Yann Lecun, Léon Bottou, Yoshua Bengio, Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86 (11), pp.2278-2324. 10.1109/5.726791 . hal-03926082

Gradient-Based Learning Applied to Document Recognition

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

Carolina Dias
Claudio Fortier

Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual do Ceará

O que veremos hoje...

- Introdução ao problema de reconhecimento de padrões
- Evolução das técnicas mais tradicionais e manuais até o uso de CNNS
- Algoritmo de *backpropagation*
- O surgimento do clássico conjunto de dados MNIST
- Apresentação da rede neural LeNet-5
- Aplicação real do sistema de reconhecimento de caracteres manuais

Ao final, aprenderemos...

- Como funcionam os algoritmos de aprendizagem baseados em gradiente
- Como classificar padrões de grandes dimensões, como caracteres escritos à mão, com o maior uso possível de passos automatizados
- Porque as Redes Neurais Convolucionais possuem o melhor desempenho nessa tarefa específica
- Como esses algoritmos podem ser utilizados em diversas outras aplicações

Introdução

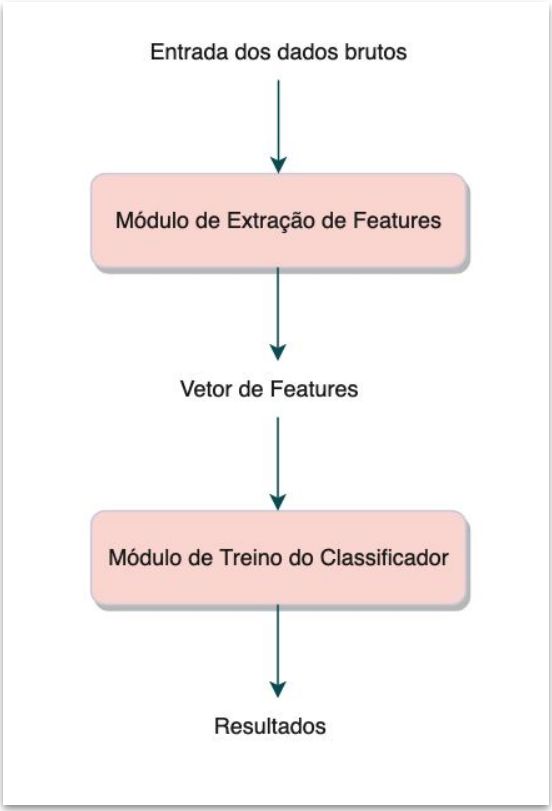
- Esse artigo é um trabalho fundamental que introduziu técnicas de aprendizado baseado em gradiente para reconhecimento de documentos.
- Sistemas de reconhecimento de padrões utilizavam* cada vez mais técnicas de aprendizado de máquina e redes neurais em seu design.
- Esses sistemas incluem reconhecimento de fala em tempo real e de caracteres escritos à mão.
- Mesmo assim, ainda eram utilizadas heurísticas manuais nesses sistemas:
  - Extração de features
  - Pré-processamento de features
- A variabilidade e riqueza dos dados transformava a tarefa de se construir um sistema completamente automático quase impossível.

*em 1998

Método Tradicional de Reconhecimento de Padrões

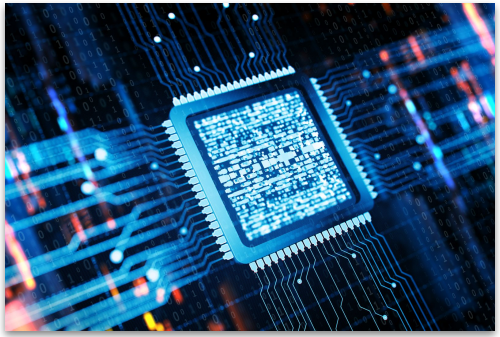
Dois módulos principais:

- **Módulo de Extração de Features:**
  - transforma os dados brutos da entrada em vetores de poucas dimensões que possam ser facilmente comparados e que preservem as transformações e distorções dos dados originais
  - é realizado manualmente 😢
  - específico para cada atividade
- **Módulo de Treino do Classificador:**
  - generalista para diversos tipos de atividades
  - é treinado automaticamente 😀

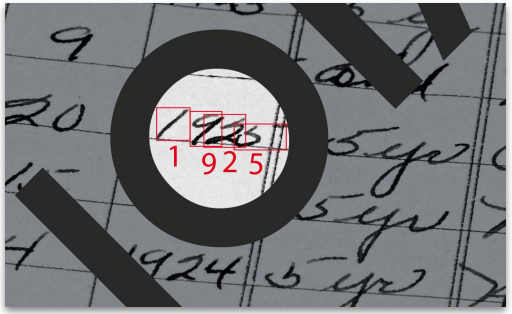


Menos técnicas manuais, pois...

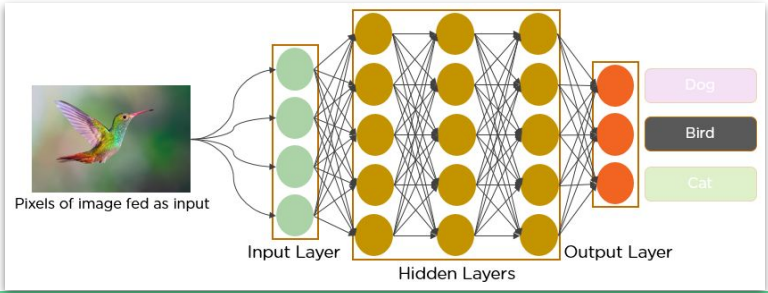
- Tivemos o avanço do poder computacional das máquinas;
- Existe a maior disponibilidade de grandes conjuntos de dados de treino;
- Houve a evolução de técnicas de aprendizado de máquina cada vez mais poderosas capazes de lidar com dados de grandes dimensões.



No entanto, nenhuma técnica automática consegue sucesso sem um mínimo de conhecimento prévio sobre a atividade a ser realizada.



- Reconhecimento de padrões de caracteres escritos à mão
  - Imagens são formas em duas dimensões (2D)
    - Redes Neurais Convolucionais são especializadas em problemas relacionados a formas 2D



Aprendendo com os Dados: Aprendizado Baseado em Gradientes

$$Y^p = F(Z^p, W)$$

- Y^p é a classe prevista para o dado Z^P
- F é a função computada
- W são os pesos (parâmetros ajustáveis)

Aprendendo com os Dados: Aprendizado Baseado em Gradientes

$$E^p = \mathcal{D}(D^p, F(Z^p, W))$$

- E^p é o erro medido entre
  - D^p, o valor esperado e
  - $F(Z^p, W)$, o valor obtido

Aprendendo com os Dados: Aprendizado Baseado em Gradientes

- $E_{train}(W)$ é a média dos erros E^p no conjunto de treino $\{(Z^1, D^1), \ldots, (Z^p, D^p)\}$

- Muito mais relevante é saber qual a média dos erros no **conjunto de teste**
- A diferença do erro entre treino e teste sempre diminui com o aumento da quantidade de dados de treino
- O objetivo é minimizar essa função de perda

Minimizando a Função de Perda

- A função de perda consegue ser minimizada ao estimarmos o impacto de pequenas variações nos valores dos parâmetros da função (os pesos).
- Isso é mensurado pelo gradiente da função de perda em relação aos parâmetros.

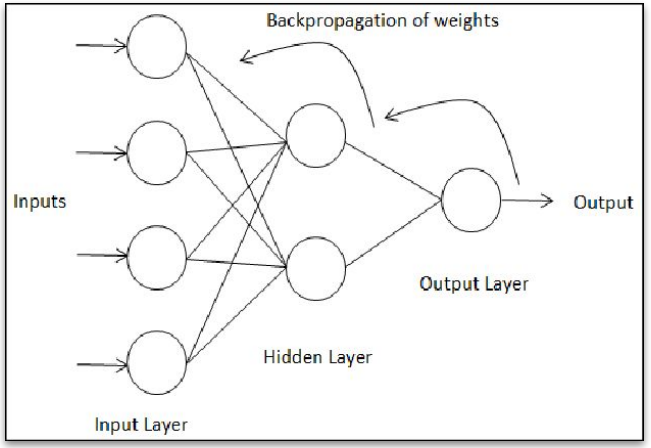
Algoritmo do Gradiente Descendente

Os pesos são ajustados iterativamente da seguinte forma:

$$W_k = W_{k-1} - \epsilon \frac{\delta E(W)}{\delta W}$$

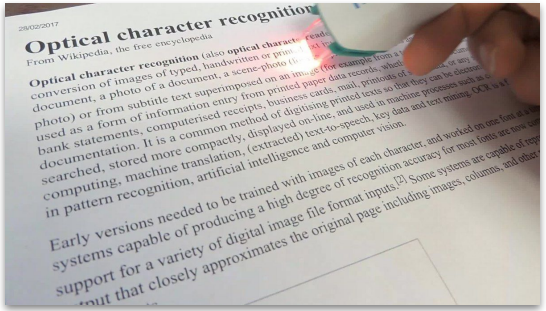
Gradient Backpropagation

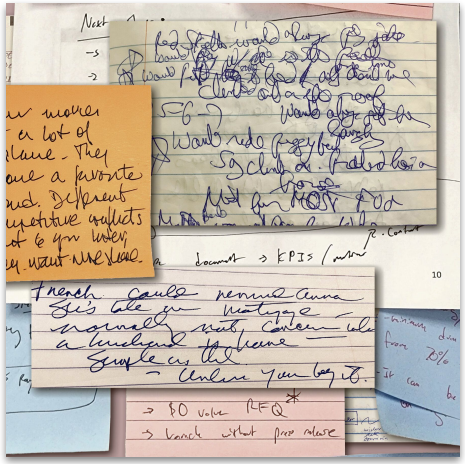
- As técnicas de aprendizado baseadas em gradiente eram bastante utilizadas para a resolução de sistemas lineares.
- Para utilizar essas técnicas em sistemas não-lineares, foi necessário o desenvolvimento do algoritmo de *backpropagation* para computar o gradiente nesses sistemas.
- A ideia principal é que os gradientes podem ser computados eficientemente ao serem propagados da saída para a entrada do problema.



Sistema de Reconhecimento de Caracteres

- O reconhecimento de documentos possui importantes aplicações:
  - Identificação e classificação de documentos, como escrita à mão ou impressa;
  - OCR (Optical Character Recognition);
  - Detecção de fraudes em assinaturas;
  - Organização de arquivos, etc.
- E também possui grandes dificuldades:
  - Variabilidade: Diferentes estilos de escrita, tamanhos de fonte, distorções.
  - Ruído: Manchas, distorções na digitalização, interferências.
  - Problemas em como separar cada caractere de uma palavra/frase.



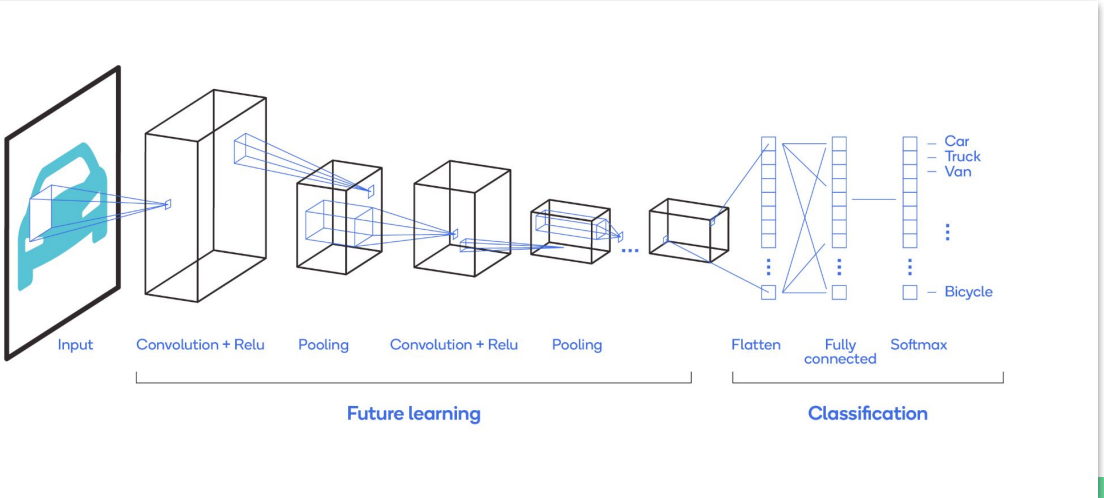


Sistemas Globalmente Treináveis

- Geralmente os sistema de reconhecimento de padrões são compostos por múltiplos módulos, que precisam ser treinados individualmente.
- Uma melhor alternativa seria treinar o sistema inteiro para minimizar o erro global e melhorar a classificação dos caracteres.
- Calcular a função de perda do sistema global e ir propagando ao invés de calcular de cada módulo individualmente é uma opção para melhorar a performance como um todo.

CNNs para Reconhecimento de Caracteres Isolados

- CNNs treinadas com gradiente descendente possuem alta capacidade de aprender funções difíceis, de grandes dimensões e não-lineares a partir de um grande conjunto de dados
- Isso torna as CNNs fortes candidatas para o reconhecimento de imagens
- Conseguem receber imagens com pouca ou nenhuma modificação.



Por que não utilizar uma rede neural tradicional completamente conectada para o reconhecimento de imagens?

- Imagens são grandes, com centenas de variáveis (os pixels).
  - Isso implica que as camadas completamente conectadas teriam milhares de pesos a serem calculados.
  - Necessitaria de um grande conjunto de dados de treino para aprender esses pesos.
  - Alto uso de memória, o que requer hardwares poderosos.
- Principal problemática: essas redes neurais não lidam com variações nos dados, como translações ou distorções, algo comum em imagens.
- Não conseguem representar as estruturas 2D de imagens ➜ não capturam que pixels próximos são relacionados ➜ não extraem arestas, quinas, etc.

Redes Neurais Convolucionais

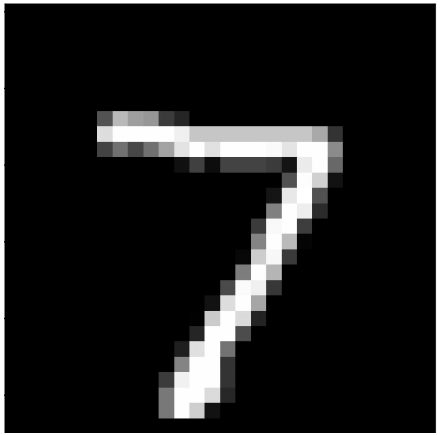
- Lidam bem com variações em escala, shift e distorções nos dados.
- Para isso, fazem uso de
  - **Campos Receptores Locais:** os neurônios são capazes de extrair elementos visuais como arestas, quinas, contornos, etc. Esses elementos são então combinados pelas camadas subsequentes.
  - **Pesos Compartilhados (ou replicados):** os contornos e arestas das imagens sofrem variações espaciais e de escala, mas continuam os mesmos na imagem completa ➜ essas unidades então são forçadas a compartilharem os mesmos pesos ➜ aplicam a mesma operação em diferentes partes da imagem ➜ *feature map*
  - **Subamostragem (temporal ou espacial):** reduzir a resolução do input ao aplicar uma função (como a média dos pixels adjacente) a fim de reduzir a sensibilidade a shifts e a distorções.

Redes Neurais Convolucionais

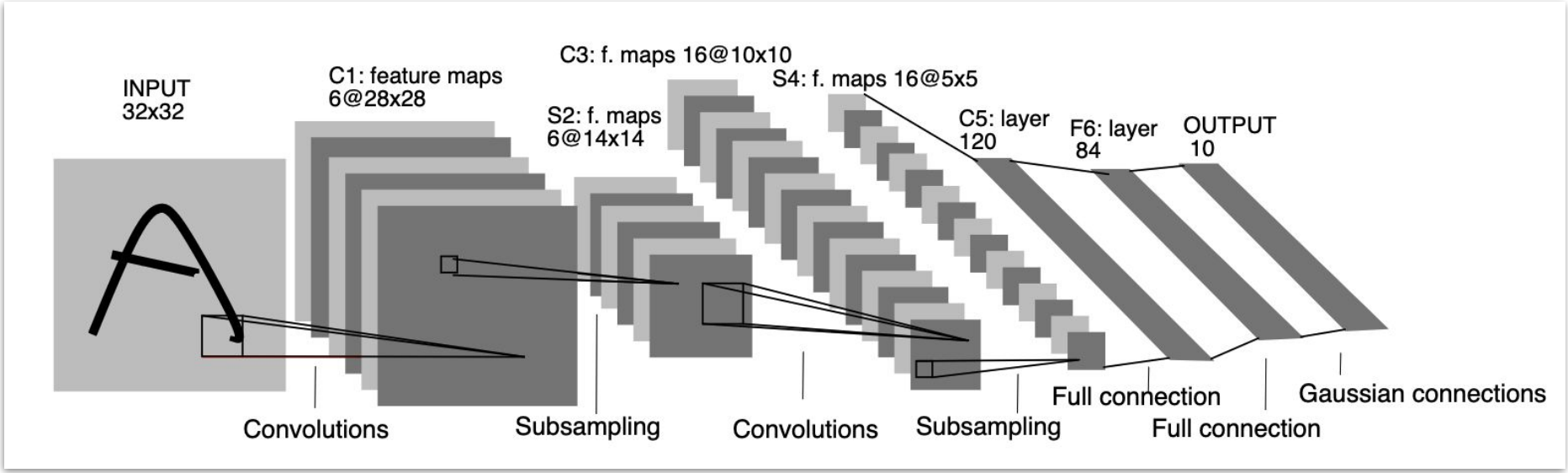
- Uma vez que uma feature (como o contorno) de uma imagem é detectada, sua posição exata perde importância. Apenas a posição *aproximada* EM RELAÇÃO as outras features é relevante.
- *Exemplo:* Qual número possui as seguintes features?
  - Ponta de um segmento horizontal no canto superior esquerdo da imagem
  - Uma quina no canto superior direito da imagem
  - Ponta de um segmento vertical na parte de baixo da imagem

Redes Neurais Convolucionais

- Uma vez que uma feature (como o contorno) de uma imagem é detectada, sua posição exata perde importância. Apenas a posição *aproximada* EM RELAÇÃO as outras features é relevante.
- *Exemplo:* Qual número possui as seguintes features?
  - Ponta de um segmento horizontal no canto superior esquerdo da imagem
  - Uma quina no canto superior direito da imagem
  - Ponta de um segmento vertical na parte de baixo da imagem
- Resposta: número 7



LeNet-5


INPUT
32x32
C1: feature maps
6@28x28
S2: f. maps
6@14x14
C3: f. maps 16@10x10
S4: f. maps 16@5x5
C5: layer
120
F6: layer
84
OUTPUT
10
Convolutions
Subsampling
Convolutions
Subsampling
Full connection
Full connection
Gaussian connections

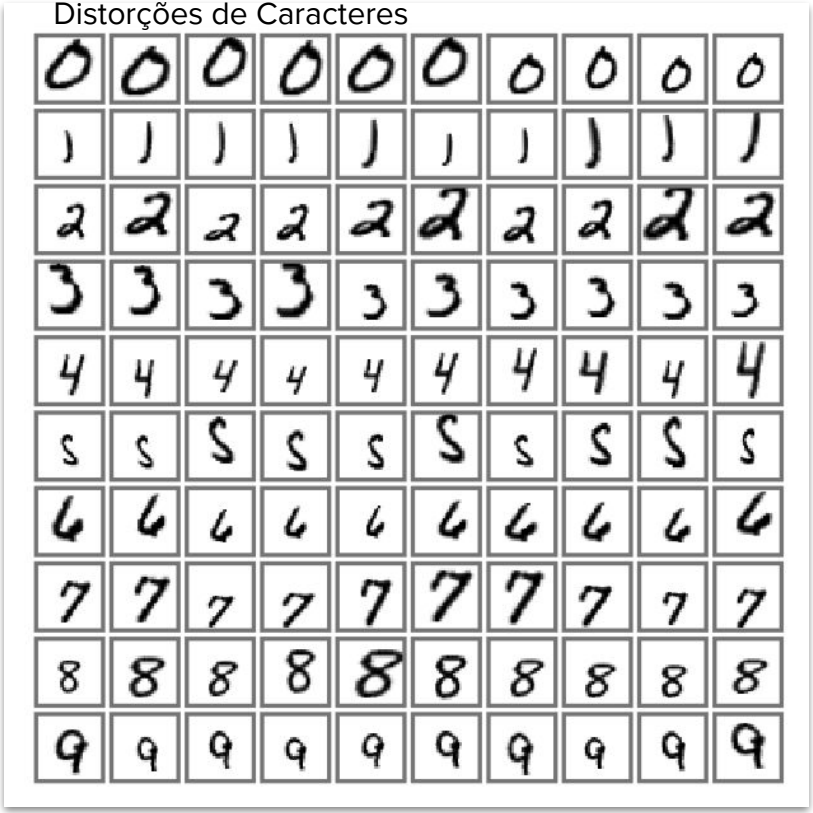
LeNet-5

- 7 camadas (+ o input):
  - Input: imagem de 32x32 pixels
  - Camada C1: camada convolucional com 6 *feature maps*
  - Camada S2: camada de subamostragem, reduzindo a imagem para 14x14 pixels
  - Camada C3: camada convolucional com 12 *feature maps*
  - Camada S4: camada de subamostragem, reduzindo a imagem para 5x5 pixels
  - Camada C5: camada convolucional com 16 *feature maps*
  - Camada F6: totalmente conectada com a C5
  - Saída: classificação dos caracteres

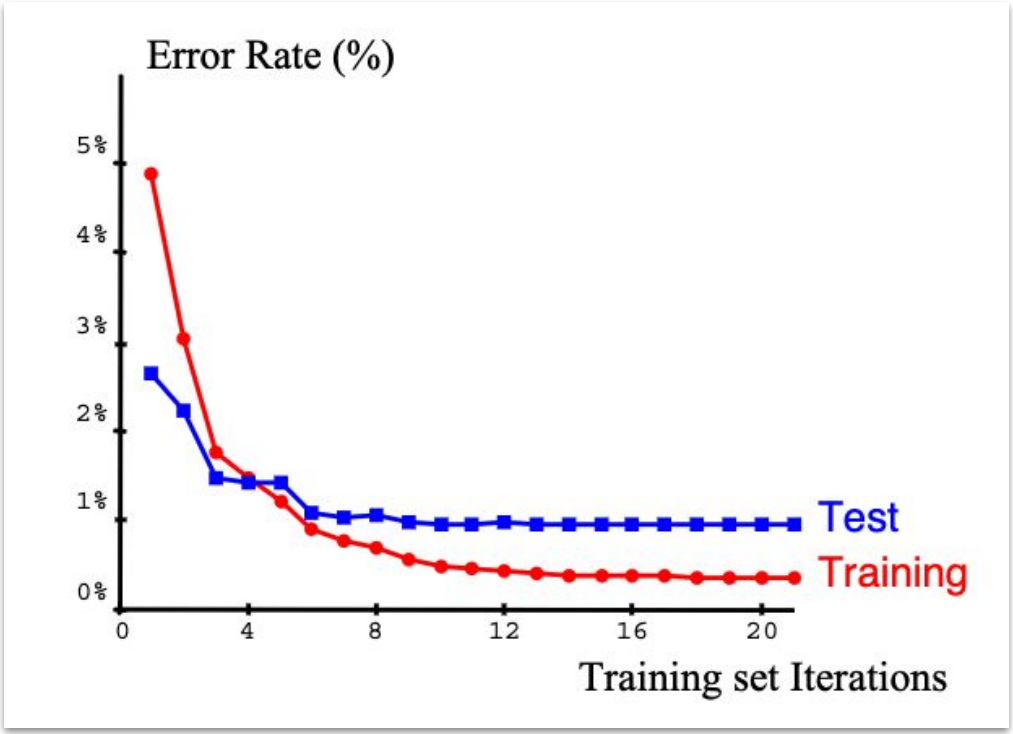
Conjunto de Dados MNIST - Modified NIST

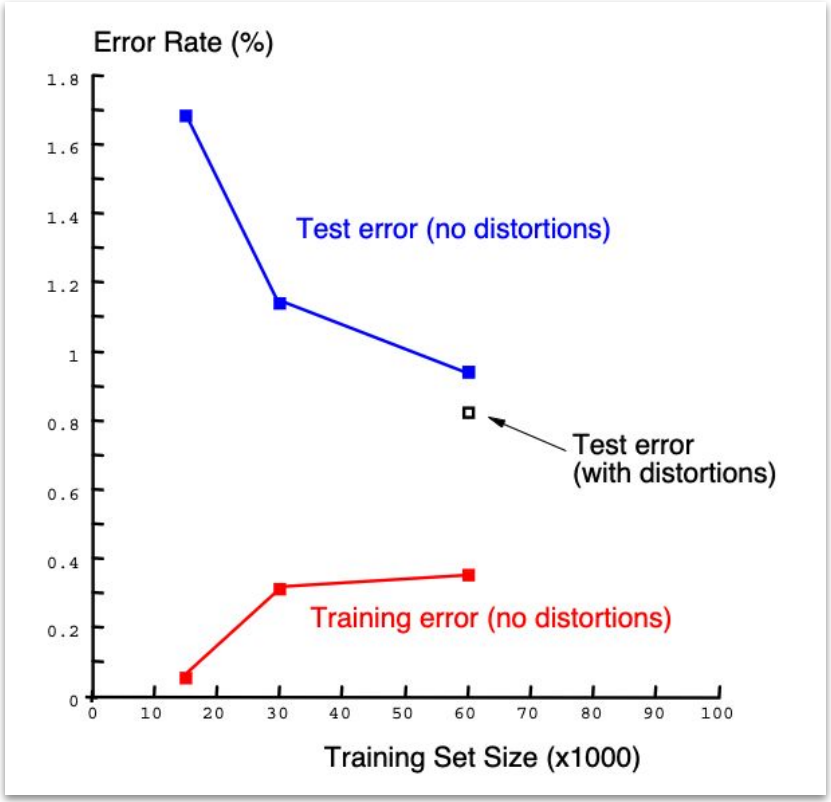
Normalização de Tamanho

Distorções de Caracteres

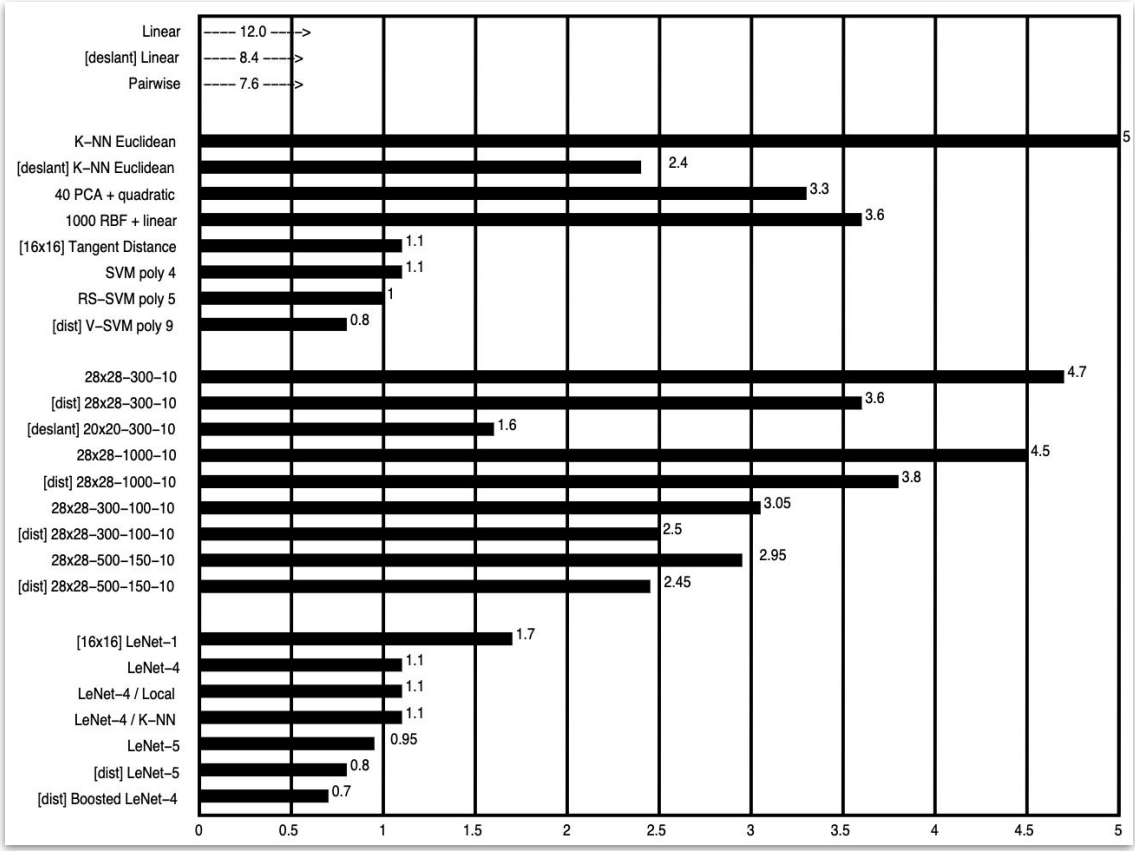


Comparação de Resultados: LeNet-5





Comparação de Resultados: Erro no conj. de teste

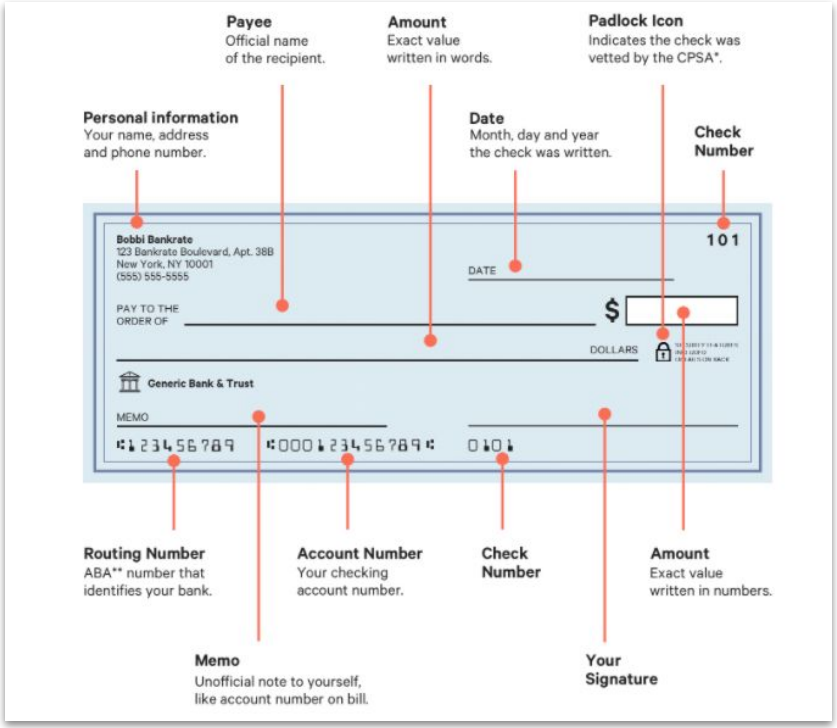


Discussão dos Resultados

- Foram realizados diversos testes com as diferentes técnicas de aprendizado de máquina para compará-las.
- Testes com os dados "limpos", com os dados com defeitos adicionados (números tortos, mal escritos, com falhas na escrita, etc).
- A melhor performance foi a do Boosted LeNet-4 (0,7%), um *ensemble* de diversas técnicas que se conectam para escolher a melhor opção dentre elas.
- Logo após temos a LeNet-5 com 0,8% de erro nos 10.000 dados do conjunto de teste. Sua principal vantagem é a relativa simplicidade se comparada com uma rede neural "*boosted*" com a anterior.

Graph Transformer Network

- Utilização de grafos direcionados acíclicos (DAGs) para realizar a "conversa" entre os diferentes módulos de um sistema de reconhecimento de padrões.
- Exemplo de um uso desse sistema para o reconhecimento de escrita de valores em cheques, utilizado em um banco.



Conclusão

- Esse artigo mostrou que utilizar sistemas automatizados superam até mesmo sistemas manuais (além de vantagens no tempo e valor).
- As CNNs lidam muito bem com imagens, pois conseguem captar distorções e ruídos nativamente.

Dúvidas?