LeNet-5
The Classic CNN Architecture for Handwritten Digit Recognition

2025-2 Introduction to Computer Vision

박고운 Park Goun

2021112436

컴퓨터공학전공
Computer Science and Engineering

INDEX

INDEX

What is LeNet-5

- A 7-layer Convolutional Neural Network.
- Proposed in the paper
  "Gradient-Based Learning Applied to Document Recognition"
  by LeCun et al. in 1998 for handwritten digit recognition.
- First successful and widely recognized **modern CNN architecture**.
- Using **weight sharing** and local **receptive fields**.

- "Gradient-Based Learning Applied to Document Recognition"
  - **LeNet-5**
  - Graph Transformer Network
  - Heuristic Over-Segmentation

PROC. OF THE IEEE, NOVEMBER 1998 1

Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

Abstract—
Multilayer Neural Networks trained with the backpropagation algorithm constitute the best example of a successful Gradient-Based Learning technique. Given an appropriate network architecture, Gradient-Based Learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional Neural Networks, that are specifically designed to deal with the variability of 2D shapes, are shown to outperform all other techniques.

Real-life document recognition systems are composed of multiple modules including field extraction, segmentation, recognition, and language modeling. A new learning paradigm, called Graph Transformer Networks (GTN), allows such multi-module systems to be trained globally using Gradient-Based methods so as to minimize an overall performance measure.

Two systems for on-line handwriting recognition are described. Experiments demonstrate the advantage of global training, and the flexibility of Graph Transformer Networks.

A Graph Transformer Network for reading bank check is also described. It uses Convolutional Neural Network character recognizers combined with global training techniques to provides record accuracy on business and personal checks. It is deployed commercially and reads several million checks per day.

Keywords— **Neural Networks, OCR, Document Recognition, Machine Learning, Gradient-Based Learning, Convolutional Neural Networks, Graph Transformer Networks, Finite State Transducers.**

NOMENCLATURE

- GT Graph transformer.
- GTN Graph transformer network.
- HMM Hidden Markov model.
- HOS Heuristic oversegmentation.
- K-NN K-nearest neighbor.
- NN Neural network.
- OCR Optical character recognition.
- PCA Principal component analysis.
- RBF Radial basis function.
- RS-SVM Reduced-set support vector method.
- SDNN Space displacement neural network.
- SVM Support vector method.
- TDNN Time delay neural network.
- V-SVM Virtual support vector method.

The authors are with the Speech and Image Processing Services Research Laboratory, AT&T Labs-Research, 100 Schulz Drive Red Bank, NJ 07701. E-mail: {yann,leonb,yoshua,haffner}@research.att.com. Yoshua Bengio is also with the Département d'Informatique et de Recherche Opérationelle, Université de Montréal, C.P. 6128 Succ. Centre-Ville, 2920 Chemin de la Tour, Montréal, Québec, Canada H3C 3J7.

I. INTRODUCTION

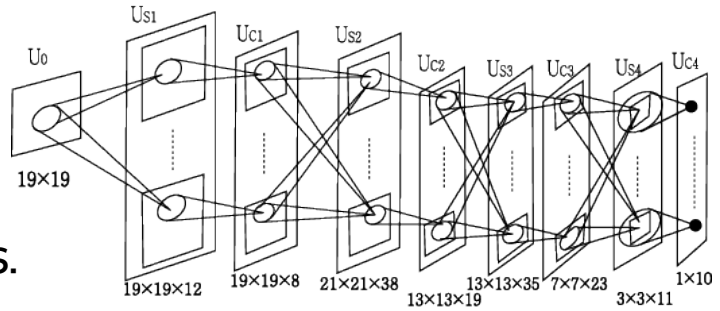
Over the last several years, machine learning techniques, particularly when applied to neural networks, have played an increasingly important role in the design of pattern recognition systems. In fact, it could be argued that the availability of learning techniques has been a crucial factor in the recent success of pattern recognition applications such as continuous speech recognition and handwriting recognition.

The main message of this paper is that better pattern recognition systems can be built by relying more on automatic learning, and less on hand-designed heuristics. This is made possible by recent progress in machine learning and computer technology. Using character recognition as a case study, we show that hand-crafted feature extraction can be advantageously replaced by carefully designed learning machines that operate directly on pixel images. Using document understanding as a case study, we show that the traditional way of building recognition systems by manually integrating individually designed modules can be replaced by a unified and well-principled design paradigm, called *Graph Transformer Networks*, that allows training all the modules to optimize a global performance criterion.

Since the early days of pattern recognition it has been known that the variability and richness of natural data, be it speech, glyphs, or other types of patterns, make it almost impossible to build an accurate recognition system entirely by hand. Consequently, most pattern recognition systems are built using a combination of automatic learning techniques and hand-crafted algorithms. The usual method of recognizing individual patterns consists in dividing the system into two main modules shown in figure 1. The first module, called the feature extractor, transforms the input patterns so that they can be represented by low-dimensional vectors or short strings of symbols that (a) can be easily matched or compared, and (b) are relatively invariant with respect to transformations and distortions of the input patterns that do not change their nature. The feature extractor contains most of the prior knowledge and is rather specific to the task. It is also the focus of most of the design effort, because it is often entirely hand-crafted. The classifier, on the other hand, is often general-purpose and trainable. One of the main problems with this approach is that the recognition accuracy is largely determined by the ability of the designer to come up with an appropriate set of features. This turns out to be a daunting task which, unfortunately, must be redone for each new problem. A large amount of the pattern recognition literature is devoted to describing and comparing the relative

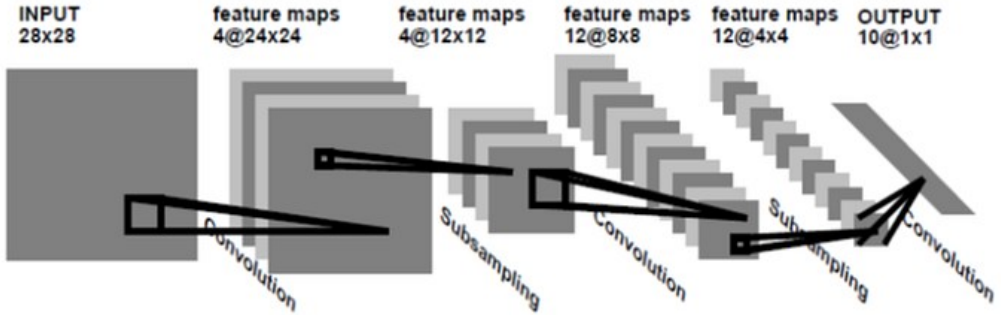
Earlier CNN-like Work Before LeNet-5

- Neocognitron (Fukushima, 1980)
  - Considered the ancestor of CNNs
  - Introduced convolution-like receptive field and pooling concepts.
  - But lacked modern gradient-based learning and wasn't practical for large datasets.

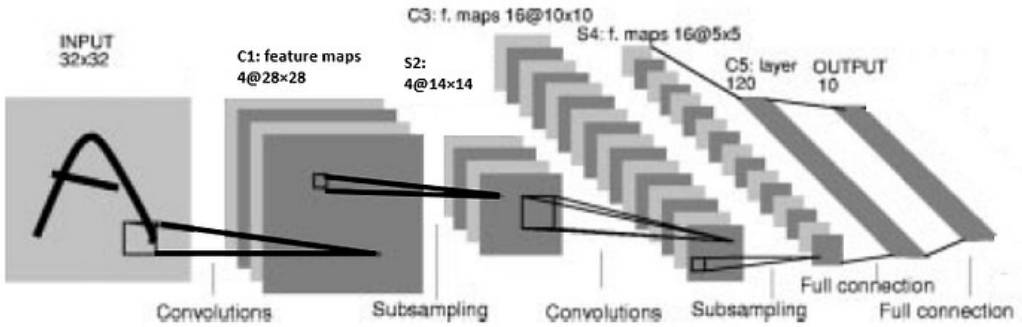


Neocognitron

- Lenet-1 to LeNet-4 (LeCun, early 1990s)
  - Earlier iterations leading to LeNet-5
  - Gradually refined convolution + pooling + backprop training structure.



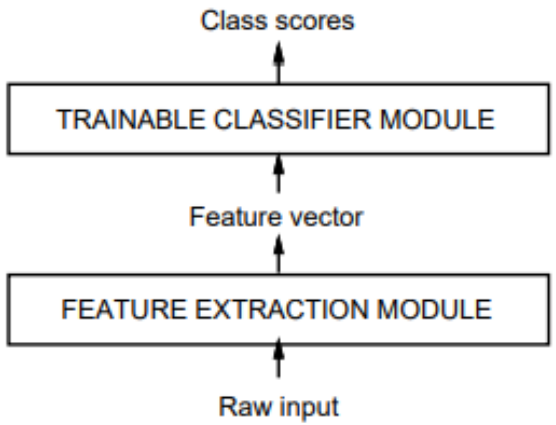
LeNet-1



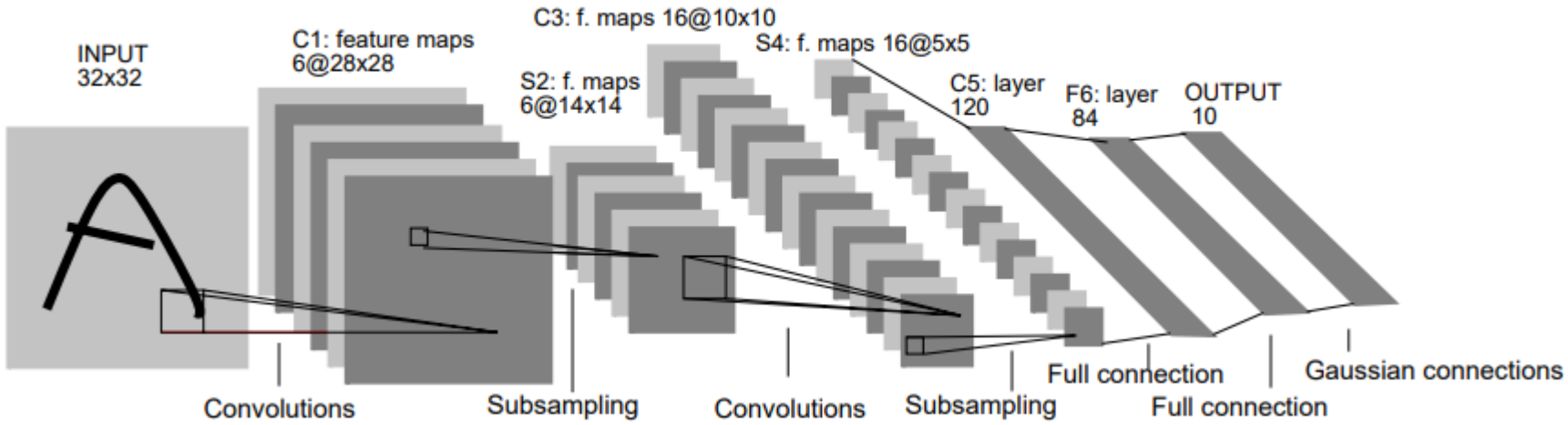
LeNet-4

Traditional pattern recognition

- Better pattern recognition systems can be built by relying more on automatic learning, and less on hand-designed heuristics.
- The traditional classifier is general-purpose and trainable, but the recognition accuracy is largely determined by the ability of the designer to come up with an appropriate set of features.
- The fully-connected network has too many weights and is difficult to consider the topology.



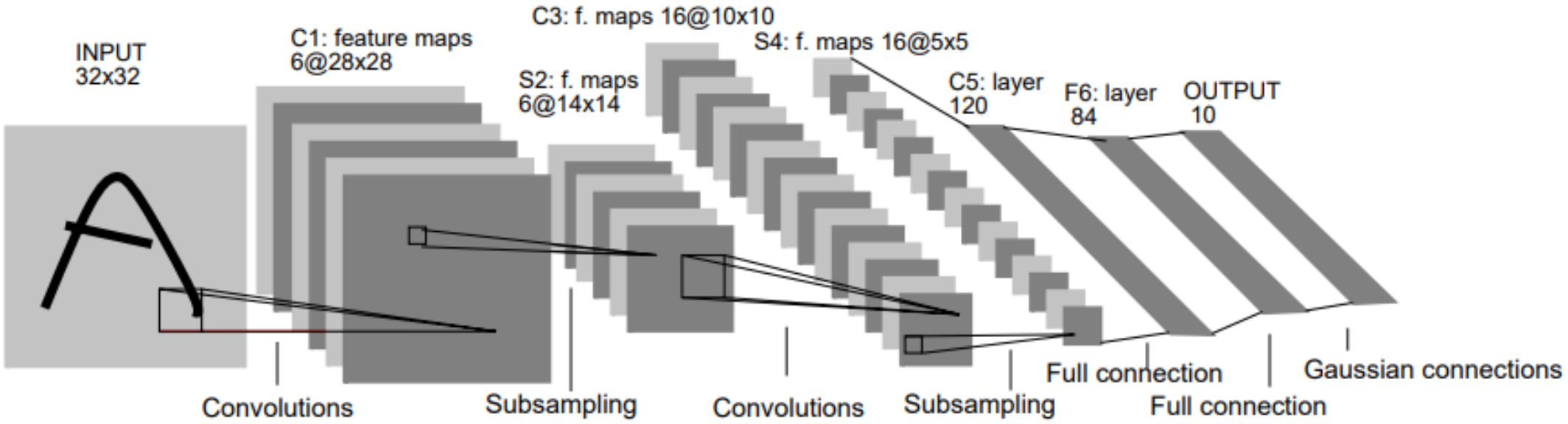
Traditional pattern recognition by feature extraction



LeNet-5

INDEX

Overview

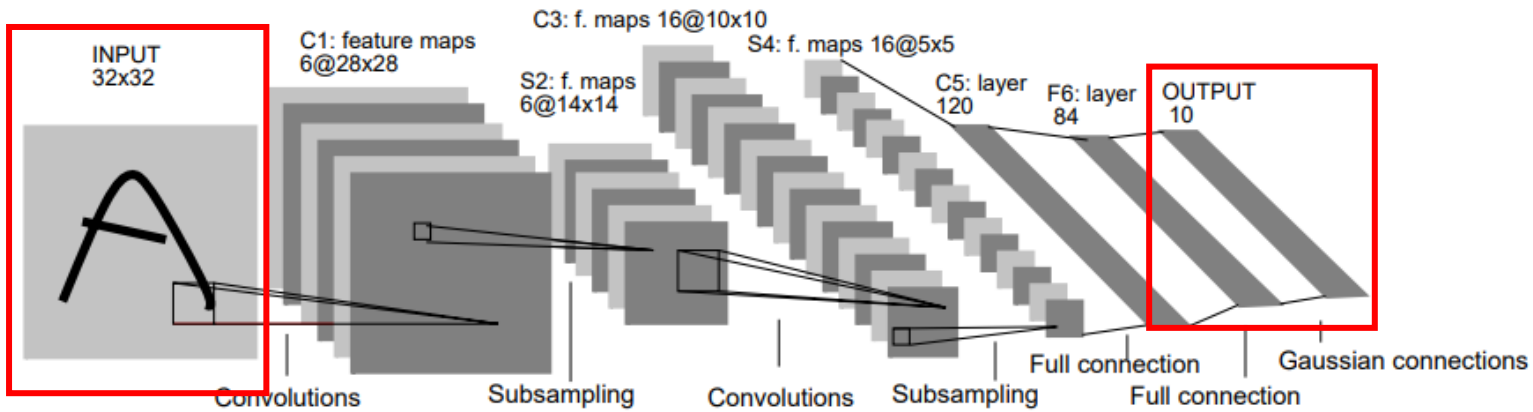


Architecture of LeNet-5, a CNN, for digits recognition
A 7-layer Convolution Neural Network

Input & Output

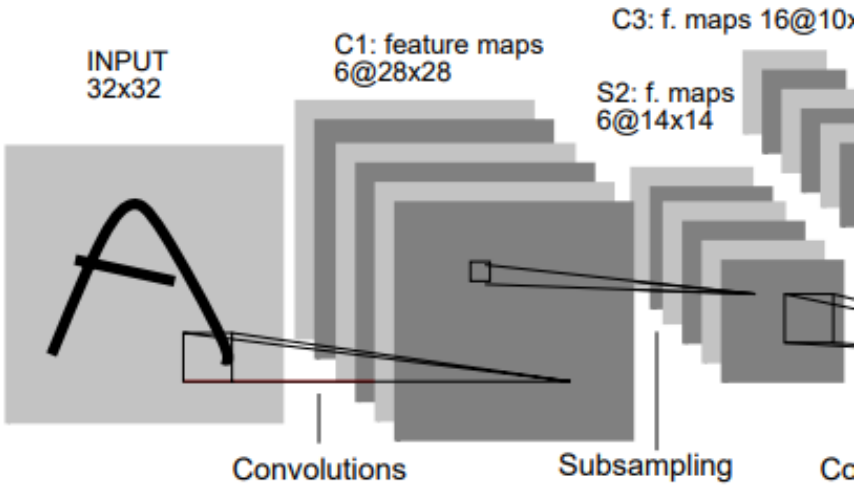
- Input
  - **32x32 image**
- MNIST Database
  - the Modified NIST set
  - A large database of handwritten digits .
  - Commonly used for training various image processing systems and machine learning.
- Output
  - **0~9 digits classification**

MNIST Database

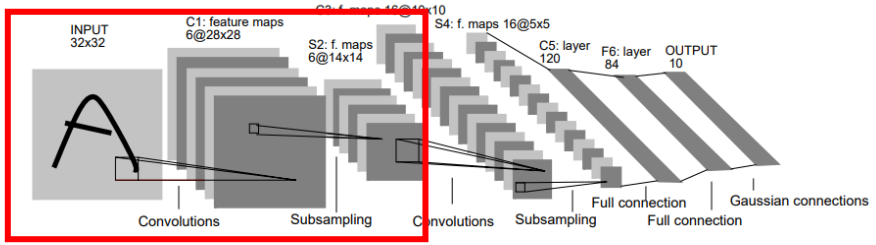


Layer

- **C1**
  - A **convolutional layer** with 6 feature maps.
  - Each feature map is connected to a **5x5 neighborhood** in the input.
  - 32x32 image → 6 feature maps @28x28
  - 156 trainable parameters, and 122,304 connection.

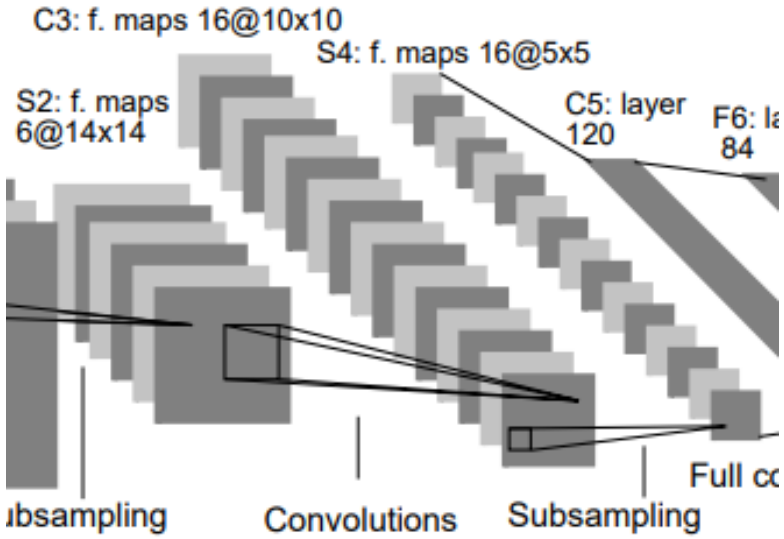


- **S2**
  - A **sub-sampling layer** with 6 feature maps.
  - **2x2 neighborhood** in corresponding feature map in C1.
  - 6 feature maps @28x28 → 6 @14x14
  - The four inputs to a unit in S2 are added, multiplied by the coefficient and added to a bias.
  - The result is passed through a **sigmoidal function**.
  - 12 trainable parameters. and 5,880 connections.



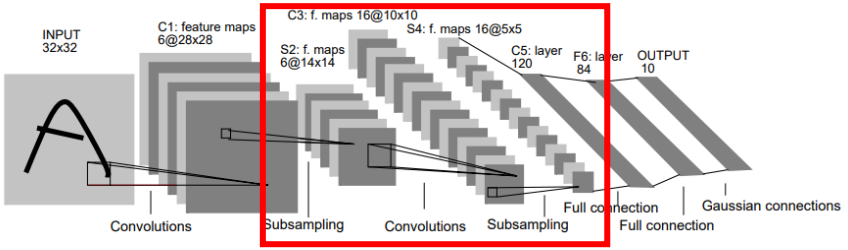
Layer

- **C3**
  - A **convolutional layer** with 16 feature maps.
  - Connected to **5x5 neighborhoods** at identical location in S2.
  - Not connect every S2 feature map to every C3 feature map. It forces a break of symmetry in the network.
  - 6 feature maps @14x14 → 16 @10x10
  - 1,516 trainable parameters, and 151,600 connections.
- **S4**
  - A **sub-sampling layer** with 16 feature maps of size 5x5
  - **2x2 neighborhoods** in corresponding feature map in C3.
  - 16 @10x10 → 16 @5x5
  - 32 trainable parameters, and 2,000 connections.



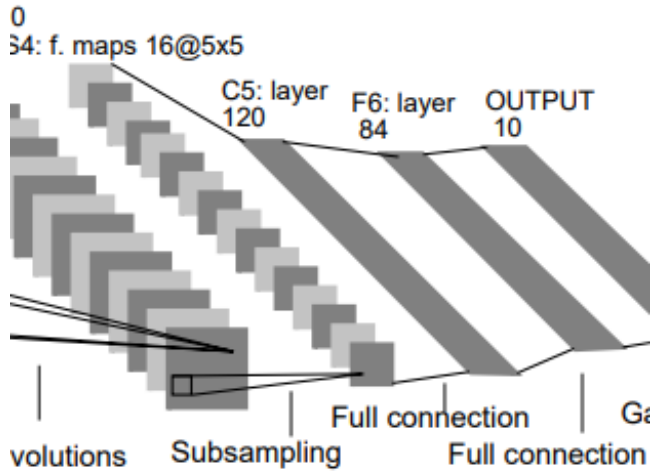
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X		X	X		X
5				X	X	X			X	X	X	X		X	X	X

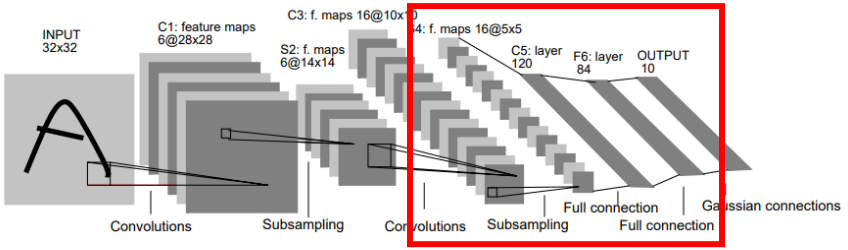
S2 to C3 Table



Layer

- **C5**
  - A **convolutional layer** with 120 feature maps.
  - Each unit is connected to a **5x5 neighborhood** on all 16 of S4's.
  - C5 is **labeled as a convolutional layer**, because if LeNet-5 input were made bigger with everything else, the dimension would be larger than 1x1
  - 16 feature map @5x5 → 120 @1x1
  - 48,120 trainable connections.
- **F6**
  - Contains 84 units and is **fully connected to C5**.
  - 10,164 trainable parameters.
- **OUTPUT**
  - Composed of Euclidean Radial Basis Function units(RBF)



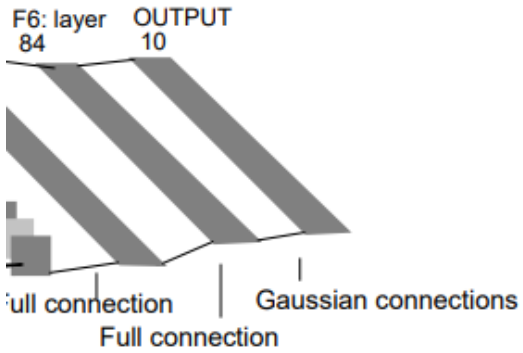


Layer

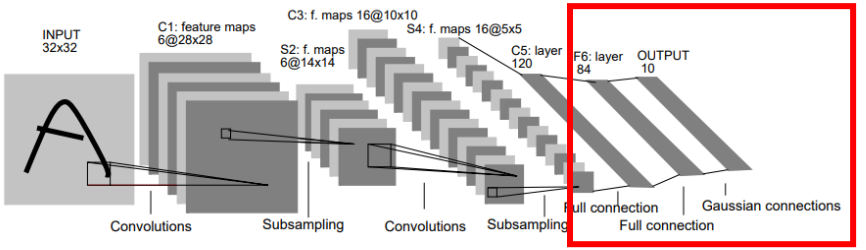
- **F6**
  - Contains **84 units** and is **fully connected to C5**.
  - 10,164 trainable parameters.
  - Designed to represent a stylized image of the corresponding character class drawn on a 7x12 bitmap.
  - Such a representation is not particularly useful for recognizing isolated digits, but quite **useful for recognizing string of character** from the full ASCII set
- **OUTPUT**
  - Composed of Euclidean Radial Basis Function units(RBF)

$$y_i = \sum_j (x_j - w_{ij})^2.$$

RBF



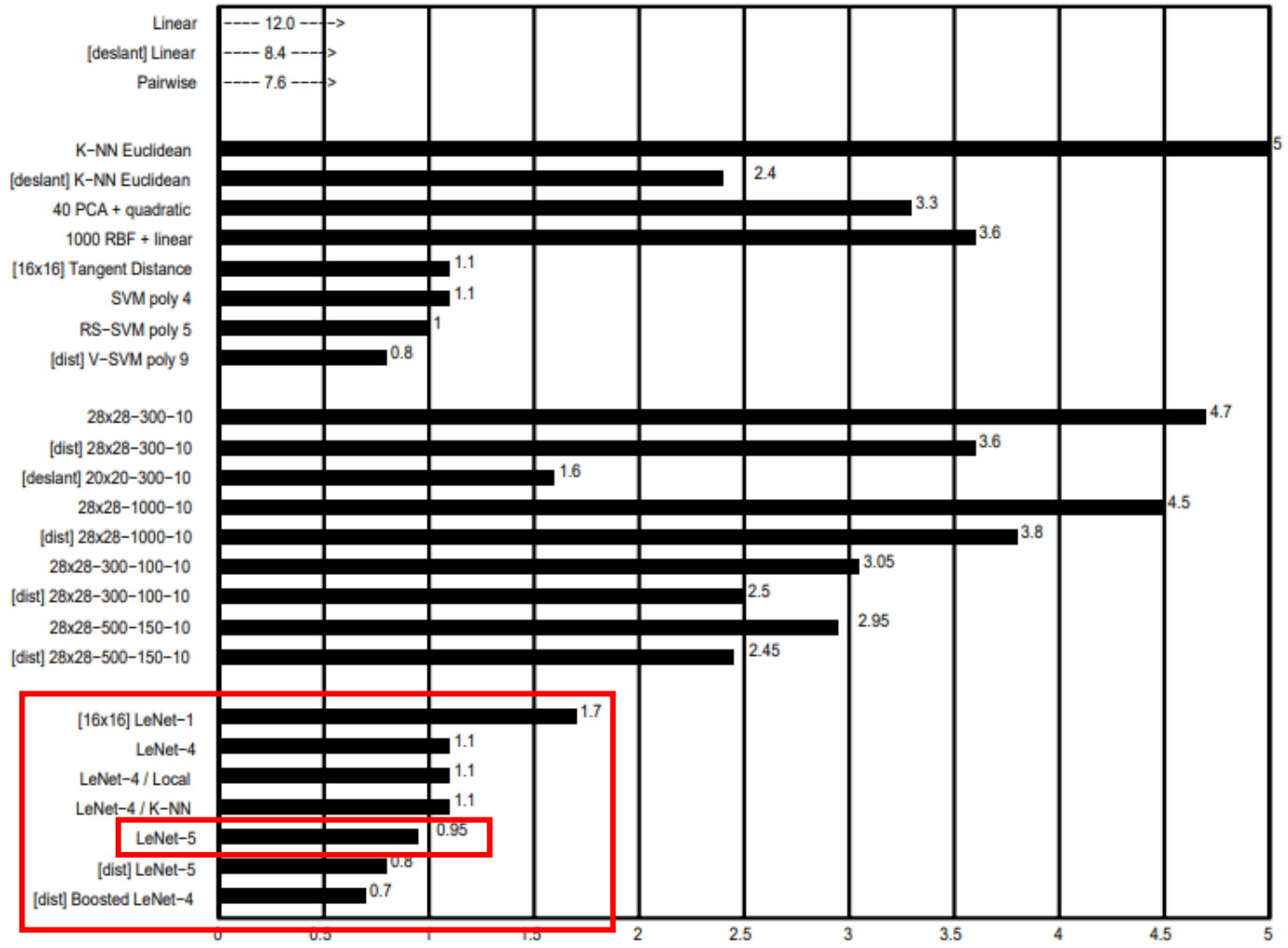
7x12 bitmap (84)



INDEX

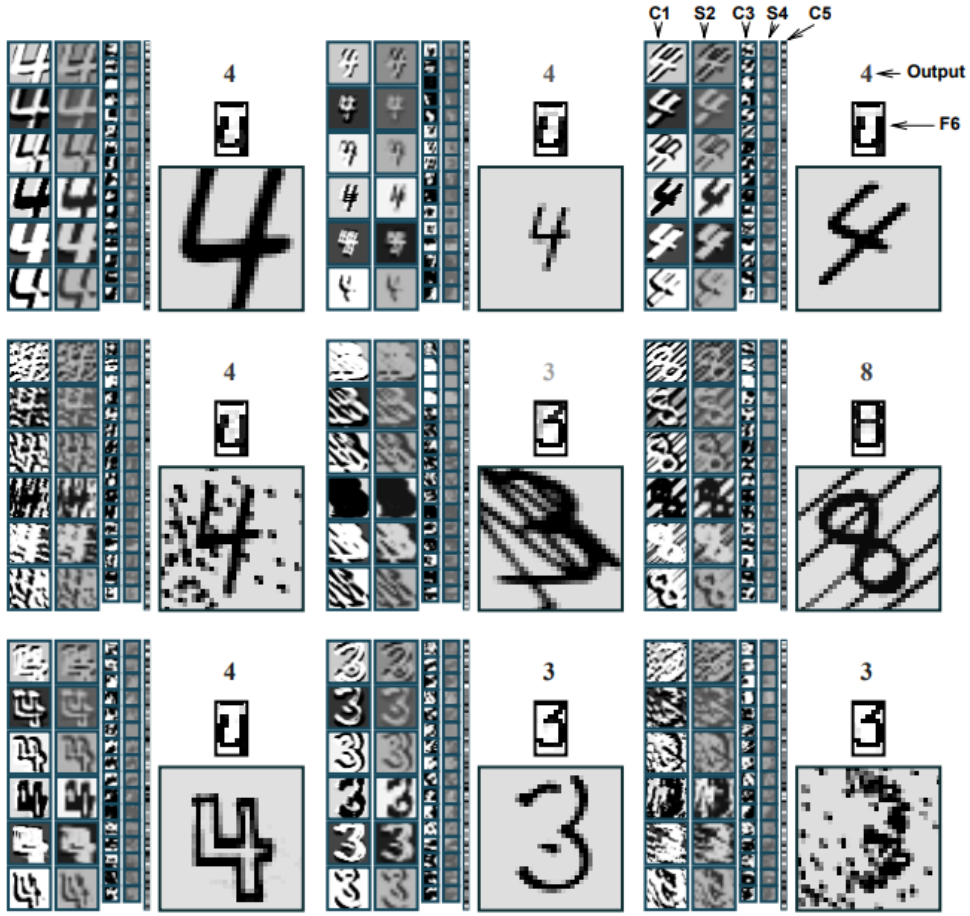
Compare with Other Classifier

Error rate on the test set
(%) for various classification methods



Robustness of LeNet-5

Examples of unusual, distorted, and noisy character correctly recognized by LeNet-5



INDEX

What Made LeNet-5 Innovative (in 1998)

- ✓ Establishing the **Modern CNN Architecture**
  - Introduced the now-standard pipeline:

    Convolution → Subsampling(pooling) → Convolution → Fully Connected
  - Demonstrated practical use of **weight sharing** and **local receptive fields**.

- ✓ End-to-End Learning Without Hand-Crafted Features
  - Replaced traditional feature engineering with direct gradient-based learning.

- ✓ Real-World, Industrial-Level Performance.
  - Applied not only to MNIST but also to real banking systems.(check reading)

- ✓ Integrated Document Processing Framework.
  - Introduced **Graph Transformer Networks (GTN)** to unify recognition, segmentation, and contextual modeling.

Lasting Impact Today

- ✓ Foundation of **All Modern CNN Architectures**
  - VGG, ResNet, Inception, and many others trace their structural roots back to LeNet-5
  - Concepts like **parameter sharing**, **locality**, and **pooling** remain fundamental.

- ✓ Start of the End-to-Ent Deep Learning Era
  - Marked the shift from hand-crafted features to representation learning.
  - Influenced **not only CNNs** but also today's **Transformers, diffusion models, and autonomous vision systems**.

- ✓ First Practical Deep Learning System
  - Showed **how deep networks could be deployed in real document-processing pipelines**.
  - Modern OCR, handwriting recognition, and document automation systems all build on this lineage.

Thank You