International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

THE ROLE OF DEEP LEARNING ARCHITECTURES IN COMPUTER VISION AND NATURAL LANGUAGE PROCESSING: A COMPREHENSIVE REVIEW

Nur Ilman[1*], Dr. Navin Prakash[2], Agravat Shardulkumar J.[3], V. Rama Krishna[4], Dr. Vipin Kumar[5]

[1*]AMIK Ibnu Khaldun Palopo, Indonesia nurilmansyah@gmail.com
[2]Professor, Department of CSE(AI), GNIOT, Greater Noida naveenshran@gmail.com
[3]Assistant Professor, I.T. Department, Shantilal Shah Engineering College, Bhavnagar, Gujarat Technological University, Ahmedabad shardulagravat@gmail.com
[4]Professor, Department of AI & Data Science Koneru Lakshmaiah Education Foundation, Vaddeswaram, India ORCID ID: 0000-0002-0514-1027 vramakrishna@kluniversity.in
[5]Associate Professor, Department of Mathematics, Faculty of Engineering, Teerthanker Mahaveer University drvipink.engineering@tmu.ac.in

Abstract

The invention of powerful neural architectures has transformed the areas of computer vision and natural language processing with deep learning. This review addresses the history and use of the most well-known deep learning models, such as convolutional neural networks, recurrent neural networks, and transformers, in both CV and NLP fields. CNN-based models have proved to be incredibly powerful in image classification, object identification and segmentation in computer vision. Transformer-based models like BERT and GPT have surpassed conventional RNNs in NLP by allowing a scalable context modeling due to the self-attention mechanisms. Along with these achievements, deep learning systems have a number of limitations, namely high computational complexity, sensitivity to data, poor explainability, and susceptibility to bias and adversarial attacks. In this paper, some important benchmark datasets and evaluation measures to compare architectures will be discussed, and a critical evaluation of the existing shortcomings and future research avenues will be provided. Most of these challenges should be addressed in the future with respect to efficient architectures, explainability and multimodal learning. Through the synthesis of recent developments, comparative analysis, and implications, this review gives a thorough picture of how deep learning architectures are still defining the field of computer vision and natural language processing.

Keywords: Deep Learning, Computer Vision, Natural Language Processing, Neural Architectures, Transformers

1. Introduction

Deep learning has emerged as a transformative technology in artificial intelligence and machine learning, enabling major advances in solving complex problems across domains. Its most prominent applications are in Computer Vision and Natural Language Processing, both of which involve high-dimensional, unstructured data and demand sophisticated pattern recognition. The ability of deep learning to learn hierarchical representations directly from raw data allows it to outperform traditional approaches in many real-world tasks [1]. While rooted in earlier

Received: August 04 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

neural network research, recent success stems from large-scale labeled data, GPU acceleration, and deep architectures such as CNNs, RNNs, and transformers [2].
In CV, CNNs have become foundational due to their ability to capture spatial hierarchies, driving progress in object detection, segmentation, and classification [3]. Their variants have proven effective in applications like facial recognition, autonomous driving, and medical imaging. More recently, vision transformers have further enhanced performance through attention mechanisms [4]. In NLP, the shift from rule-based and statistical models to neural architectures such as word embeddings, LSTMs, and GRUs addressed challenges like vanishing gradients and long-range dependencies [5]. Transformer-based models like BERT and GPT brought a new era of pre-trained models achieving state-of-the-art results across various NLP tasks [6].
Multimodal learning has emerged at the intersection of CV and NLP, enabling applications like image captioning, visual question answering, and robotics through integrated visual-linguistic models [7]. Despite these achievements, challenges remain—deep models are data- and compute-intensive, often lack interpretability, and raise ethical concerns in sensitive domains like healthcare and law. This has prompted growing interest in efficient, transparent, and responsible AI methods, including model compression, explainable AI, and federated learning [8]. This review provides a detailed analysis of deep learning architectures, their roles in CV and NLP, ongoing challenges, and emerging research directions.

2. Literature Review

The development and use of deep learning architecture in computer vision and natural language processing have been studied many times. Initial work, e.g. the invention of multilayer perceptrons already developed some of the principles of deep learning although there were problems to deal with, e.g. the vanishing gradient problem and the lack of computational power [9]. Convolutional neural networks were a major breakthrough in CV, and the AlexNet model showed a breakthrough performance on ImageNet in 2012, and increasingly deeper networks such as VGG, ResNet, and DenseNet since showed further gains [10][11].
The initial rule-based and statistical approaches in the NLP field were replaced by neural approaches, when word embeddings such as Word2Vec were developed, allowing meaningful representation of the language in vectors. The recurrent neural networks and its adaptations LSTM and GRU solved the sequence modeling problem more successfully yet were not able to capture long-range dependencies. The Transformer architecture provided self-attention, which could be used to more effectively model sequences in both NLP and CV. Recent models (BERT, GPT, T5) used these mechanisms to pretrain and they have set state-of-the-art performance in many NLP tasks [12]. Cross-modal models that combine text and vision are also the result of this shift, allowing applications like image captioning, visual QA and robotics [13]. These trends reflect the merging of CV and NLP, which has been facilitated by the similarity in interpretability, data efficiency and model robustness issues [14].

3. Deep Learning Architectures

Deep learning architectures form the foundation of modern AI systems, enabling models to learn complex patterns from large datasets. These architectures vary in how they process different input types, such as images, sequences, or multimodal data. The most widely used types include Convolutional Neural Networks, Recurrent Neural Networks, Transformers, and Generative Adversarial Networks. Each offers unique mechanisms and strengths across various

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

applications. Figure 1 presents a simplified schematic illustrating the core structure and information flow of these four architectures.

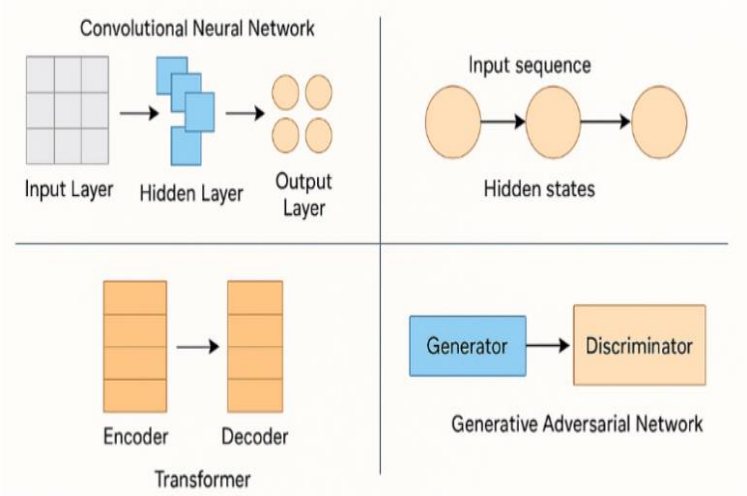

Figure 1: Schematic representation of core deep learning architectures

3.1 Convolutional Neural Networks (CNNs)

CNNs are specialized to process visual data in the form of two-dimensional grids, e.g. images. They employ the use of convolutional filters with hierarchy of layers that capture spatial hierarchies and then are down streamed through the pooling and fully connected layers. The emergence of classic CNN architectures such as AlexNet, VGG, and ResNet changed the discipline of computer vision by boosting the performance of tasks like object detection and classification by huge margins [15]. CNNs have a computational footprint that is efficient and saleable by way of weight sharing and local connectivity [16].

3.2 Recurrent Neural Networks (RNNs)

The RNNs are meant to process sequential data and are largely applied in natural language processing and time-series forecasting. Feedback loops are provided in these networks, enabling them to remember past inputs, thus they can be used where context is important. Nevertheless, classic RNNs have problems with long-term dependencies, so LSTM and GRU structures were created. LSTMs propose gated operations that control the information flow to make the model forget or remember previous data (as required). Such enhancements are especially helpful in translation of languages and speech recognition [17].

3.3 Transformers

Transformers are abandoned in the whole idea of recurrence and are based on self-attention mechanisms. This enables them to process all tokens in sequence to a single pass, resulting in much faster training and more context modeling. Transformers power models like BERT, GPT or T5, are state-of-the-art on most NLP benchmarks [18]. Vision is also adopting transformers with architectures such as ViT that cut images into patches and treat them as tokens, like words. This trend to the attention-based model has reestablished both CV and NLP scenery.

3.4 Generative Adversarial Networks (GANs)

GANs consist of two models that compete against each other, they are a generator that generates new information, and a discriminator that verifies the authenticity of the information. Such a

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

competitive mechanism enables GANs to generate outputs that are highly realistic in such tasks as image generation, super-resolution, and text-to-image translation [19]. Although GANs have brought about the advancements in the generation of synthetic data, they are infamously challenging to train and tend to exhibit such problems as mode collapse and unstable gradients. To make these architectures comparisons more structured, Table 2 presents a brief analysis of their common uses, main strong points, and primary drawbacks.

Table 1. Comparative Summary of Major Deep Learning Architectures

Architecture	Key Use Cases	Advantages	Limitations
CNN	Image classification, object detection	Spatial feature learning, parameter sharing	Fixed input size, limited for sequential data
RNN (LSTM/GRU)	Text generation, language modeling	Temporal memory, good for sequential data	Difficult to train on long sequences, slow inference
Transformer	NLP, Vision (ViT), Multimodal models	Parallel training, global context via self-attention	Resource-intensive, needs large datasets
GAN	Image generation, super-resolution	High-quality synthetic data, creative tasks	Mode collapse, unstable training

4. Applications in Computer Vision

Deep learning has transformed computer vision by superseding the prior hand-engineered pipelines that train end-to-end models. Deep learning architectures and in particular CNNs and their variants have transformed the performance of the image classification task, real-time object detection as well as medical imaging among many others.

4.1 Image Classification

The initial big success of deep learning in vision was an image classification problem. Such CNN architectures as AlexNet, VGGNet, and ResNet managed to deliver a breakthrough performance on the ImageNet dataset which is superior to the traditional feature-engineered models [20]. These nets are trained to learn filters in an automatic, hierarchical manner, i.e. edges in initial layers, parts of objects and semantics in later layers. The idea of residual networks (ResNet) was to add identity shortcut connections in order to compensate the vanishing gradient problem in very deep models.

4.2 Object Detection and Recognition

During detection, models should classify as well as localize an object in an image. Single-shot detectors such as YOLO and SSD, or region-based CNNs (R-CNN, Fast R-CNN, Faster R-CNN) have become de-facto industry standards in terms of speed and accuracy. Specifically, YOLO (You Only Look Once) transforms the problem of detection into a single regression task, which allows making inferences in real-time, which is critical to such applications as autonomous driving and surveillance [21].

4.3 Image Segmentation and Scene Understanding

The classification at pixel level is needed in tasks like segmentation, and encoder-decoder architectures like U-Net and DeepLab have been addressing the problem successfully. U-Net is an original biomedical image segmentations network that use skip connections to combine both spatial and contextual features [22].

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Applications These architectures are common in medical imaging, robotics and environmental monitoring applications where the shape of the object and the accuracy of its boundary are important.

4.4 Model Performance Metrics in CV

To evaluate model performance in classification and detection tasks, common metrics include accuracy, IoU (Intersection over Union), mean Average Precision, and Receiver Operating Characteristic curves. These metrics provide insights into how well a model distinguishes between classes under different thresholds. The ROC curve, for instance, compares the true positive rate to the false positive rate and is especially useful in binary classification problems such as disease diagnosis or pedestrian detection [23]. Figure 2 below shows a typical ROC curve used to assess model behavior in such scenarios.

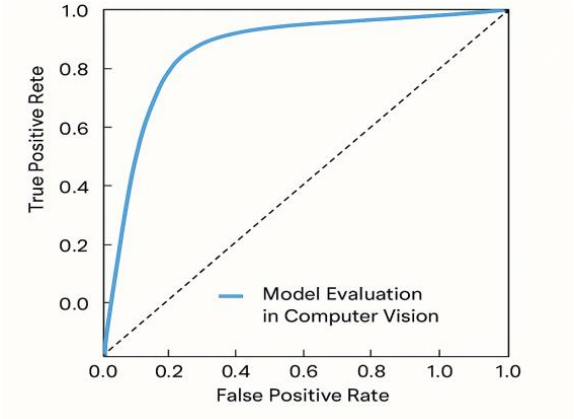

Figure 2: ROC curve showing true vs. false positive rates in binary classification.

4.5 Comparative Training Behavior

Other than precision, it is also worth considering the rate at which models converge and the stability of the convergence throughout training. Tracking training loss across epochs assists in identifying the effectiveness of a model or overfitting. In the following, Figure 3 illustrates training loss curves of three distinct architecture types, CNN (standard in CV), LSTM (sequential) and Word2Vec (embedding), to highlight the difference in learning behavior and convergence rate. Such patterns are particularly applicable to cases when the CV systems deal with sequential data (e.g. video frames or scene descriptions).

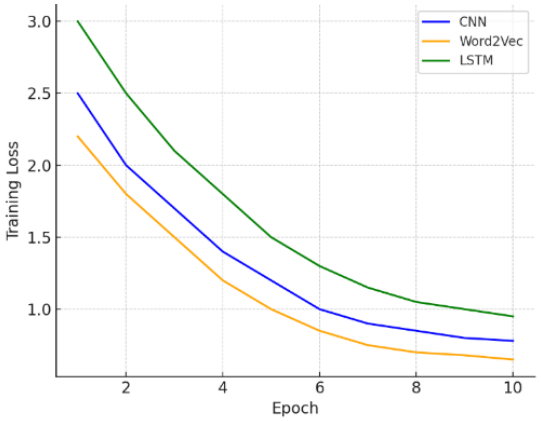

Figure 3: Training loss comparison for CNN, Word2Vec, and LSTM over 10 epochs.

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.6 Popular Datasets in Computer Vision

Table 3 shows the overview of popular databases in CV which are used in the benchmarking and training of deep learning models:

Table 2. Commonly Used Computer Vision Datasets

Dataset	Domain	Description	Typical Use
ImageNet	Object Classification	1,000-class object dataset with 1.2M images	Classification, Transfer Learning
COCO	Object Detection	Large-scale detection and captioning dataset	Detection, Segmentation, Captioning
Pascal VOC	General Vision Tasks	Annotated images for object classes and actions	Multi-task vision
Cityscapes	Urban Scene Segmentation	Annotated street scenes	Autonomous Driving, Scene Parsing
LUNA	Medical Imaging	CT scans for lung cancer screening	Biomedical detection/segmentation

5. Applications in Natural Language Processing

NLP is one of the most impactful domains transformed by deep learning. NLP used to be dominated by rule-based algorithms and statistical models, but now has the advantage of neural architecture with the ability to capture dependencies, contextual understanding, and produce meaningful language. This section discusses the applications of various deep learning models to key NLP tasks, including such architectures as CNNs, RNNs, and Transformers.

5.1 Text Classification and Sentiment Analysis

Text classification tasks such as spam filtering, topic detection, and sentiment analysis were among the first to benefit from deep learning. Convolutional Neural Networks are able to capture local n-gram features and are considered ideal in short text classification tasks whereas RNN particularly when using LSTM are able to capture long-term dependencies in word sequences. One of such applications is movie review sentiments classification, topic classification (AG News), and social media. CNNs are more accurate at spotting important phrases, whereas LSTMs have a more accurate picture of the order of words and syntax [24].

5.2 Machine Translation and Sequence Modeling

Sequence-to-sequence models (Seq2Seq) were a major leap in machine translation. Using encoder-decoder networks built with RNNs or GRUs, these models encode an input sentence and then decode it into a target language. However, they struggled with long-range dependencies and parallel computation. This issue was overcome by the introduction of the Transformer architecture which substituted recurrence with self-attention. Transformers handle entire sequences at once and have obtained state of the art performance in multilingual translation systems like Google Translate and M2M-100 [25].

Received: August 08 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5.3 Pre-trained Language Models and Transfer Learning

The appearance of pre-trained language models introduced a paradigm shift in NLP. These models are pretrained on huge corpora and are fine-tuned on downstream tasks, and hence perform better with less task-specific data. BERT (Bidirectional Encoder Representations from Transformers) is pre-trained with masked language modeling and therefore can be effective in any task that demands deep contextual representation like question answering and named entity recognition. Unlike that, GPT-2 and GPT-3 employ an autoregressive decoding structure that produces fluent, conversational text and are quite adept at creative or generative tasks. T5 (Text-to-Text Transfer Transformer) resorts to unification of NLP tasks by transforming all tasks into text-to-text format [26].

5.4 Attention Mechanisms and Interpretability

Attention mechanisms are the backbone of the transformer revolution. They allow models to evaluate the significance of every word in a sentence compared to other words and hence produce dynamic contextual representations. Such attention weights can be plotted to comprehend model behavior. The following heatmap (Figure 4) shows the attention scores in BERT for the sentence *"The quick brown fox jumps over the lazy dog."* Each cell shows how much attention a word gives to another. Darker squares represent higher attention values, indicating stronger relevance between words.

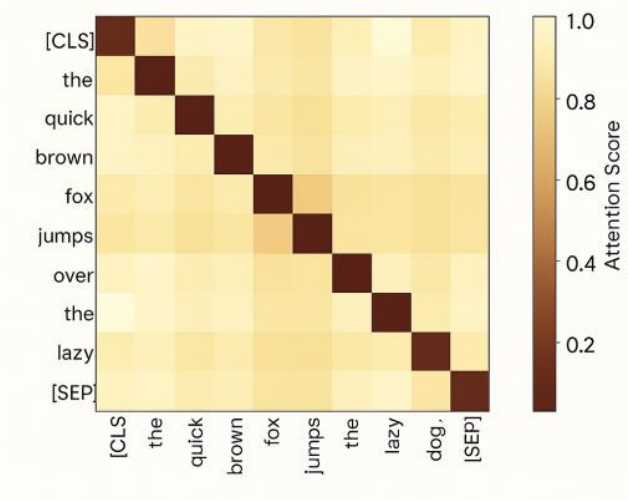

Figure 4: BERT attention heatmap showing token-to-token attention weights for a sample sentence.

This explainability is particularly useful in such high-stakes contexts as legal document interpretation or medical question answering where it is as important to know why the model made a call as the call itself.

5.5 Challenges in Deep NLP

Despite its remarkable success, deep NLP faces several limitations. Bias is one of the biggest challenges to learn and reproduce societal stereotypes using datasets of large sizes on the internet. The other is hallucination; in which models produce plausible (but not necessarily true) results. The models such as GPT-3 are quite efficient and their use demands massive computation powers, which makes them inaccessible to many and the models create environmental issues. Also, a majority of the pre-trained models are English biased which

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

creates difficulty in cross-lingual applications. Efforts like XLM-R (Cross-lingual RoBERTa) aim to address this by training models on multilingual corpora, enabling better performance in low-resource languages.

5.6 Overview of Leading Language Models

The current NLP environment has seen the optimization of several models to particular tasks and architectural designs. They differ in their characteristics and strengths in terms of their processing of the context, production of outputs, and cross-language/domain adaptation. To summarize the differences, a comparative overview is shown in the Table 4 below:

Table 3. Overview of Deep Learning-Based Language Models

Model	Architecture	Core Strengths	Typical Tasks
BERT	Transformer (Encoder)	Deep bidirectional context	Classification, QA, NER
GPT-2/3	Transformer (Decoder)	Coherent text generation	Storytelling, Completion, Dialogue
T5	Encoder-Decoder	Unified text-to-text framework	Translation, QA, Summarization
RoBERTa	BERT variant	Optimized pre-training, robust features	Sentence classification, QA
XLM-R	Multilingual BERT	Cross-lingual understanding	Translation, Multilingual NLP

The models have become the default backbone of nearly all significant NLP tasks, including summarization and automated reasoning. Their elasticity, scalability, and adaptability have brought a new age of smart language interfaces. As we continue, we'll explore how deep learning architectures, initially developed for vision or language, are now converging in multimodal systems that can reason across text, images, and even audio.

6. Benchmark Datasets and Evaluation Metrics

During the comparative analysis of deep learning architectures, benchmark datasets and standardized evaluation measures are required. They provide a controlled setting with which to quantify model performance to achieve reproducibility between studies. Both in CV and NLP, the performance of a model is mostly measured by its performance on popular benchmarks with pre-defined evaluation metrics [27].

6.1 Computer Vision Benchmarks

ImageNet dataset has been critical in the development of architecture in computer vision. It contains more than one million labeled images and 1,000 categories and has been traditionally used as the main benchmark of image classification tasks [28]. Models like AlexNet, VGGNet, ResNet, and ViT were first tested on this dataset based on top-1 and top-5 accuracy. In object detection and captioning, the COCO (Common Objects in Context) dataset is commonly used. It also contains complicated visual scenes and can be evaluated with mean Average Precision (mAP), BLEU, and CIDEr measurements. Additional well-known datasets are Pascal VOC (object detection/segmentation, mAP, IoU), Cityscapes (scene parsing, mean IoU, pixel accuracy), and LUNA (medical image analysis, AUC, sensitivity). The following table 5 presents the summary of these main benchmarks and their respective assessment strategies.

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 4. Computer Vision Benchmarks and Metrics

Dataset	Task	Metrics	Common Models
ImageNet	Classification	Top-1, Top-5 Accuracy	AlexNet, ResNet
COCO	Detection, Captioning	mAP, BLEU, CIDEr	YOLO, Faster R-CNN
Pascal VOC	Detection, Segmentation	mAP, IoU	SSD, FCN
Cityscapes	Scene Parsing	Pixel Accuracy, mean IoU	DeepLab
LUNA	Medical Imaging	AUC, Sensitivity	U-Net

6.2 NLP Benchmarks

In natural language processing, general-purpose language understanding is often measured using the GLUE and SuperGLUE benchmarks. GLUE measures performance on a number of tasks; sentiment analysis, paraphrase detection and natural language inference, in terms of accuracy and F1 score. SuperGLUE is an extension of a greater difficulty and has tasks that involve more reasoning, and is measured in the same way [29]. Question answering tasks as in the SQuAD dataset are evaluated with the help of Exact Match and F1 scores. The datasets used in language modeling tasks are, e.g., WikiText and Common Crawl, with the common metrics being perplexity, BLEU, and ROUGE [30]. Table 6 presents these datasets and metrics in a more convenient form.

Table 5. NLP Benchmarks and Metrics

Dataset	Task	Metrics	Common Models
GLUE	Multi-task NLU	Accuracy, F1	BERT, RoBERTa
SuperGLUE	Reasoning, Question Answering	Accuracy, Exact Match (EM)	T5, DeBERTa
SQuAD	Question Answering	Exact Match (EM), F1	ALBERT, XLNet
WikiText	Language Modeling	Perplexity	GPT, LSTM
Common Crawl	Pretraining (LM)	BLEU, ROUGE	GPT-2, GPT-3

These benchmarks serve as standardized yardsticks for comparing the capabilities of different architectures and continue to evolve in response to new model developments and task demands.

Discussion

Computer vision and natural language processing have been ameliorated by deep learning to great extents, and modalities-specific architectures, such as CNNs, RNNs, and Transformers, can provide modality-specific advantages. The spatial feature identification combined with scalability of CNNs has made them effective in image tasks whereas Transformers have emerged because RNNs and LSTMS did not work well on capturing long-range dependencies on sequentially ordered data [31]. Transformers derived on NLP and including BERT, GPT, and ViT are originally designed to use attention mechanisms to capture global context in text, as well as vision recognition tasks [32]. CNNs are not suitable enough when it comes to long-range dependencies, even though they appear to be computationally effective and perform well in spatial learning. RNNs do temporal modelling but are susceptible to vanishing gradients and are therefore less applicable in deep contextual tasks. Conversely, Transformers do provide better global contextual modeling and scaling as well but at the cost of efficiency [33].

Received: August 08 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Transformers are successful, but due to high data and computational requirements, they are not very accessible in low-resource settings. Interpretability is minimal as well; there is limited data in attention maps of model choices, which is of vital concern in healthcare, finance, and law. Models are also susceptible to adversarial inputs which have semantic meaning but confuse predictions [34]. Ethical issues carry on, particularly those where discriminatory training data can enforce inequity in versatile or demographic delicate settings [35]. But there are still gaps on creating models adapted to low-resource languages, domain-specific CV tasks (agriculture, remote sensing), and powerful multilingual NLP systems. Future work should prioritize efficient models through compression, pruning, and distillation for edge deployment, alongside self-supervised, few-shot, and zero-shot learning to reduce dependence on labeled data. Developing explainable AI (XAI) is equally vital for transparency and accountability. As deep learning evolves, addressing its limitations in scalability, fairness, and interpretability will be key to responsible and impactful deployment.

Conclusion

Deep learning has been a paradigm-shifting force in computer vision and natural language processing because of the invention of the successful architectures such as CNNs, RNNs, and Transformers. The models have been shown to be exceptional feature extractors, sequence modellers and contextual knowledge in a wide range of tasks. CNNs still dominate spatial data processing, and RNNs have performed well in traditional sequence modeling, but transformer-based model has set new performance and end-to-end standards in both fields. However, along with these achievements, issues are still present, including high data and computing demands, low interpretability, and vulnerability to bias and adversarial examples. The higher requirement of effective, transparent, and ethically accountable AI has led to new trends in self-supervised learning, multimodal systems, model compression, and explainability. This summary demonstrates how architectural progress contributed to the development of CV and NLP with the help of benchmarks, and evaluation metrics. The discipline will in the future be forced to balance between effectiveness and accessibility, power, and social impact. As the deep learning continues its evolution, its utilization in the real-life decision-making will be scaled up and will demand additional research in the field of modeling as well as in the field of fairness, accountability, and sustainability.

References

[1] A. Sarraf, M. Azhdari, and S. Sarraf, "A comprehensive review of deep learning architectures for computer vision applications," *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, vol. 77, no. 1, pp. 1–29, 2021.

[2] J. Bharadiya, "A comprehensive survey of deep learning techniques natural language processing," *European Journal of Technology*, vol. 7, no. 1, pp. 58–66, 2023.

[3] A. Bouraoui, S. Jamoussi, and A. B. Hamadou, "A comprehensive review of deep learning for natural language processing," *International Journal of Data Mining, Modelling and Management*, vol. 14, no. 2, pp. 149–182, 2022.

[4] J. Guo, H. He, T. He, L. Lausen, M. Li, H. Lin, A. L. Maas, A. Smola, L. Song, X. Wang, and Y. Zhu, "GluonCV and GluonNLP: Deep learning in computer vision and natural language processing," *Journal of Machine Learning Research*, vol. 21, no. 23, pp. 1–7, 2020.

Received: August 08 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[5] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, pp. 1–20, 2021.
[6] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," *arXiv preprint arXiv:2003.01200*, 2020.
[7] I. D. Mienye and T. G. Swart, "A comprehensive review of deep learning: Architectures, recent advances, and applications," *Information*, vol. 15, no. 12, p. 755, 2024.
[8] J. Chai, H. Zeng, A. Li, and E. W. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Machine Learning with Applications*, vol. 6, p. 100134, 2021.
[9] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer*, vol. 38, no. 8, pp. 2939–2970, 2022.
[10] R. K. Mishra, G. S. Reddy, and H. Pathak, "The understanding of deep learning: A comprehensive review," *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 5548884, 2021.
[11] V. A. Devi and M. Naved, "Dive in deep learning: Computer vision, natural language processing, and signal processing," in *Machine Learning in Signal Processing*, Chapman and Hall/CRC, pp. 97–126, 2021.
[12] N. Manakitsa, G. S. Maraslidis, L. Moysis, and G. F. Fragulis, "A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision," *Technologies*, vol. 12, no. 2, p. 15, 2024.
[13] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Deep learning for natural language processing in radiology—fundamentals and a systematic review," *Journal of the American College of Radiology*, vol. 17, no. 5, pp. 639–648, 2020.
[14] A. Upadhyay, N. S. Chandel, K. P. Singh, S. K. Chakraborty, B. M. Nandede, M. Kumar, A. Alam, A. Elbeltagi, and A. S. Alghamdi, "Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models, and trends in precision agriculture," *Artificial Intelligence Review*, vol. 58, no. 3, p. 92, 2025.
[15] M. Gheisari, F. Ebrahimzadeh, M. Rahimi, M. Moazzamigodarzi, Y. Liu, P. K. Dutta Pramanik, R. Ranjan, and S. Kosari, "Deep learning: Applications, architectures, models, tools, and frameworks: A comprehensive survey," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 3, pp. 581–606, 2023.
[16] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604–624, 2020.
[17] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.
[18] R. M. Samant, M. R. Bachute, S. Gite, and K. Kotecha, "Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions," *IEEE Access*, vol. 10, pp. 17078–17097, 2022.
[19] O. Sen, M. Fuad, M. N. Islam, J. Rabbi, M. K. Hasan, A. A. Fime, M. M. Rahman, and M. A. Iftee, "Bangla natural language processing: A comprehensive review of classical machine learning and deep learning based methods," *arXiv preprint arXiv:2102.10402*, 2021.
[20] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, N. Tanwar, P. R. Manogaran, and H. Ghayvat, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, 2021.

Received: August 08 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[21] M. M. Taye, "Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions," *Computers*, vol. 12, no. 5, p. 91, 2023.
[22] N. Le, V. S. Rathour, K. Yamazaki, K. Luu, and M. Savvides, "Deep reinforcement learning in computer vision: A comprehensive survey," *Artificial Intelligence Review*, vol. 55, no. 4, pp. 2733–2819, 2022.
[23] A. Younesi, M. Ansari, M. Fazli, A. Ejlali, M. Shafique, and J. Henkel, "A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends," *IEEE Access*, vol. 12, pp. 41180–41218, 2024.
[24] G. Rafiq, M. Rafiq, and G. S. Choi, "Video description: A comprehensive survey of deep learning approaches," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13293–13372, 2023.
[25] M. Memari, "Advances in computer vision and image processing for pattern recognition: a comprehensive review," *International Journal of Engineering and Applied Sciences*, vol. 11, no. 05, pp. 2896–2901, 2023.
[26] Y. A. Christobel and R. J. Suji, "A comprehensive review on neural network architectures," *Journal of Computational Analysis & Applications*, vol. 33, no. 7, 2024.
[27] L. Aziz, M. S. B. H. Salam, U. U. Sheikh, and S. Ayub, "Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review," *IEEE Access*, vol. 8, pp. 170461–170495, 2020.
[28] B. Chinnaiyan, S. Balasubaramanian, M. Jeyabalu, and G. S. Warrier, "AI applications–Computer vision and natural language processing," in *Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications*, pp. 25–41, 2025.
[29] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.
[30] X. Liu, Y. Su, and B. Xu, "The application of graph neural network in natural language processing and computer vision," in *Proc. 2021 3rd Int. Conf. Machine Learning, Big Data and Business Intelligence (MLBDBI)*, Dec. 2021, pp. 708–714. IEEE.
[31] S. Singh, S. Kumar, and B. K. Tripathi, "A comprehensive analysis of quaternion deep neural networks: architectures, applications, challenges, and future scope," *Archives of Computational Methods in Engineering*, pp. 1–28, 2024.
[32] M. J. Basha, S. Vijayakumar, J. Jayashankari, A. H. Alawadi, and P. Durdona, "Advancements in natural language processing for text understanding," in *E3S Web of Conferences*, vol. 399, p. 04031, 2023. EDP Sciences.
[33] A. J. Obaid, B. Bhushan, and S. S. Rajest, Eds., *Advanced Applications of Generative AI and Natural Language Processing Models*. IGI Global, 2023.
[34] J. Liu, K. Li, A. Zhu, B. Hong, P. Zhao, S. Dai, ... and H. Su, "Application of deep learning-based natural language processing in multilingual sentiment analysis," *Mediterranean Journal of Basic and Applied Sciences (MJBAS)*, vol. 8, no. 2, pp. 243–260, 2024.
[35] S. Balasubramaniam, V. Chirchi, S. Kadry, M. Agoramoorthy, S. P. Gururama, K. K. Satheesh, and T. A. Sivakumar, "The road ahead: emerging trends, unresolved issues, and concluding remarks in generative AI—a comprehensive review," *International Journal of Intelligent Systems*, 2024.

Received: August 08 2025