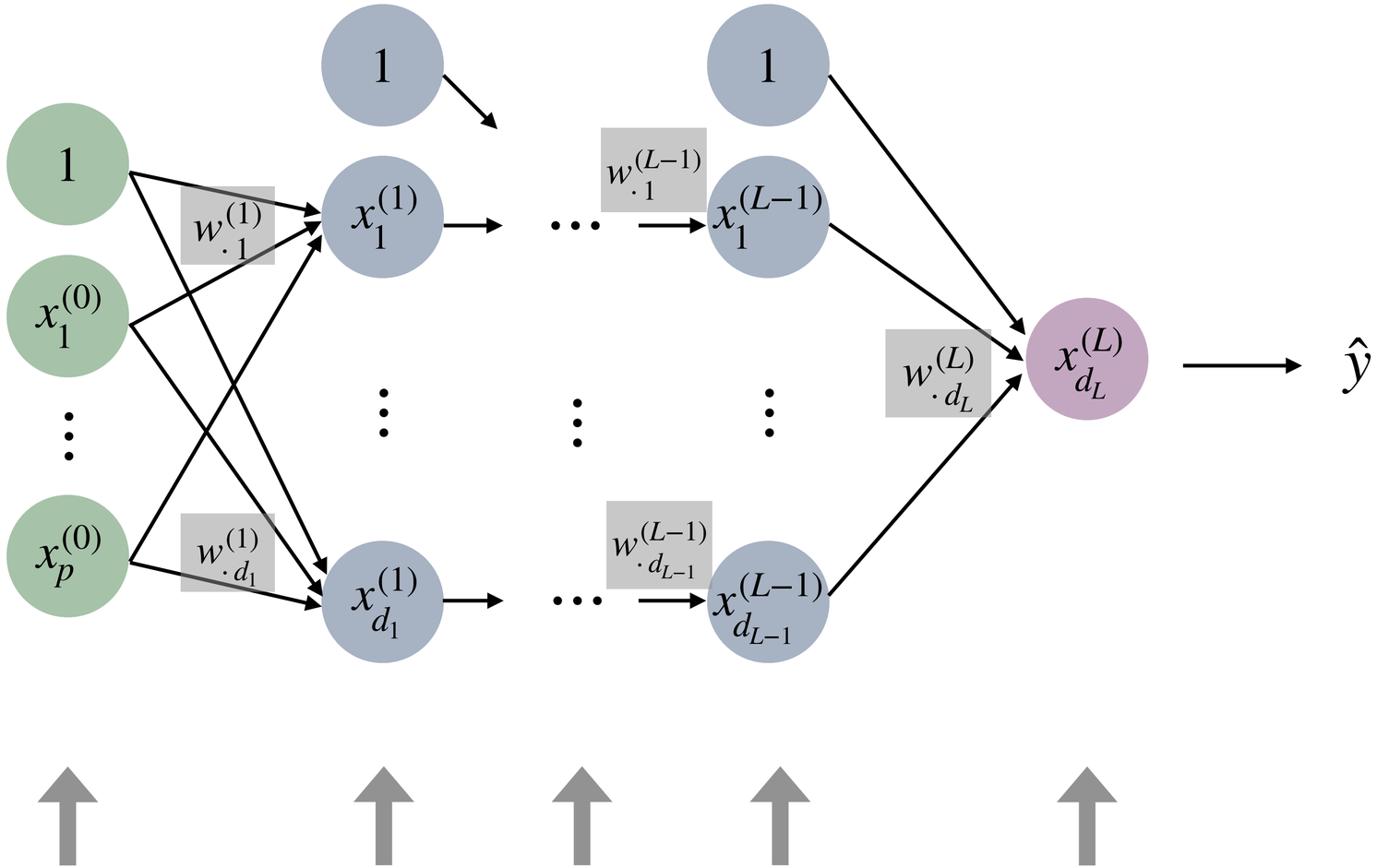
Aprendizagem estatística em altas dimensões

Florencia Leonardi

Conteúdo

- Redes neurais convolucionais
- Redes neurais recorrentes
- Exemplos de aplicação

Rede neural de múltiplas camadas ocultas

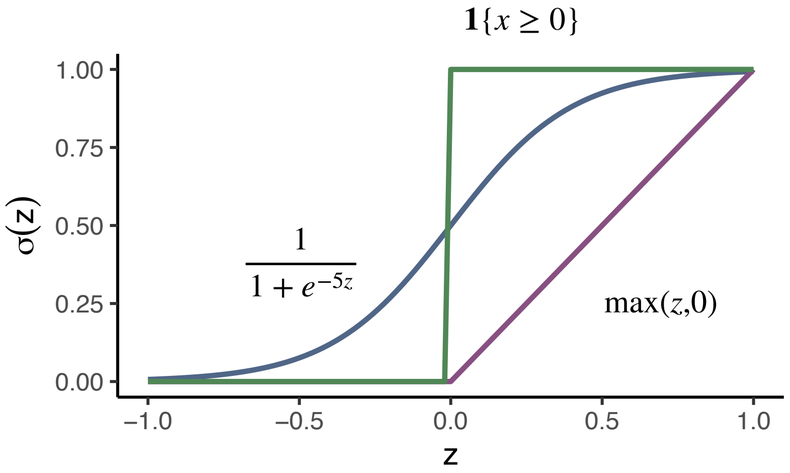


Camada de entrada

Camadas ocultas

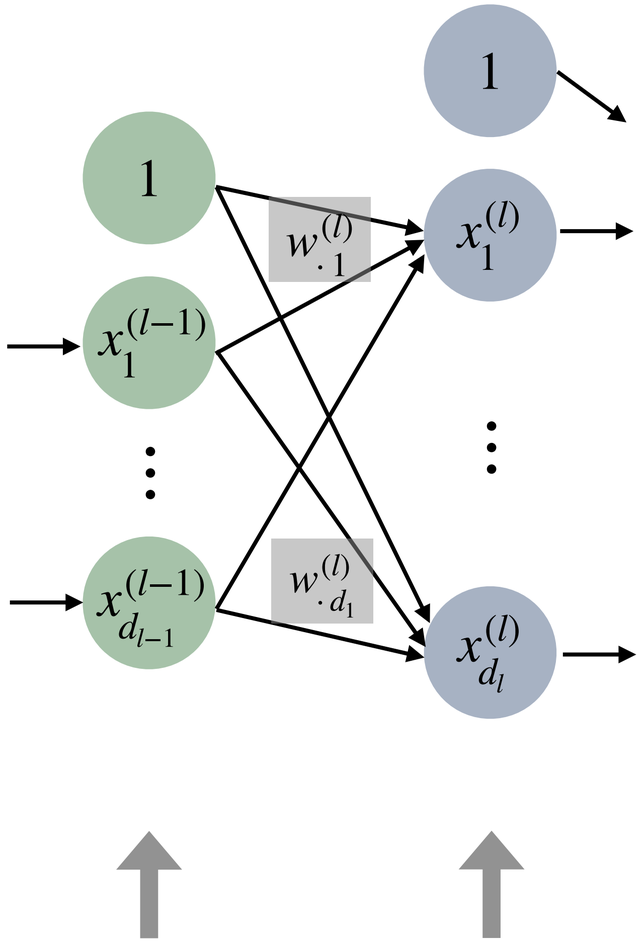
Camada de saída

$$x_j^{(l)} = \sigma(x^{(l-1),T} w_{\cdot j}^{(l)})$$



Funções de ativação

Rede neural de múltiplas camadas ocultas

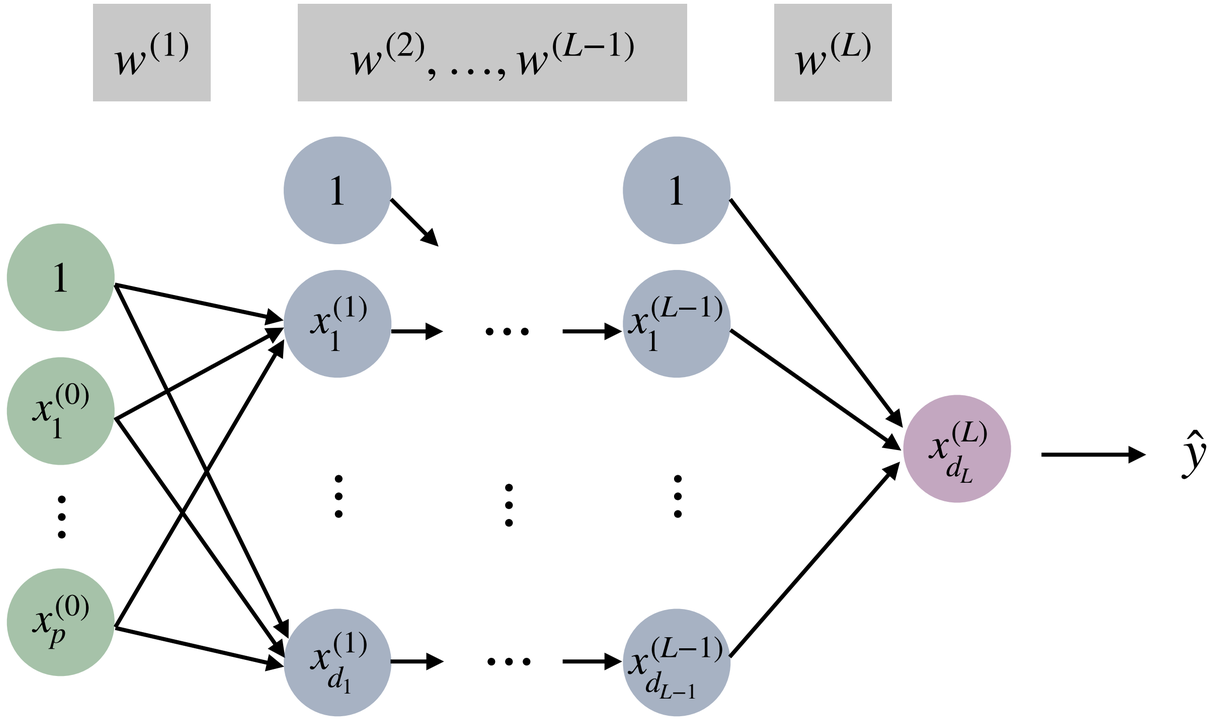


$$\left.\begin{aligned} z_j^{(l)} &= x^{(l-1),T} w_{\cdot j}^{(l)} \\ x_j^{(l)} &= \sigma(z_j^{(l)}) = \sigma(x^{(l-1),T} w_{\cdot j}^{(l)}) \end{aligned}\right\} \quad 1 \le j \le d_l$$

$$w_{ij}^{(l)} \quad \begin{cases} 1 \le l \le L & \text{camadas} \\ 0 \le i \le d_{l-1} & \text{entradas} \\ 1 \le j \le d_l & \text{saídas} \end{cases}$$

$$w^{(l)} \in \mathbb{R}^{(d_{l-1}+1) \times d_l}$$

Propagação para a frente e para atrás



$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \frac{\partial L}{\partial z_j^{(l)}} \times \frac{\partial z_j^{(l)}}{\partial w_{ij}^{(l)}}$$

$$\delta_j^{(l)} \qquad x_i^{(l-1)}$$

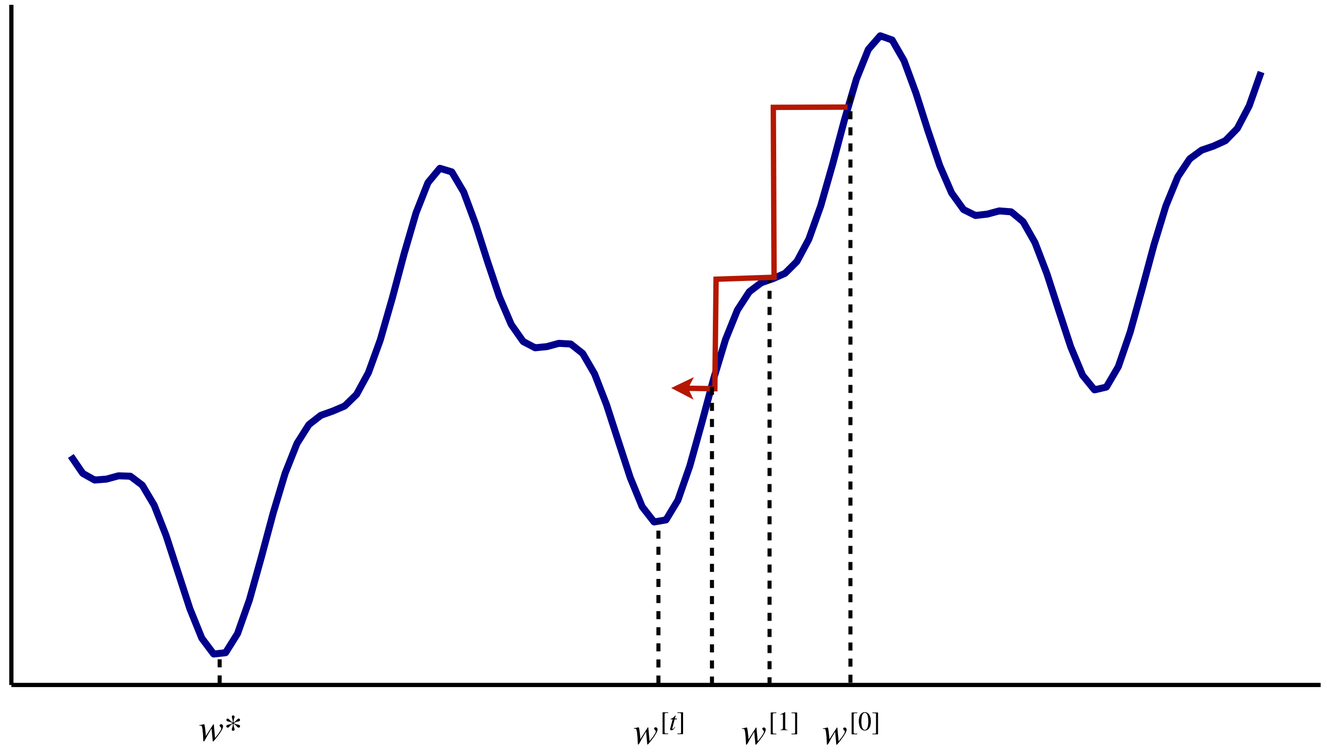
$$x = x^{(0)} \rightarrow x^{(1)} \cdots \rightarrow x^{(L-1)} \rightarrow x^{(L)} \rightarrow L(g(x), y)$$

$$\delta^{(1)} \leftarrow \cdots \delta^{(L-1)} \leftarrow \delta^{(L)} = \frac{\partial L}{\partial z_j^{(L)}}$$

$$\nabla \hat{L}(w) = \begin{pmatrix} \frac{\partial L}{\partial w_{01}^{(1)}}(w) \\ \vdots \\ \frac{\partial \hat{L}}{\partial w_{d_{L-1}\, d_L}^{(L)}}(w) \end{pmatrix}$$

Algoritmo descendente do gradiente

Objetivo: achar w que minimize $\widehat{E}_D(w) = \frac{1}{n}\sum_{i=1}^{n} L(g(x_i), y_i)$



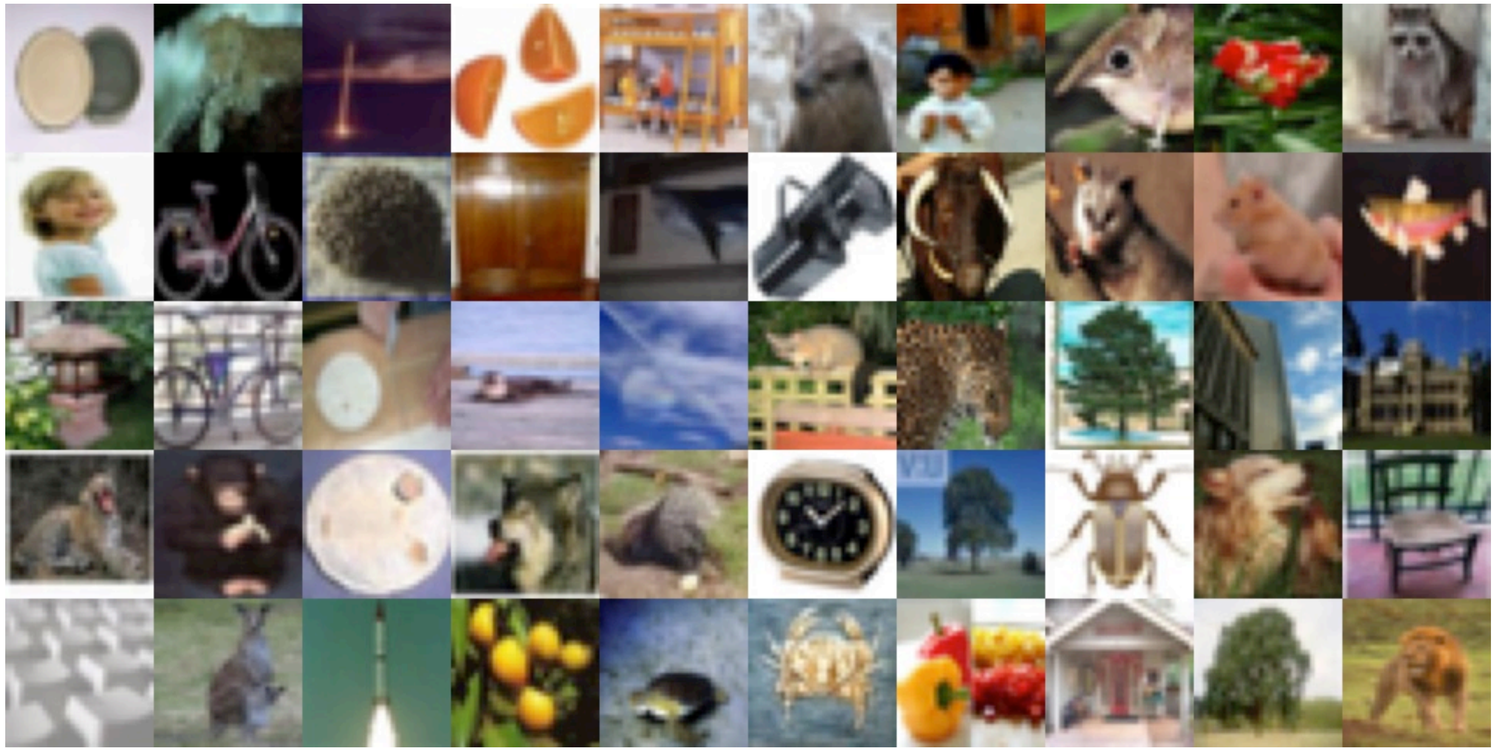
$w^{[0]} \sim \mathcal{N}(0,1)$ (inicialização)

$w^{[t]} = w^{[t-1]} - \eta \nabla \widehat{E}_D(w^{[t-1]})$ (iteração)

η (taxa de aprendizagem)

Redes neurais convolucionais

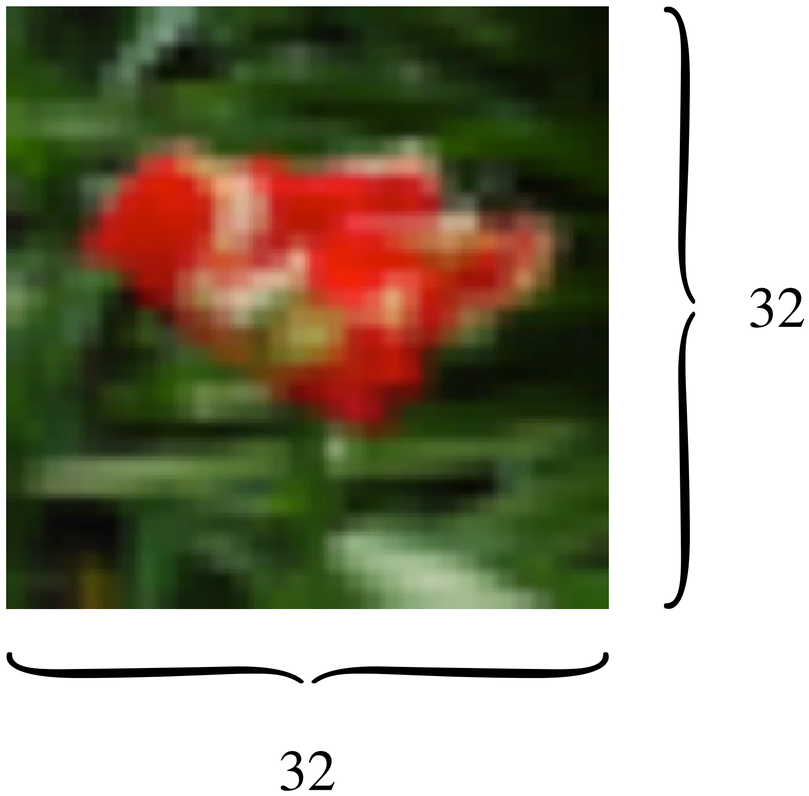
- Redes neurais convolucionais são redes especificamente desenhadas para análise (classificação) de imagens
- Elas são construídas com base em transformações específicas das imagens
- As principais transformações são as *convoluções,* que são combinadas com outras transformações para reduzir o tamanho das imagens, chamadas de *pooling*

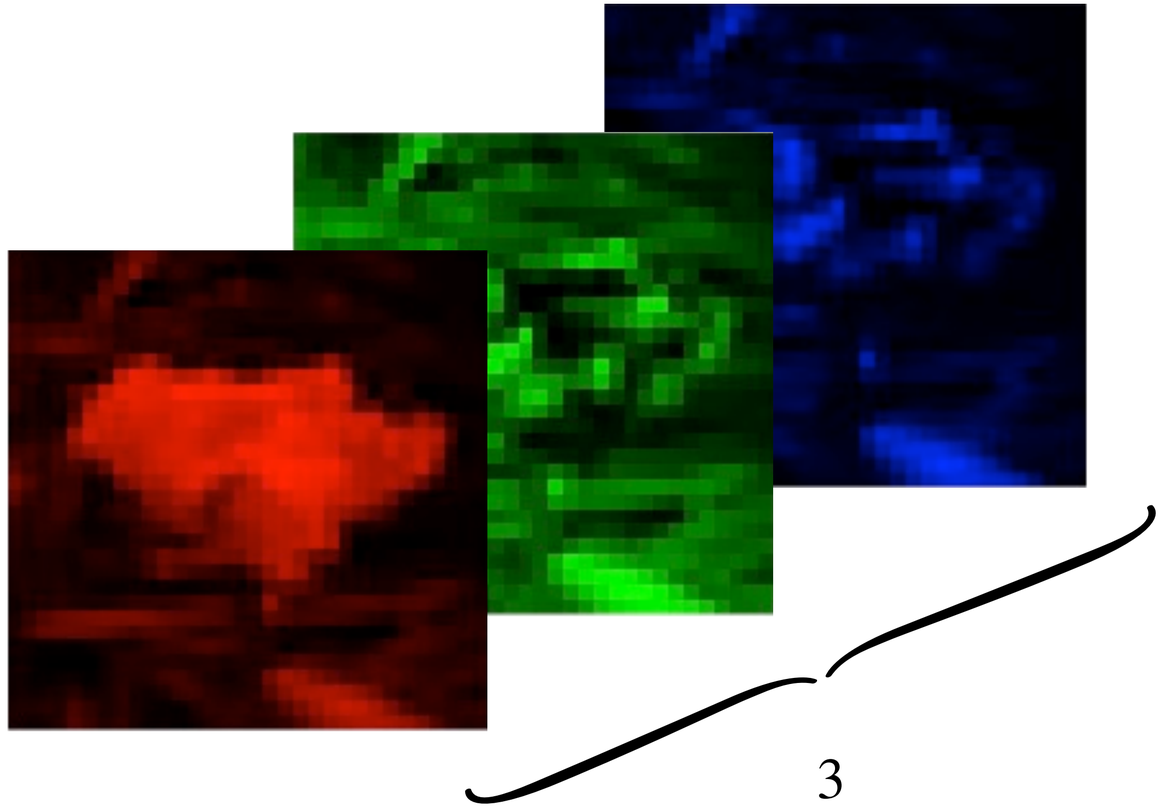


Base de dados `CIFAR100`, com 50.000 imagens de treinamento e 10.000 de teste

As imagens tem dimensão 32 x 32 x 3 e estão classificadas em 20 *superclasses*

Cada *superclasse* está dividida ainda em 5 classes.


32
32


3

Convolução

$$\text{Imagem original} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \\ j & k & l \end{bmatrix}$$

$$\text{Filtro de convolução} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$$

$$\text{Imagem convolucionada} = \begin{bmatrix} a\alpha + b\beta + d\gamma + e\delta & b\alpha + c\beta + e\gamma + f\delta \\ d\alpha + e\beta + g\gamma + h\delta & e\alpha + f\beta + h\gamma + i\delta \\ g\alpha + h\beta + j\gamma + k\delta & h\alpha + i\beta + k\gamma + l\delta \end{bmatrix}$$

Convolução

$$\text{Imagem original} = \begin{bmatrix} a & a & a & 0 & 0 & 0 \\ a & a & a & 0 & 0 & 0 \\ a & a & a & 0 & 0 & 0 \\ a & a & a & 0 & 0 & 0 \end{bmatrix}$$

$$\text{Filtro de convolução} = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

$$\text{Imagem convolucionada} = \begin{bmatrix} 0 & 3a & 3a & 0 \\ 0 & 3a & 3a & 0 \end{bmatrix}$$

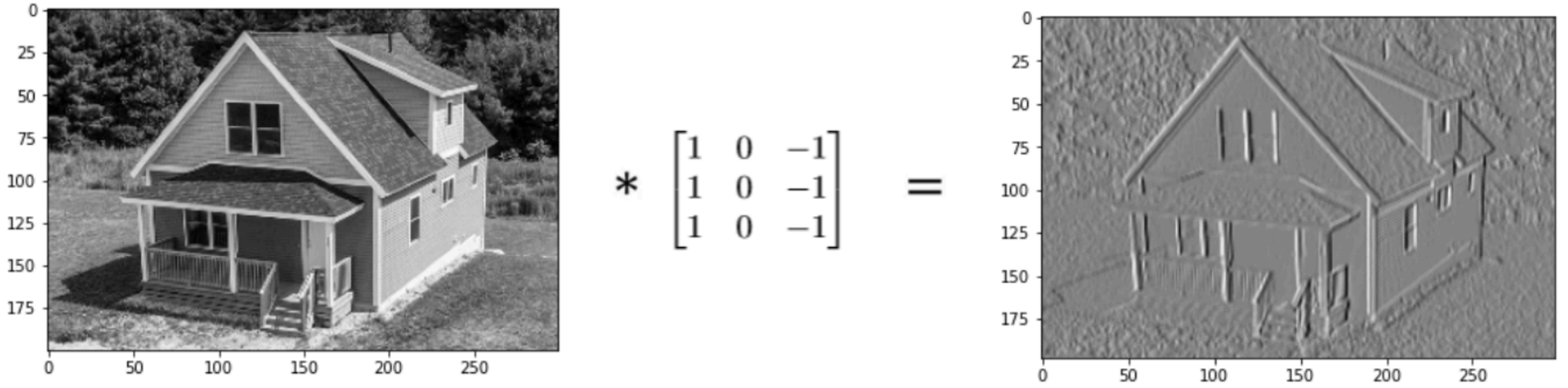


Imagem do TCC do aluno do IME-USP Diego Marques

Convolução

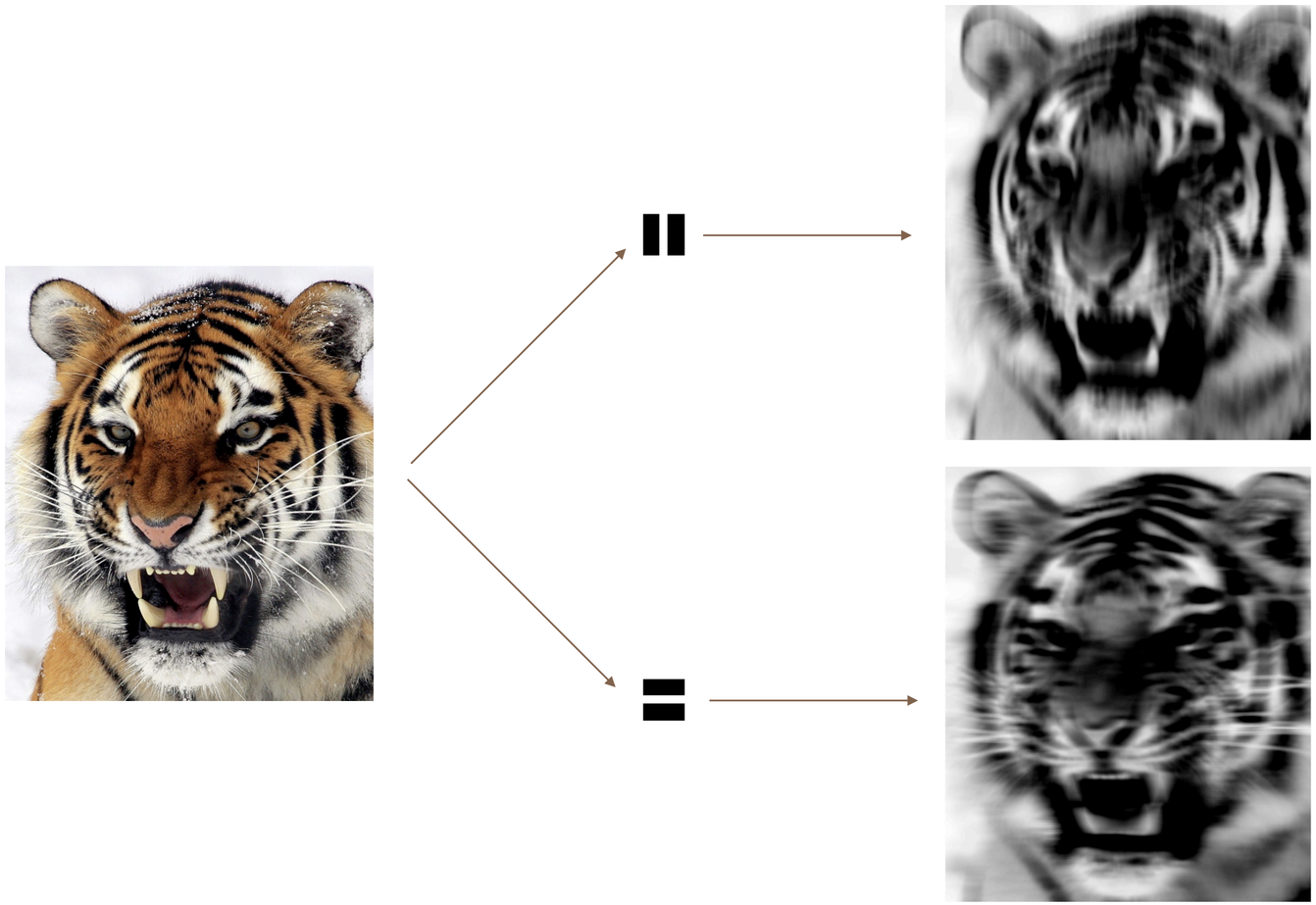


Imagem do livro *An Introduction to Statistical Learning*

Padding

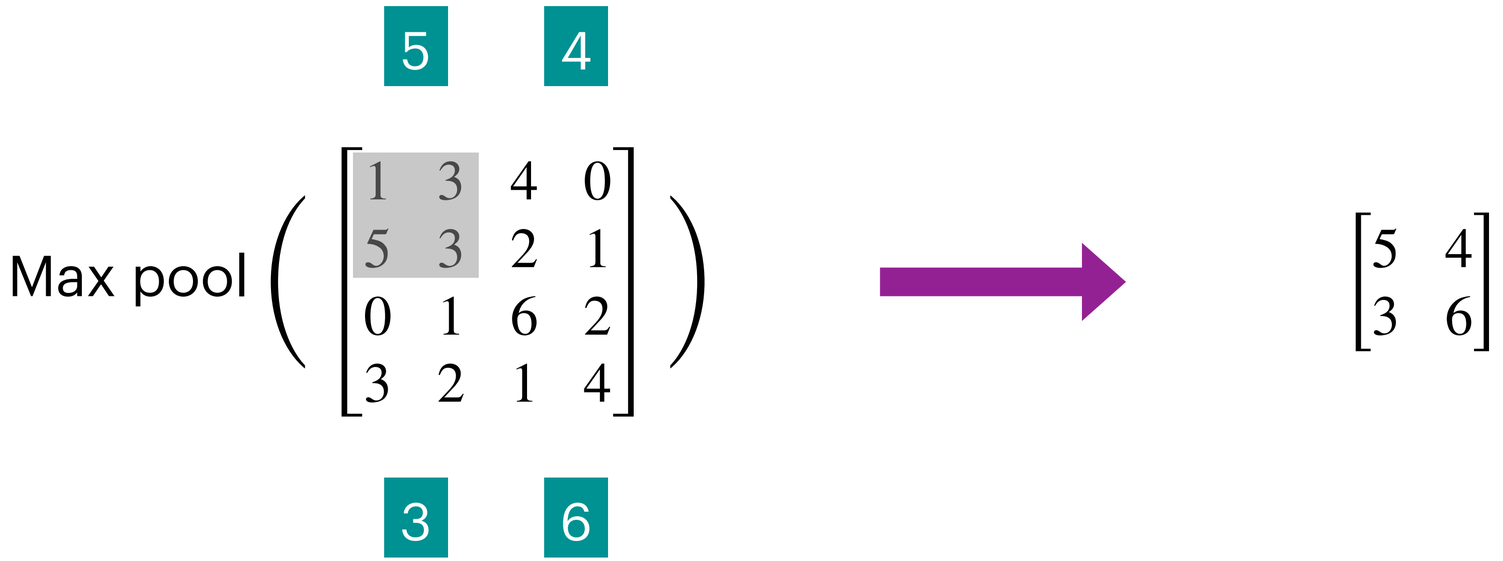
Para que a imagem convolucionada mantenha a mesma dimensão da imagem original, muitas vezes é realizada uma operação de *padding,* que consiste em adicionar uma borda de 0's ao redor da imagem de entrada para depois aplicar a convolução

$$\text{Imagem original} = \begin{bmatrix} a & a & a & 0 & 0 & 0 \\ a & a & a & 0 & 0 & 0 \\ a & a & a & 0 & 0 & 0 \\ a & a & a & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\text{Padding}} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a & a & a & 0 & 0 & 0 & 0 \\ 0 & a & a & a & 0 & 0 & 0 & 0 \\ 0 & a & a & a & 0 & 0 & 0 & 0 \\ 0 & a & a & a & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\text{Filtro de convolução} = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \quad \text{Imagem convolucionada} = \begin{bmatrix} -2a & 0 & 2a & 2a & 0 & 0 \\ -3a & 0 & 3a & 3a & 0 & 0 \\ -3a & 0 & 3a & 3a & 0 & 0 \\ -2a & 0 & 2a & 2a & 0 & 0 \end{bmatrix}$$

Pooling

Esta operação tem como objetivo reduzir a informação na camada de entrada e produzir uma camada com dimensão menor



Arquitetura de uma rede neural convolucional

Redes neurais convolucionais são redes especificamente desenhadas para análise (classificação) de imagens.

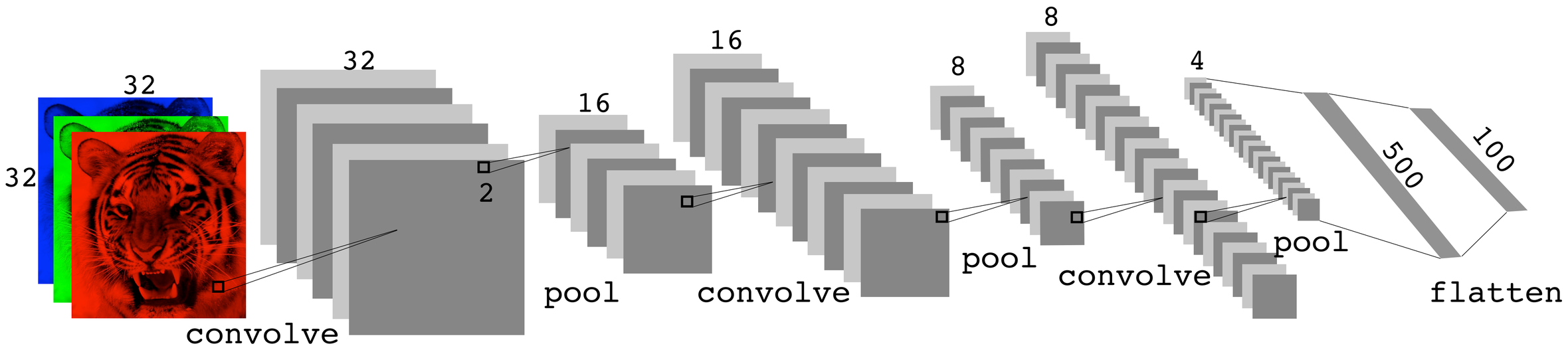
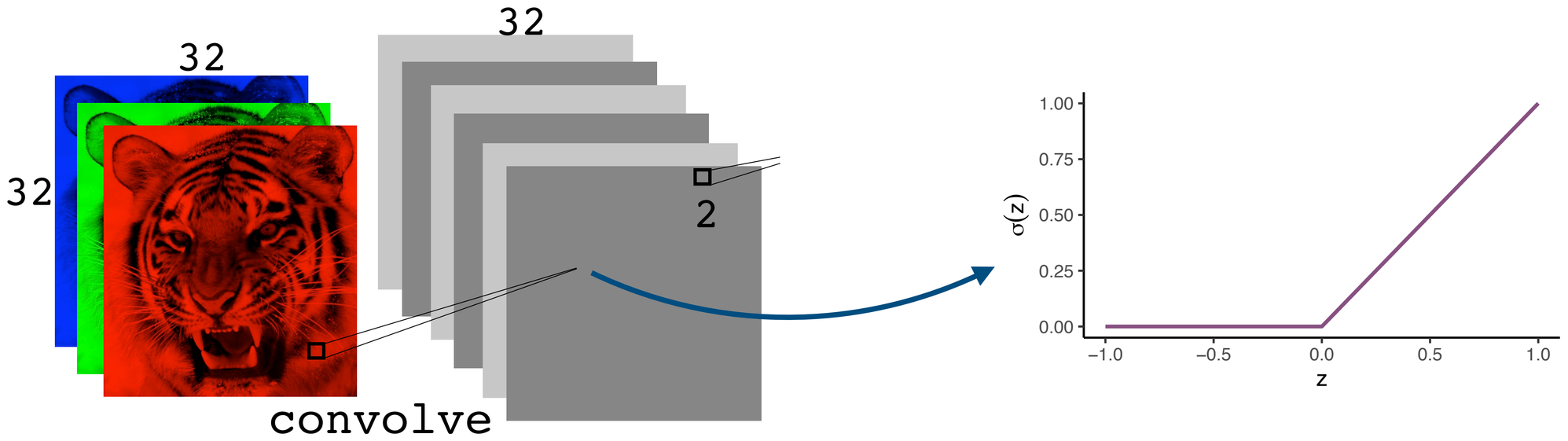


Imagem do livro *An Introduction to Statistical Learning*

Rede neural convolucional

Como as imagens tem três canais (um por cor) em geral cada filtros de convolução também tem três canais (podem ser iguais ou diferentes) e o resultado se soma



Em geral, após realizar a convolução é aplicada a função de ativação ReLU a cada entrada da imagem convolucionada

Imagem do livro *An Introduction to Statistical Learning*

Rede neural convolucional

Assim é obtida a primeira camada oculta nesta rede, que é constituída de K imagens, obtidas pela utilização de K filtros de convolução 3×3, criando um "mapa de atributos" (*feature map*) de dimensão $32 \times 32 \times K$ (na figura $K = 6$)

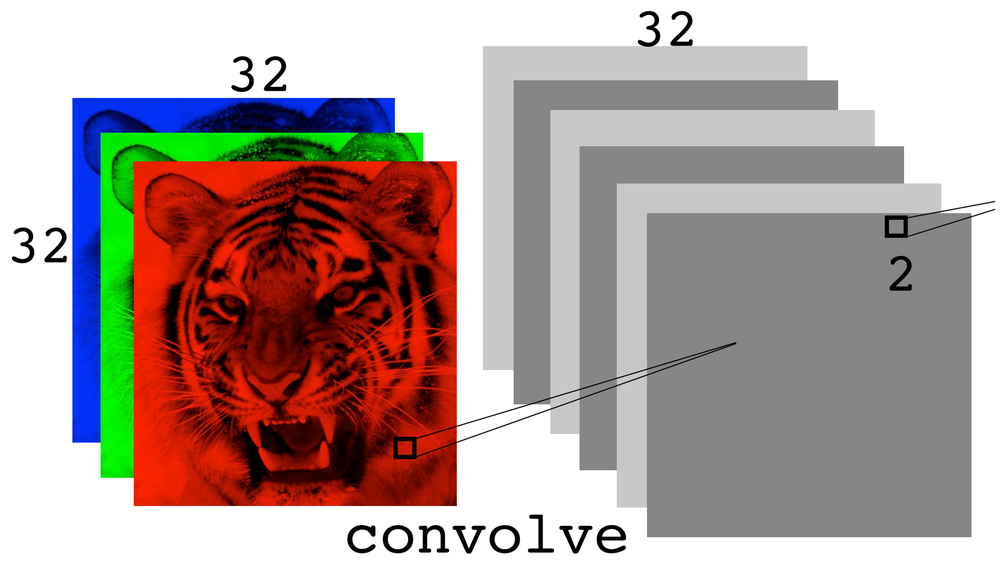


Imagem do livro *An Introduction to Statistical Learning*

Rede neural convolucional

As segunda camada oculta é obtida com a operação de *pooling* com um filtro 2×2

que reduz a dimensão do mapa de atributos para $16 \times 16 \times K$

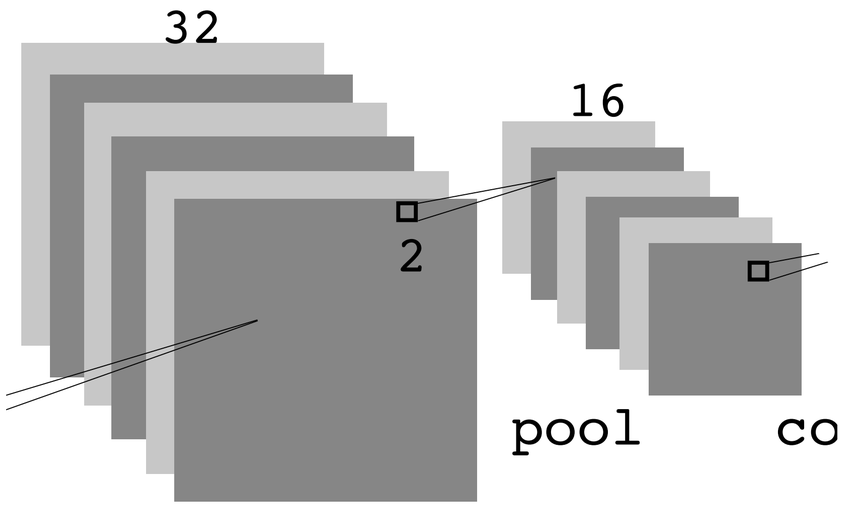


Imagem do livro *An Introduction to Statistical Learning*

Arquitetura de uma rede neural convolucional

As operações de convolução e pooling são iteradas duas vezes mais nesta rede

Finalmente, o mapa de atributos é “empilhado" em camadas totalmente conectadas (esta operação é chamada de *flatten* e corresponde às últimas duas camadas desta rede). A última camada usa a ativação *softmax* para as 100 classes de imagens

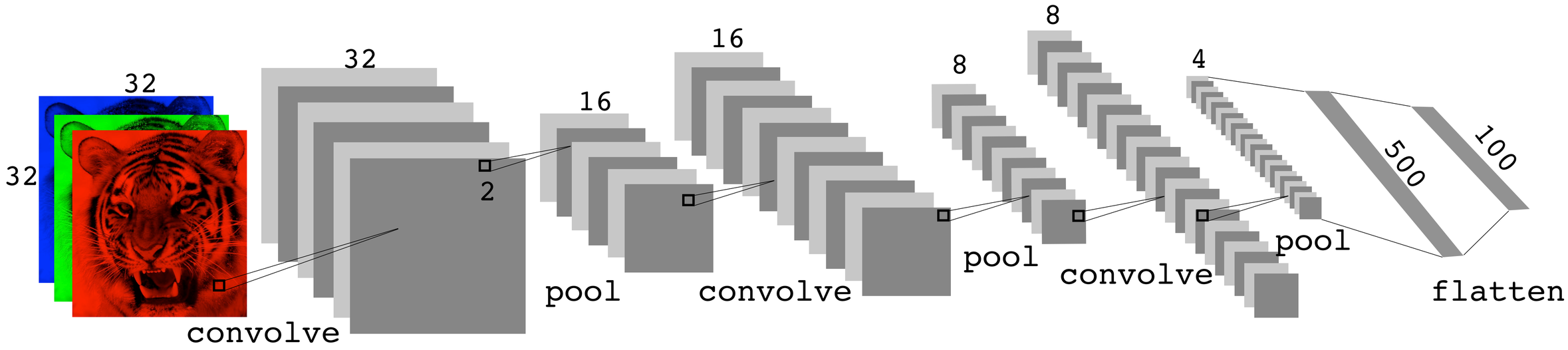


Imagem do livro *An Introduction to Statistical Learning*

Redes neurais recorrentes

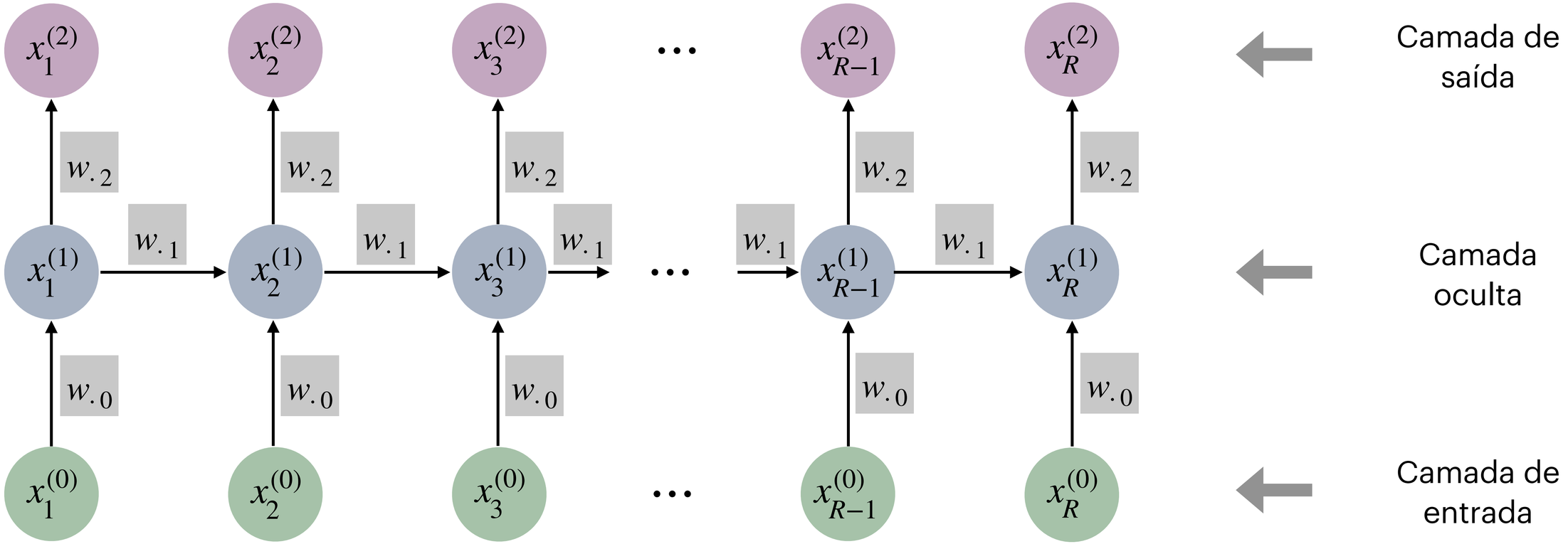
- Estas redes são especificamente desenhadas para dados sequenciais
- Numa rede neural recorrente, as covariáveis de entrada x são sequencias, isto é $x = (x_1, \ldots, x_l)$, onde cada x_i pode ser uma variável ou um vetor
- Neste caso, a ordem das coordenadas é importante e portanto a sequencia deve ser tratada como uma unidade
- A variável resposta pode também ser uma sequência, ou pode ser uma etiqueta como nos problemas de classificação

Redes neurais recorrentes

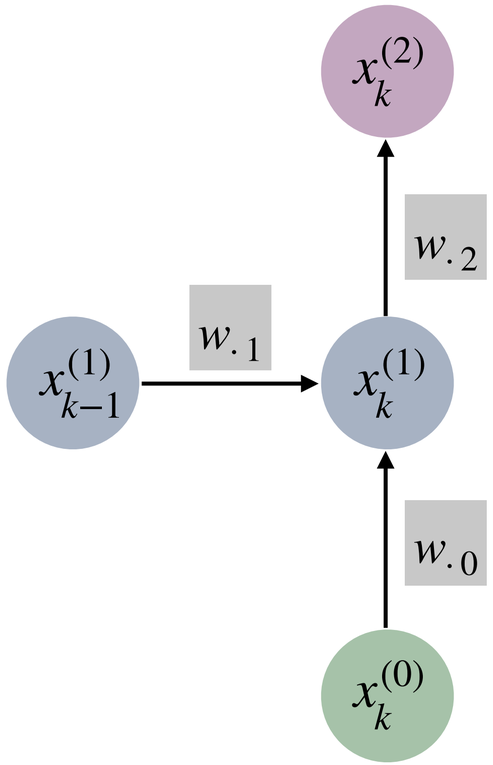
Alguns exemplos de aplicação das redes neurais recorrentes são

- Documentos escritos como livros ou avaliações de filmes, artigos de jornais, e postagens em redes sociais
- Séries temporais de temperatura, chuva, velocidade do vento, qualidade do ar, etc
- Séries temporais financeiras: índices de ações, taxas de câmbio, etc
- Gravações de voz, música e outros arquivos de áudio
- Textos manuscritos

Estrutura das redes neurais recorrentes



Redes neurais recorrentes

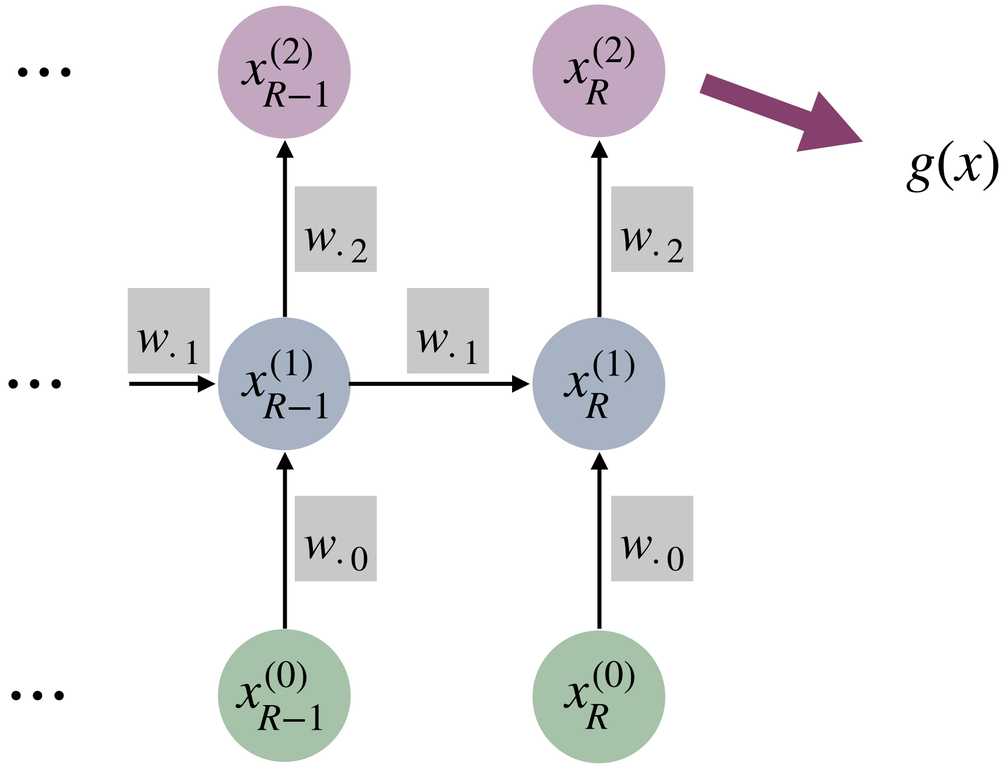


$$x_k^{(2)} = z_k^{(2)}$$

$$z_k^{(2)} = x^{(1),T} w_{\cdot 2}$$

$$\left.\begin{aligned} x_{ks}^{(1)} &= \sigma(z_{ks}^{(1)}) = \sigma(x_k^{(0),T} w_{s\,0} + x_{k-1}^{(1),T} w_{s\,1}) \\ z_{ks}^{(1)} &= x_k^{(0),T} w_{s\,0} + x_{k-1}^{(1),T} w_{s\,1} \end{aligned}\right\} \quad 1 \leq s \leq S$$

Redes neurais recorrentes



$$\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Encontrar $w_{\cdot,0}$, $w_{\cdot,1}$ e $w_{\cdot,2}$ que minimizem

$$\widehat{E}_D(w) = \frac{1}{n}\sum_{i=1}^{n} L(g(x_i), y_i)$$

Exemplo: análise de avaliações de filmes

This has to be one of the worst films of the 1990s. When my friends & I were watching this film (being the target audience it was aimed at) we just sat & watched the first half an hour with our jaws touching the floor at how bad it really was. The rest of the time, everyone else in the theater just started talking to each other, leaving or generally crying into their popcorn . . .

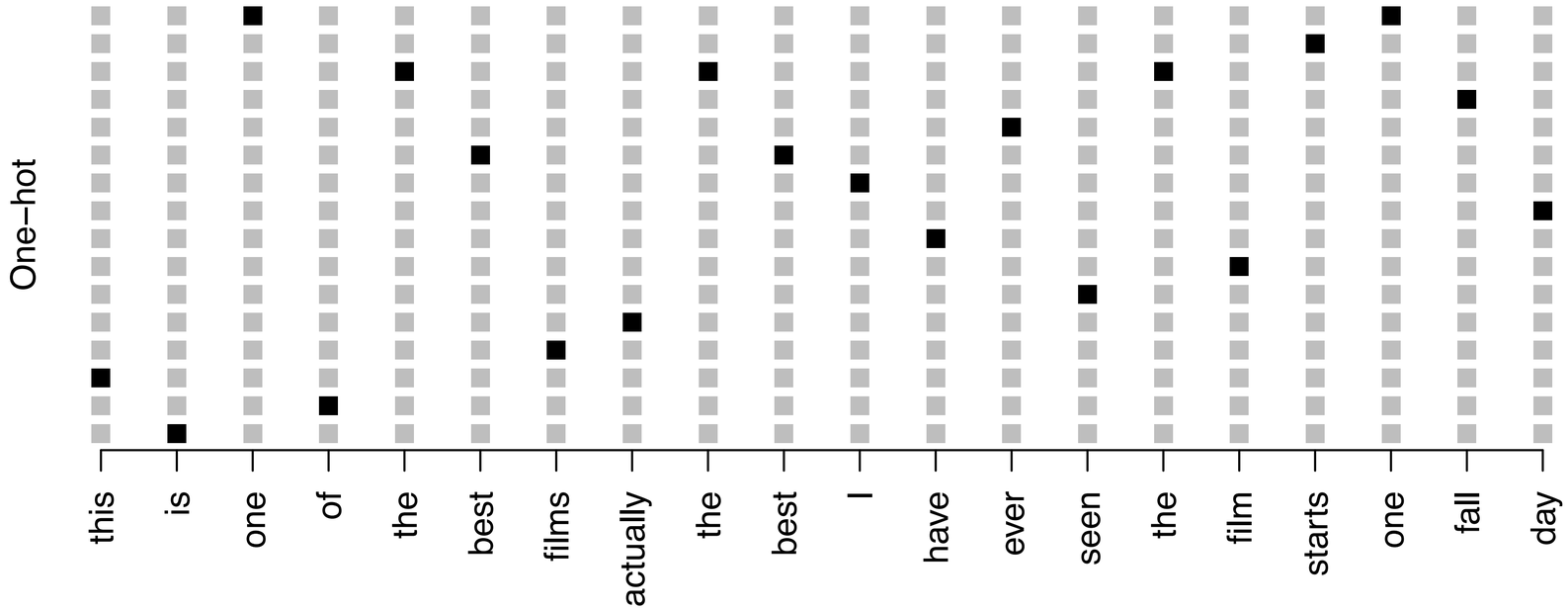
⟨START⟩ this film was just brilliant casting location scenery story direction everyone's really suited the part they played and you could just imagine being there robert ⟨UNK⟩ is an amazing actor and now the same being director ⟨UNK⟩ father came from the same scottish island as myself so i loved . . .

Exemplo do livro *An Introduction to Statistical Learning*

Exemplo: análise de avaliações de filmes

O modelo mais simples na análise de textos consiste em representar cada documento como um vetor de 0's e 1's com dimensão igual ao tamanho de um dicionário (no exemplo, um dicionário de palavras em inglês com 10.000 palavras)

Cada entrada do vetor é a função indicadora de a palavra estar presente no texto



Exemplo: análise de avaliações de filmes

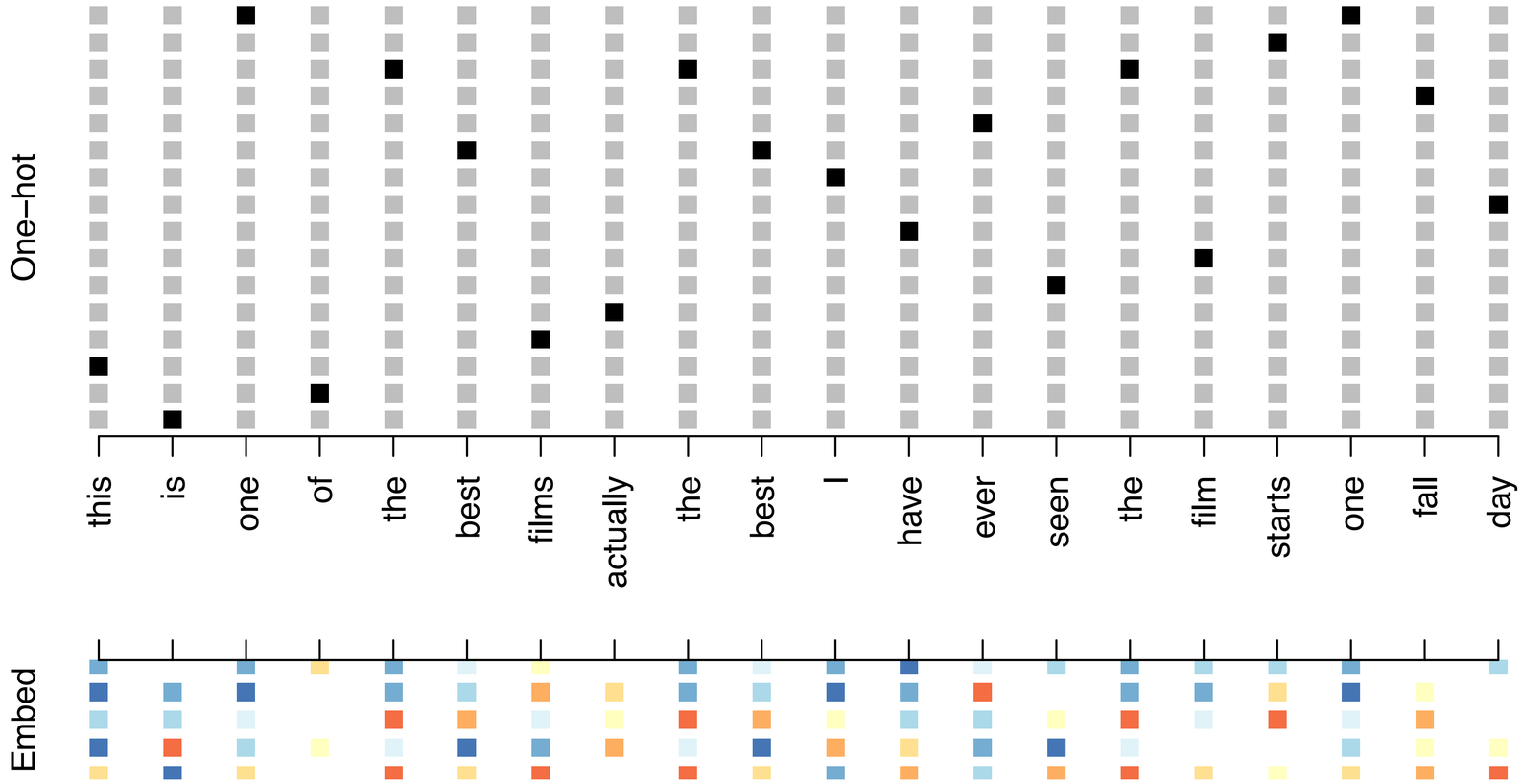
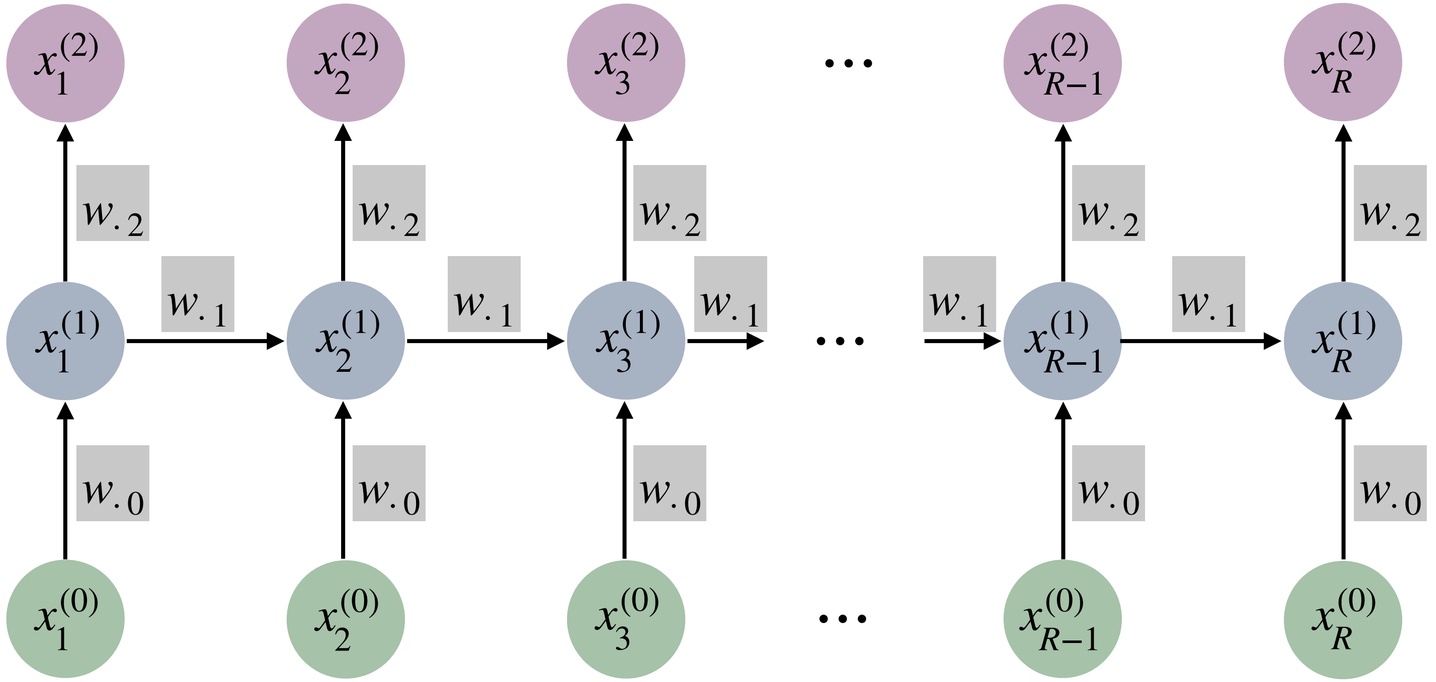


Imagem do livro *An Introduction to Statistical Learning*

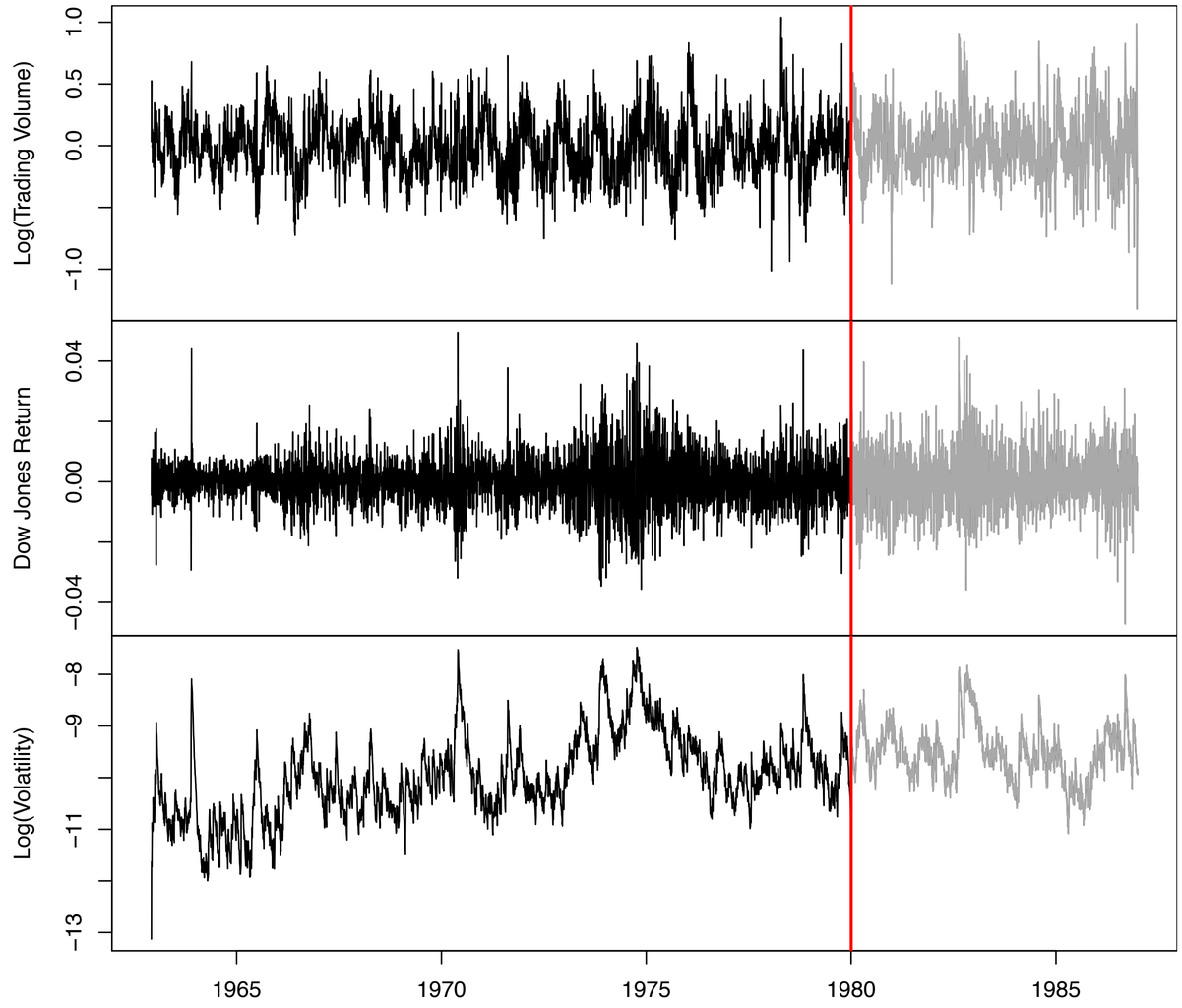
Exemplo: análise de avaliações de filmes



A aplicação desta rede na base de dados `IMDb` de avaliações de filmes para classificá-los como positivos ou negativos alcançou uma acurácia de 76% na base de teste

A maior marca até agora teve uma acurácia de 97,4%, usando modificações de Naive Bayes!

Exemplo: predição em séries temporais



Dados de três séries temporais da bolsa de valores de Nova Iorque, correspondentes ao período Dezembro 1962 a Dezembro 1986

- Log trading volume
- Dow Jones return
- Log volatility

O objetivo neste caso é predizer a futuro a primeira variável em função do histórico das séries

Imagem do livro *An Introduction to Statistical Learning*

Exemplo: predição em séries temporais

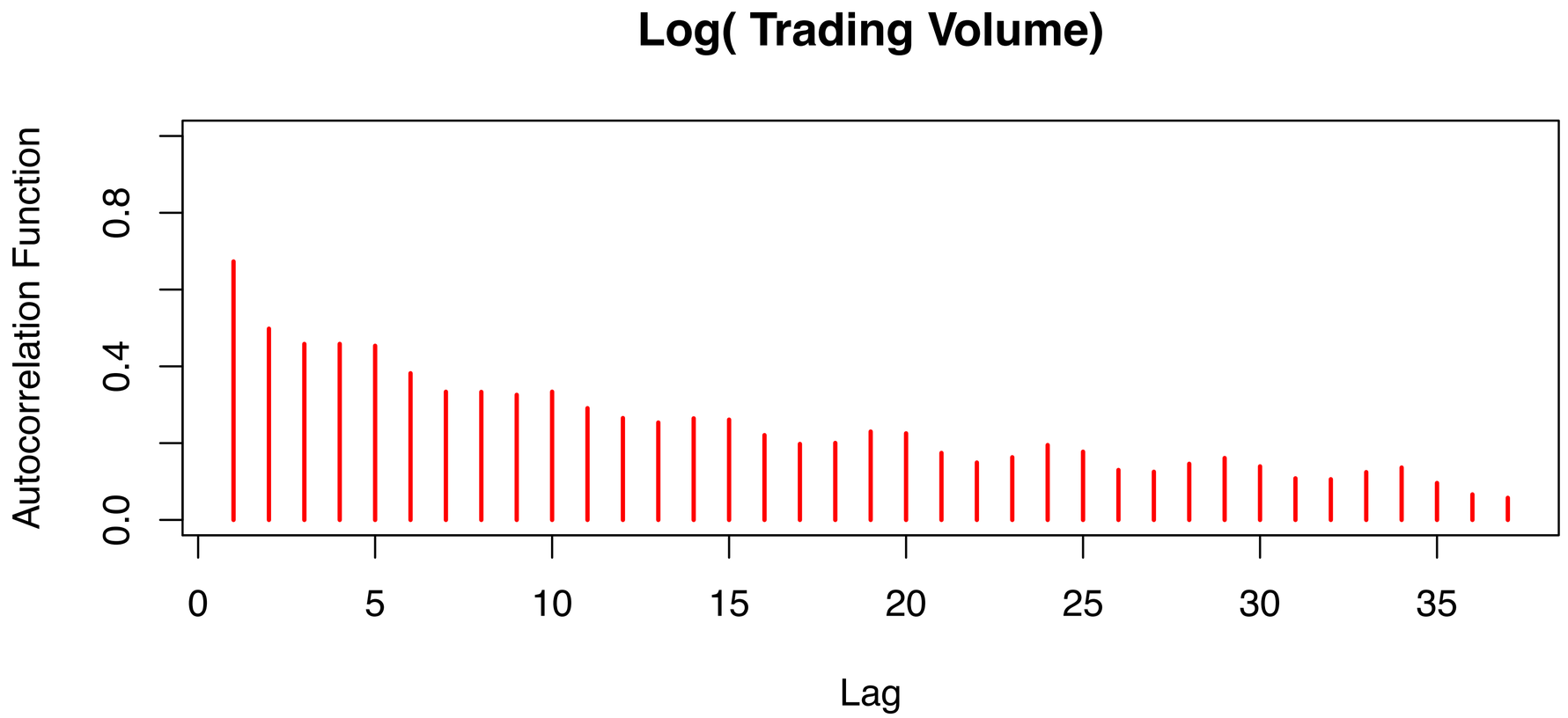
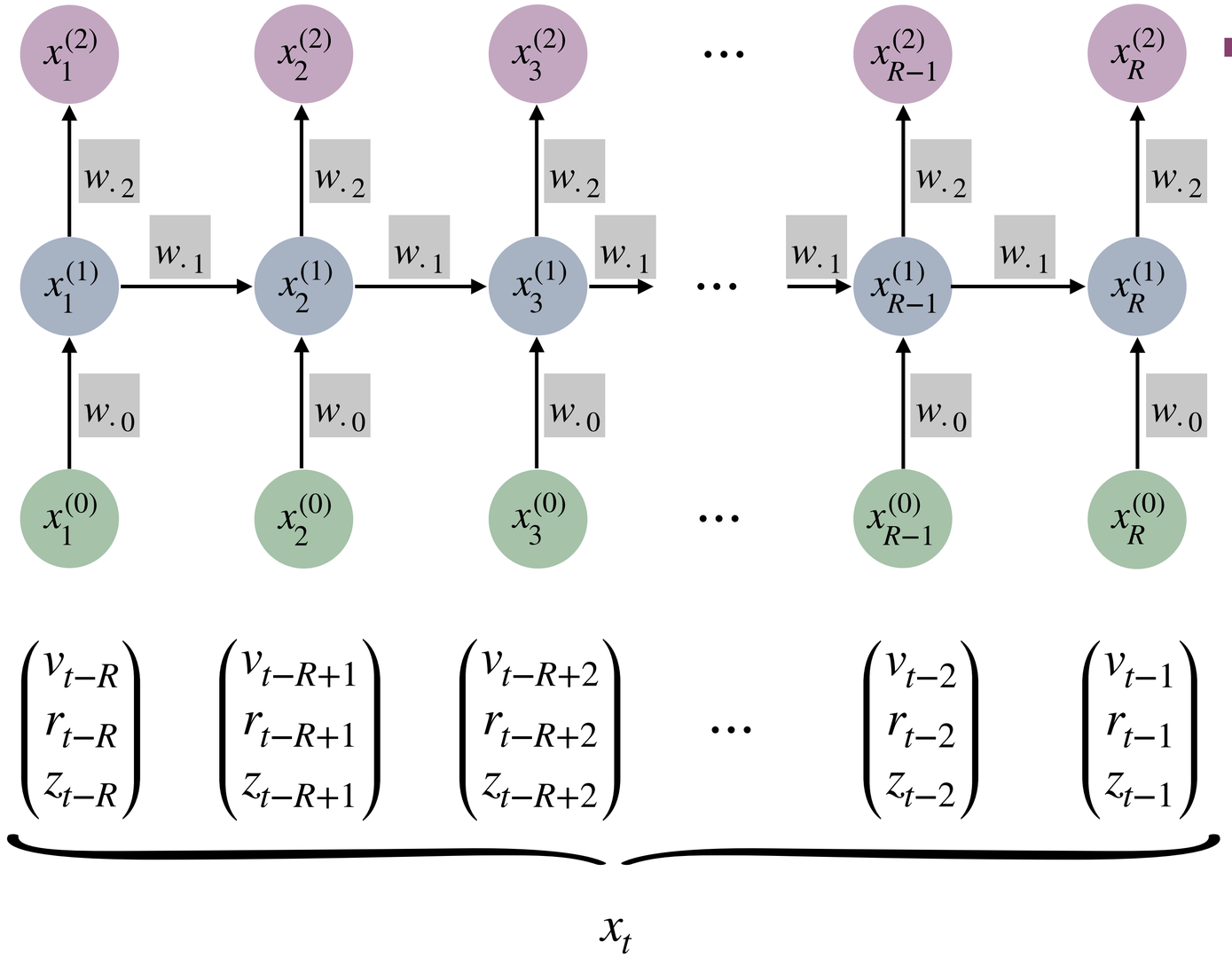


Imagem do livro *An Introduction to Statistical Learning*

Exemplo: predição em séries temporais



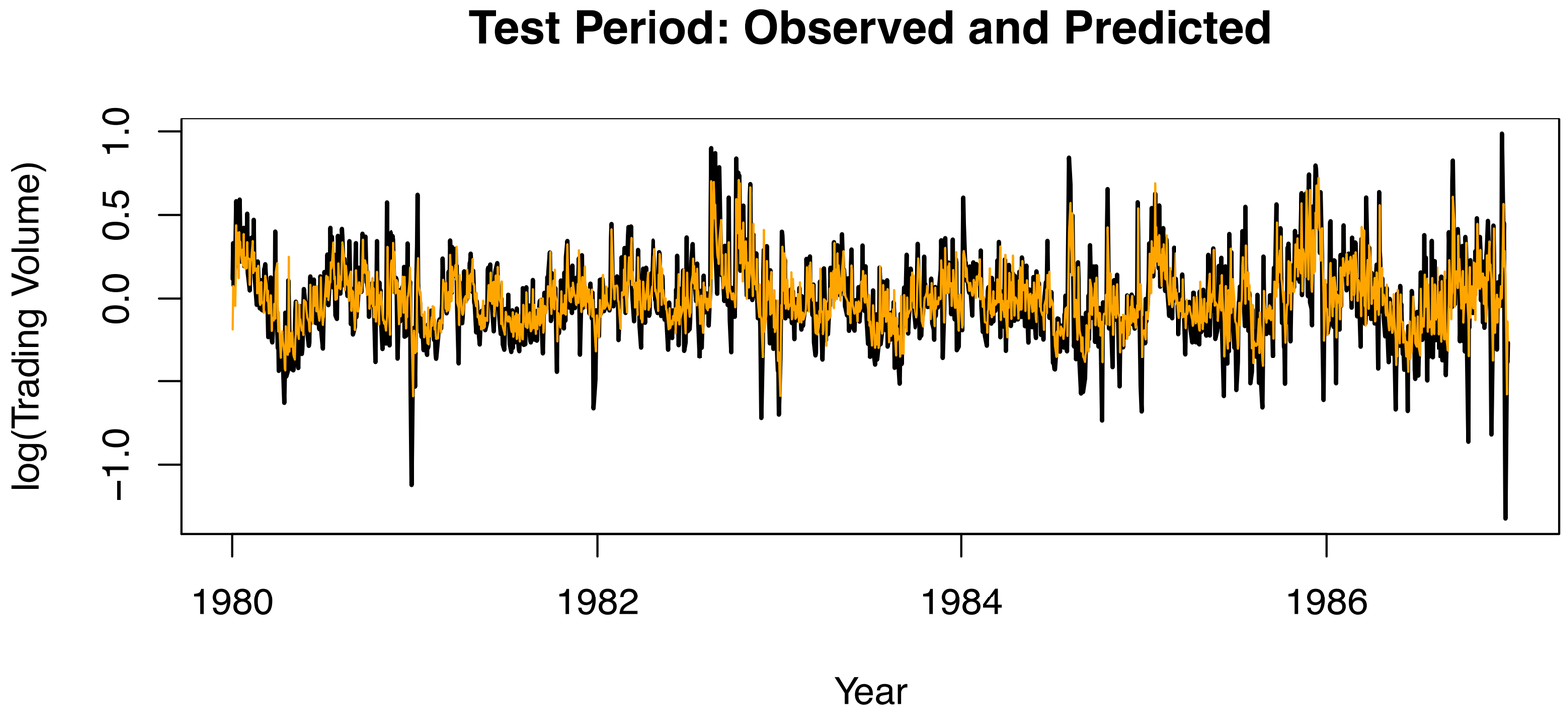
$y_t = v_t$

v_t: Log trading volume no tempo t

r_t: Dow Jones return no tempo t

z_t: Log volatility no tempo t

Exemplo: predição em séries temporais



Resultados da predição da variável Log trading volume no período após 2 de janeiro de 1980 como função das três séries em 5 dias anteriores ($R = 5$), usando uma rede neural com 12 unidades em cada nó oculto da rede neural recursiva

Imagem do livro *An Introduction to Statistical Learning*