in real-world applications. Further research and development in this sector is vital for this.

References (APA Style)

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML)*.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine, 29*(6), 82–97.

Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *EMNLP*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

Amodei, D., et al. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. *International Conference on Machine Learning (ICML)*.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

© The author. Published by ***Revista Infoacceso***, ISSN 2311-7605.