Response to LeCun et al., (1989)

Matthew Evans

February 19, 2025

Approach

In *Backpropagation Applied to Handwritten Zip Code Recognition* [1], LeCun et al. propose a novel machine learning architecture for recognizing digitized hand written numbers, specifically USPS Zip codes. Their design intentionally constrains the system, forcing it to recognize patterns that are likely to occur in multiple parts of a digit (e.g., lines and curves). This strategic constraint enables building a system that requires no manually tuning and which can be actualized (i.e., trained) faster than previous, more complex systems.

Predecessors

Prior systems (described in [2]) for digit recognition relied heavily on hand-designed features and manually tuned constants in their first few layers, requiring significant human effort in preprocessing and feature extraction before the neural network could process the data. Though these systems leverage convolution, not all of their parameters were systematically learned.

Novelty

“Unlike previous results reported by our group on this problem (Denker et all 1998), the learning network is directly fed with images, rather than feature vectors, thus demonstrating the ability of backpropagation networks to deal with large amounts of low-level information.”

Input and Output

Unlike prior systems, which required extensive manual preprocessing and feature selection, the authors’ system requires minimal preprocessing: locating the zip code on the envelope, separating digits, removing noise, and transforming the image to a standardized 16x16 pixel format. From this point, the entire recognition process is handled by a constrained multilayer network with adaptive connections trained through backpropagation, eliminating the need for hand-chosen constants in the early layers.

Feature Maps and Weight Sharing

The authors introduce two key architectural constraints: feature maps and weight sharing.

Feature maps identify and define specific features (e.g., lines, curves, etc.) occurring throughout the input image. They achieve this by convolving **receptive fields** across the input image; this can be thought of as moving a “sliding window” across the rows of pixels in the input image. As the window moves, it identifies similar patterns occurring in different locations throughout the input image. For a given feature map (of which there are many), all of the receptive fields (i.e., each position of the sliding window) are *forced* to learn the same shared set of weights. Consequently, the feature map learns a *single* feature, indifferent to that feature’s location within the input image.

These constraints serve multiple purposes:

- The network is robust to feature location (i.e., shift-invariance)
- The number of free parameters is dramatically reduced, requiring less training time and data
- Geometric and topological information such as the relevance of neighboring pixels is captured
- They maintain the ability to detect local features while allowing their combination into higher-order features

Network Architecture

“The network has three hidden layers named H1, H2, and H3 respectively. Connections entering H1 and H2 are local and heavily constrained.”

The network architecture is as follows.

- The **Input layer** contains the preprocessed image with $16 \times 16 = 256$ total input units.
- **Hidden layer** H_1 consisting of feature maps $H_{1,1}, \ldots H_{1,12}$ each with $8 \times 8 = 64$ hidden units, each of which receives input from a receptive field over $5 \times 5 = 25$ input units. All 64 receptive fields of a given feature map, though receiving input from a *different* set of 25 *input units*, are constrained to share the *same* 25 *weights*. Thus H_1 has the following.
  - $\underset{maps}{12} \times \underset{units}{(8 \times 8)} = 768$ units
  - $\underset{units}{768} \times (\underset{r.field}{5 \times 5} + \underset{bias}{1}) = 19,968$ connections
  - $\underset{maps}{12} \times \underset{r.field}{(5 \times 5)} + \underset{biases}{768} = 1,068$ free parameters
- **Hidden Layer** H_2 similarly consists of 12 feature maps, in this case, with $4 \times 4 = 16$ hidden units each. Each of these units combines information from 8 of the 12 feature maps in H_1. Each receptive field is composed of eight 5×5 neighborhoods centered around units in identical positions within each of the eight selected H_1 feature maps. As before, all units in a given feature map are constrained to have identical weight vectors.”[1]. Thus H_2 has the following.
  - $\underset{maps}{12} \times \underset{units}{(4 \times 4)} = 192$ units
  - $\underset{units}{192} \times (\underset{r.field}{5 \times 5} \times \underset{maps}{8}) + \underset{biases}{192} = 38,592$ connections
  - $\underset{maps}{12} \times (\underset{r.field}{5 \times 5} \times \underset{maps}{8}) + \underset{biases}{192} = 1,068$ free parameters
- **Hidden Layer** H_3 has 30 units and is fully connected to H_2. Thus H_3 has the following.
  - $\underset{units}{30}$ units
  - $\underset{units}{30} \times \underset{H_2\ units}{192} = 5,760$ connections
  - $\underset{units}{30} \times \underset{H_2\ units}{192} + \underset{biases}{30} = 5,790$ free parameters
- The **Output layer** has 10 units and is fully connected to H_3, adding another $30 \times 10 = 300$ units.

In summary, the network has 1,256 units, 64,660 connections, but only 9,760 free parameters.

Considerations

Costs

The authors’ approach, while innovative, comes with several notable limitations.

- Performance depends heavily on quality of preprocessing and segmentation
- System may fail on writing styles not represented in training data
- Low 16×16 resolution input may lose important details in some cases

Benefits

The system introduces numerous advantages over prior approaches.

- Faster training times compared to previous approaches due to fewer parameters
- No need for manual feature engineering or hand-tuned constants
- Robust to variations in digit position and style through weight sharing
- Achieves state-of-the-art performance while requiring minimal preprocessing
- Architecture is simpler and more general than previous approaches

Impact

This paper introduced several key concepts that became foundational to modern deep learning and computer vision.

- The concept of convolutional neural networks (CNNs) with learned features, demonstrating that end-to-end learning was possible without hand-engineered features
- Weight sharing and feature maps, which dramatically reduced the number of parameters while maintaining shift invariance
- Practical evidence that backpropagation could work effectively on large-scale pattern recognition tasks

These innovations laid crucial groundwork for modern deep learning architectures. The concepts introduced here - particularly CNNs with weight sharing - have become standard components in image recognition, computer vision, and many other machine learning applications. The success of this approach helped spark renewed interest in neural networks and contributed to the deep learning revolution of the following decades.

References

[1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[2] J. S. Denker, W. R. Gardner, H. P. Graf, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, H. S. Baird, and I. Guyon. Neural network recognizer for hand-written zip code digits. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, pages 323–331, San Mateo, CA, 1989. Morgan Kaufmann.