GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Review

Application Exploration of Machine Learning in Natural Language Processing and Computer Vision

Peiheng Qin [1,*]

[1] School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, Australia
* Correspondence: Peiheng Qin, School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, Australia

Abstract: Machine learning in general, and deep learning in particular, has revolutionised both natural language processing (NLP) and computer vision (CV) over the last decade. Performance on benchmark tasks consistently exceeds previous state of the art, and in many cases approaches or exceeds human-level accuracy. This paper provides a structured exploration of the main applications of machine learning in these two domains, exploring the architectural innovations -- from convolutional neural networks to the Transformer -- that have propelled progress, and analysing representative applications including text classification, machine translation, large language models, image classification, object detection, and generative visual modelling. The paper also reviews the convergence of NLP and CV through multimodal architectures such as CLIP and BLIP-2 that have resulted in cross-modal reasoning and new application domains. The main obstacles are noted as computing expense, data bias and restrictions in interpretability and the paper discusses future research paths focusing on efficient models and multimodal reasoning. We seek to give a thorough comparative analysis of ML applications in the two domains, and to discover shared architectural paths that indicate a similar future for the research on intelligent systems.

Keywords: machine learning; natural language processing; computer vision; deep learning; transformer; convolutional neural network; multimodal learning

1. Introduction

1.1. Research Background

The use of machine learning methods to natural language processing and computer vision has been one of the most important technological advances of the early twenty-first century. Before the deep learning era, the two fields were mostly reliant on hand-crafted features and rule-based systems. In particular, NLP was based on syntactic parsers, lexical databases and statistical n-gram models. CV was based on engineering descriptors such as SIFT and HOG for feature extraction. The advent of deep neural networks has brought about a fundamental change in both professions, where feature engineering has been replaced with end-to-end learning representations that generalise better across tasks and domains. Deep learning is part of a broader family of machine learning methods based on learning representations of data, rather than task-specific algorithms. Learning may be supervised, semi-supervised or unsupervised. As LeCun, Bengio and Hinton wrote in their landmark review in 2015, deep learning "allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction." The technique has dramatically improved the state-of-the-art in speech recognition, visual object recognition and object detection simultaneously.

The barrier between NLP and CV has been blurred since the introduction of the Transformer architecture in 2017 and the advent of huge pre-trained models such as BERT, GPT and Vision Transformer (ViT). In recent years multimodal architectures have been developed that work on text and images in joint representation spaces, allowing

Received: 19 April 2026
Revised: 28 May 2026
Accepted: 08 June 2026
Published: 11 June 2026

applications that neither domain could solve alone. This convergence allows comparative analysis of ML applications in both disciplines timely and academically productive.

1.2. Research Scope and Contributions

In this study, we focus on the period from 2012 to 2024, covering the deep learning revolution from AlexNet to today's huge multimodal models. Our contributions are twofold: (i) we offer a systematic survey of representative ML applications in NLP and CV, along with benchmark performance data from well-known public evaluation frameworks; (ii) we provide a comparative analysis of architectural trends in both domains, pointing to the Transformer as a common structure and multimodal integration as the emerging frontier.

2. Machine Learning Foundations for NLP and CV

2.1. Core Learning Paradigms

Machine learning is a class of computational methodologies in which models learn representations and decision functions from data rather than explicit rules. Three paradigms have been notably fruitful in NLP and CV alike. Supervised learning, which optimises model parameters using labelled training data and a specified target, is still the dominating paradigm for task-specific applications. Self-supervised learning -- in which models learn from the intrinsic structure of unlabelled data, such as predicting masked words in text or reconstructing corrupted image patches -- has formed the basis for huge pre-trained models in both domains [1]. With the advent of reinforcement learning from human feedback (RLHF), a technique used to align the output of language models with human preferences, instruction-tuned models have grown in popularity: Ouyang et al. revealed that human assessors favoured InstructGPT models trained with RLHF with 1.3 billion parameters over the base 175-billion-parameter GPT-3, implying that alignment training can replace raw scale [2].

2.2. The Transformer Architecture

Vaswani et al. proposed the Transformer in "Attention Is All You Need", which replaced the recurring operation with a self-attention mechanism that evaluates the relationships between all points in the sequence at once [3]. This architectural innovation brought two decisive advantages: it eliminated the sequential Bottleneck of recurrent neural networks, allowing efficient parallel training on large datasets, and it introduced a flexible inductive bias, suitable for both discrete token sequences (text) and continuous patch sequences (images). Directly stemming from the scalability of the Transformer, the pre-training and fine-tuning paradigm was established, where a model is pre-trained with large-scale unlabelled data and then fine-tuned for specific downstream tasks, becoming the dominant framework in both NLP and CV by 2020 [1].

2.3. Convolutional Neural Networks in Computer Vision

In computer vision, the predominant architecture was convolutional neural networks (CNNs) before the arrival of Vision Transformer. LeCun, Bengio and Hinton laid the foundations of deep learning by showing that hierarchical trained representations may capture complicated structure in high-dimensional data across many domains [4]. The convolutional process embodies spatial locality and translation equivariance inductive biases that are well aligned with natural image statistics, enabling the fast learning of hierarchical visual characteristics from raw pixel inputs. The architectural lineage from AlexNet (2012) to VGGNet (2014) and ResNet (2015) is a development where increasing depth and residual connections continuously improved performance on large-scale picture benchmarks [5,6]. Vision Transformer and its offspring have outperformed CNN-based models on various benchmarks since then, but CNNs still have practical advantages in low-data regimes and resource-constrained deployment scenarios and hybrid architectures with a combination of convolutional and attention layers are still competitive [6].

3. ML Applications in Natural Language Processing

3.1. Text Classification and Sentiment Analysis

Text classification is a vast set of tasks that includes topic categorisation, spam detection and sentiment analysis. Devlin et al. presented BERT in 2018, establishing a new standard of classification performance: fine-tuned BERT achieved state-of-the-art performance on the GLUE benchmark, a collection of nine NLP classification and inference tasks, with an overall score of 80.4 at release, 7.6 points higher than the previous best [1]. Following this, models such as RoBERTa, ALBERT, and DeBERTa have continually advanced GLUE scores, with DeBERTa-v3 attaining 91.9 in 2023, exceeding the human benchmark of 87.1 on the same suite [1]. In real-world applications, pre-trained Transformer encoder-based models for sentiment analysis have become the de facto components of pipelines for financial news, customer feedback, and social media monitoring systems. Financial news corpora have been used to train BERT-based models that have shown that sentiment scores derived from news headlines contain predictive information on short-term movements in equity prices, with F1 scores for positive/negative classification above 0.88 on the Financial PhraseBank dataset [1].

3.2. Machine Translation

Neural machine translation (NMT) based on the Transformer architecture has become the industry standard, displacing phrase-based statistical systems on all major commercial translation platforms. The pace of advancement is best illustrated by the results on the WMT translation benchmarks. On the WMT14 English-to-German test, the original Transformer achieved a BLEU score of 28.4, beating the top convolutional sequence-to-sequence model by more than 2 BLEU points [3]. By 2022 the best models on the same benchmark achieved 35 BLEU. The later switch to other metrics, like as COMET and BLEURT, has been a reflection of increased awareness of BLEU's failure to represent translation quality at high performance levels. When evaluated on WMT22 shared task evaluations, large language models fine-tuned for translation have shown competitive performance with specialist neural translation systems across several language pairs [7].

3.3. Large Language Models and Scaling Laws

The most significant trend in NLP is the creation of autoregressive big language models based on the GPT architecture. GPT-3 showed that increasing the scale of models and data according to predictable power laws leads to predictable and transferable improvements on a wide range of few-shot NLP tasks, with 175 billion parameters [7]. Kaplan et al. formalised these correlations as scaling laws, empirically demonstrating that language model loss scales as a power law with model size, dataset size and compute budget, with some trends spanning over seven orders of magnitude [8]. In March 2023, GPT-4 was reported to have scored at or above the 90th percentile of human test takers on the Uniform Bar Examination (298/400) and on components of the medical license examinations, providing external confirmation of LLM competence against standardised human benchmarks [9]. RLHF-based instruction tweaking (as used in InstructGPT and later models) is now a typical component of LLM development processes [2].

4. ML Applications in Computer Vision

4.1. Image Classification

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is the de facto benchmark for progress in picture categorisation. Deep CNNs became the preferred strategy and the deep learning era in CV was launched when AlexNet achieved a top-5 error rate of 15.3% in 2012, about 10 percentage points below the previous best [5]. ResNet achieved a top-5 error of 3.57% which is below the observed human error rate of around 5% on the same test [6]. He et al. ascribed this performance to residual connections that pass gradients straight through identity mappings and train networks with over 100 layers without deterioration [6]. In 2020, Dosovitskiy et al. proposed the Vision Transformer (ViT) and showed that pure Transformer architectures pre-trained on

sufficiently large datasets can perform on par or better than CNNs: ViT-H/14 pre-trained on JFT-300M reached 88.55% top-1 accuracy on ImageNet, outperforming the best CNN-based models at similar scale [10].

4.2. Object Detection and Instance Segmentation

Object detection extends image classification to locate and classify many objects in a scene. YOLO (You Only Look Once) detectors, which perform image processing in a single forward pass as opposed to region proposal pipelines, have become the dominant framework for real-time detection applications. In January 2023, Ultralytics released YOLOv8, which achieves 53.9 mean Average Precision (mAP) on the COCO benchmark at 80 frames per second on standard hardware. Mask R-CNN expanded Faster R-CNN with a parallel mask prediction branch for instance segmentation, attaining 37.1 mask AP on COCO and it has become a standard architecture widely used in medical image analysis and autonomous driving pipelines [11]. In medical imaging, detection and segmentation models have been applied for polyp detection in colonoscopy video, pulmonary nodule detection in CT scans and retinal lesion analysis in fundus photography, with several systems achieving diagnostic sensitivity comparable to specialist clinicians (source: Ultralytics YOLOv8 [Version 8.0.0, 2023]; Mask R-CNN [11]).

4.3. Generative Visual Models

In CV, generative modelling has witnessed two consecutive paradigm revolutions. Generative Adversarial Networks (GANs) were proposed by Goodfellow et al. in 2014, and they allowed for high-fidelity picture synthesis via an adversarial training objective, and were the main generative framework roughly until 2021 [12]. Denoising Diffusion Probabilistic Models (DDPMs), defined by Ho et al. in 2020, was later able to outperform GANs in conventional picture quality assessments [13]. Both diffusion-based models, DALL-E 2 (2022) and Stable Diffusion (2022), achieved significantly lower Frechet Inception Distance scores than top GAN baselines on standard evaluation sets. The open-sourcing of Stable Diffusion in August 2022 has paved the way for wide research and commercial usage [13]. The rapid rise of generative visual models has enabled new creative uses in design, advertising and entertainment, while also precipitating the copyright issues that are now the topic of active legal scholarship [12].

5. Multimodal Convergence: NLP and CV Integration

5.1. Vision-Language Models

The horizon of contemporary ML research is the convergence of NLP and CV in unified multimodal architectures. CLIP (Contrastive Language-Image Pre-training) was proposed by Radford et al. in 2021 and learned a dual-encoder architecture using 400 million image-text pairs, learning a common embedding space where semantically related images and captions are close to each other [14]. On zero-shot ImageNet classification, CLIP attained 76.2% top-1 accuracy with no task-specific training data, rivalling a fully supervised ResNet-50 and showing that large-scale contrastive pre-training can replace task-specific supervision for a wide array of visual understanding tasks [14]. Inspired by architectures such as BLIP-2 and LLaVA, an extension of the vision-language paradigm has been to connect frozen image encoders to large language models with lightweight adapter layers to enable instruction-following, visual question answering, and image captioning in a unified generative framework [14,15].

5.2. Cross-Domain Applications and Challenges

Multimodal models have been used in medical imaging, and automatic radiological picture report generation has demonstrated performance comparable to radiologist-generated reports on structured assessment metrics. Vision-Language Foundation based Video Understanding Systems have shown promising performance on benchmarks such as ActivityNet-QA and MSVD-QA, enabling applications for content moderation, sports analytics, and accessible captioning [15]. Despite these advances, multimodal models still suffer from systematic failures on compositional reasoning tasks that demand precise

spatial understanding or object counting, indicating that current architectures have not yet fully bridged the semantic gap between continuous visual features and discrete linguistic concepts [14]. Vision-language model training is scaled up to require hardware resources that most research groups do not have access to. This raises problems regarding equitable distribution of research capacity across the field.

6. Discussion and Conclusion

6.1. Comparative Analysis and Common Trajectories

This paper's survey shows a consistent architectural convergence across NLP and CV: both fields have evolved from domain-specific architectures to the Transformer as a unified foundation, and from task-specific training to large-scale pre-training followed by task adaptation. This convergence is a testament to the strength of self-attention as a general-purpose relational inductive bias, and the empirical regularities of scaling laws, which anticipate steady advances in capability with increasing model size and training data. The next step in this trajectory is the development of multimodal architectures: If Transformers are a universal architecture and pre-training a universal learning approach, then integrating different modalities within a single framework is a natural next step.

6.2. Challenges and Future Directions

Computational sustainability is a major challenge for ML applications in both NLP and CV, as is data quality and bias which propagates into model behaviour with well-documented consequences for fairness and reliability. Interpretability remains limited, despite progress in attention visualisation and mechanistic interpretability research. Promising directions for future research include parameter-efficient training methods that adapt large pre-trained models to new tasks with minimal computational overhead, few-shot and zero-shot generalisation to novel domains and languages, and evaluation frameworks that more thoroughly assess model reasoning, robustness, and alignment with human values.

References

1. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*, Association for Computational Linguistics, Minneapolis, 2019, pp. 4171–4186.
2. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, NeurIPS, New Orleans, 2022, pp. 27730–27744.
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, NeurIPS, Long Beach, 2017, pp. 5998–6008.
4. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
5. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, NeurIPS, Lake Tahoe, 2012, pp. 1097–1105.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE CVPR 2016*, IEEE, Las Vegas, 2016, pp. 770–778.
7. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, NeurIPS, 2020, pp. 1877–1901.
8. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint*, arXiv:2001.08361, 2020.
9. OpenAI, "GPT-4 technical report," *arXiv preprint*, arXiv:2303.08774, 2023.
10. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of ICLR 2021*, OpenReview, 2021.
11. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

12. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, NeurIPS, Montreal, 2014, pp. 2672–2680.
13. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, NeurIPS, 2020, pp. 6840–6851.
14. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of ICML 2021*, vol. 139, PMLR, 2021, pp. 8748–8763.
15. J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of ICML 2023*, vol. 202, PMLR, 2023, pp. 19730–19742.