Fundamentos de Deep Learning para Processamento de Linguagem Natural Parte 2

Dr. Cristian E. Muñoz Villalobos

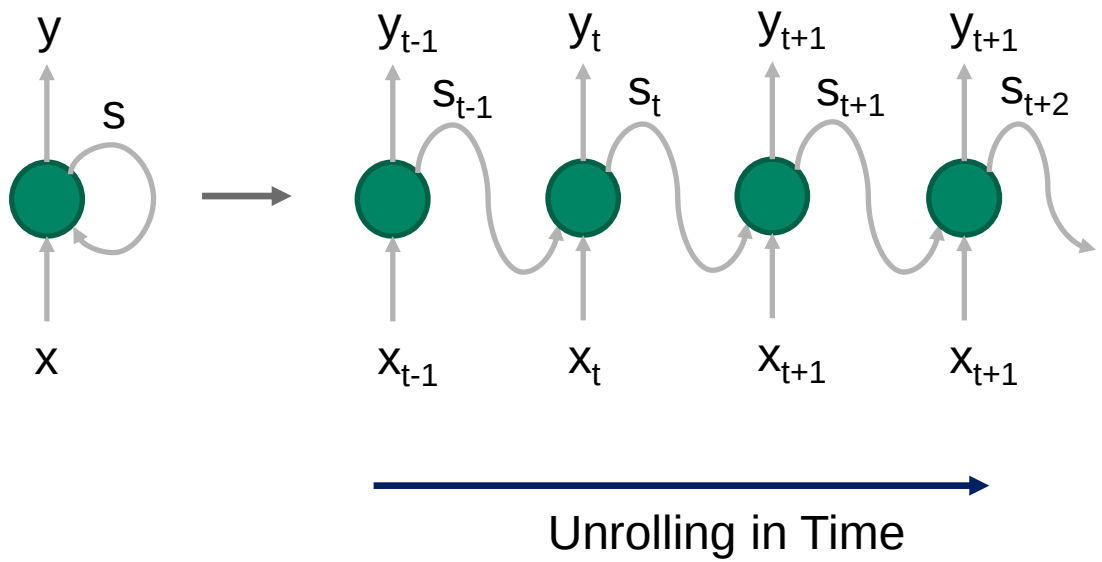
TEXT CLASSIFICATION

SUMÁRIO

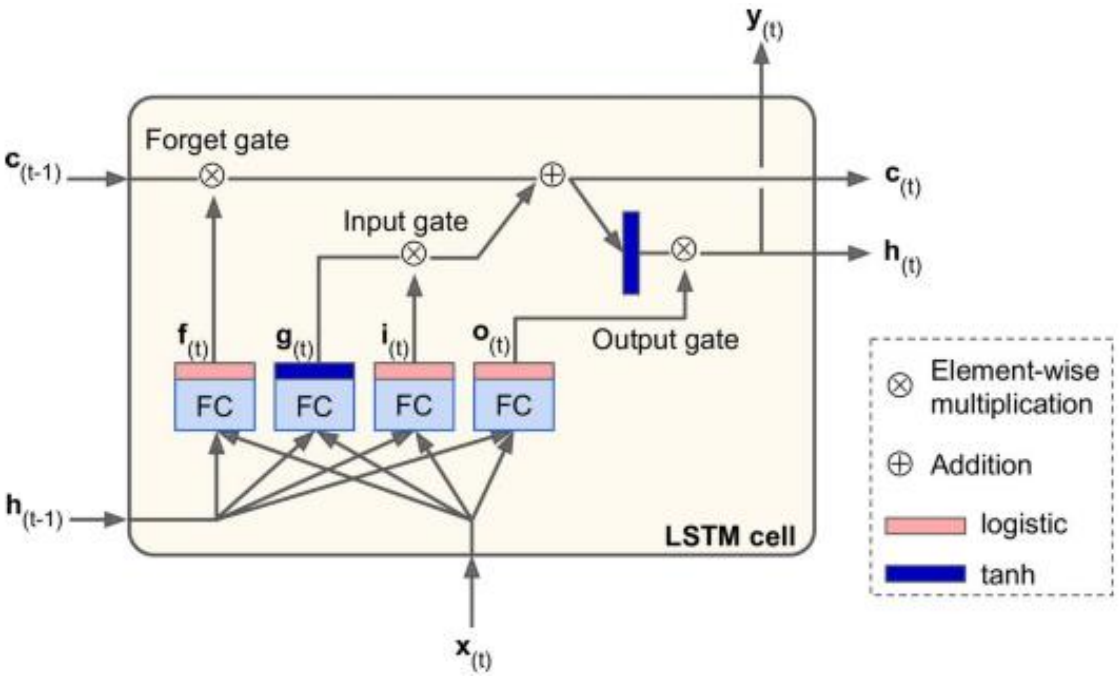
1. Redes Neurais Recorrentes e LSTMs
2. O Problema de Atribuição de Autoria
3. The Federalist Papers Lab - Suposições e Arquitetura da Rede

https://github.com/crismunoz/NLP_Course/tree/master/3.%20Fundamentals%20of%20Deep%20Learning%20for%20Natural%20Language%20Processing%20(Portuguese)

NEURONIO RECORRENTE



LONG SHORT TERM (LSTM)



$$
\begin{aligned}
\mathbf{i}_{(t)} &= \sigma(\mathbf{W}_{xi}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hi}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{i}) \\
\mathbf{f}_{(t)} &= \sigma(\mathbf{W}_{xf}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hf}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{f}) \\
\mathbf{o}_{(t)} &= \sigma(\mathbf{W}_{xo}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{ho}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{o}) \\
\mathbf{g}_{(t)} &= \tanh(\mathbf{W}_{xg}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hg}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{g}) \\
\mathbf{c}_{(t)} &= \mathbf{f}_{(t)} \otimes \mathbf{c}_{(t-1)} + \mathbf{i}_{(t)} \otimes \mathbf{g}_{(t)} \\
\mathbf{y}_{(t)} &= \mathbf{h}_{(t)} = \mathbf{o}_{(t)} \otimes \tanh(\mathbf{c}_{(t)})
\end{aligned}
$$

ATRIBUIÇÃO DE AUTORIA

- Dado um conjunto de documentos e um conjunto de autores, quais autores escreveram quais documentos?
- Precisamos aprender o "estilo" de cada autor: palavras e frases comuns, comprimento da frase, uso de pontuação, etc. Isso é conhecido como a ciência da estilometria.

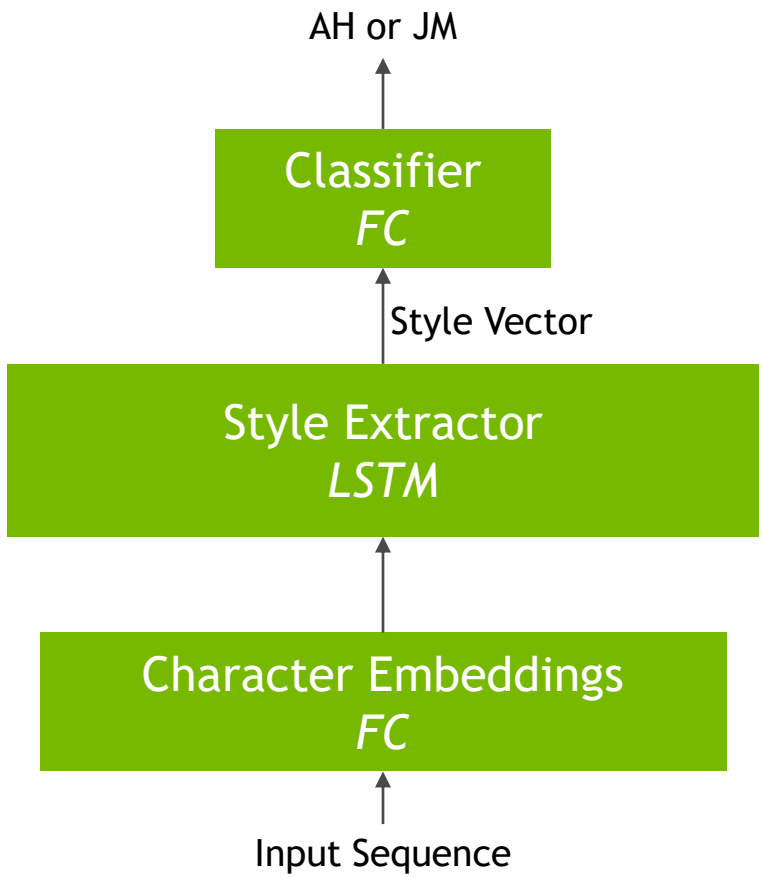
THE FEDERALIST PAPERS

- Conjunto de 85 documentos escritos entre 1787 e 1788 por Alexander Hamilton, James Madison e John Jay sob um pseudônimo e com o objetivo de promover a ratificação da Constituição dos Estados Unidos.
- A autoria de 12 desses artigos é disputada entre Hamilton e Madison
- Tentaremos usar o Deep Learning para resolver a questão!

SUPOSIÇÕES

- Apenas um único autor por documento - Hamilton ou Madison
- O estilo do autor é consistente em vários documentos

ARQUITETURA DA REDE



CONCLUSÕES

- RNNs podem ser usados para treinamento em sequências - LSTMs fornecem uma variante fácil de treinar.
- Usamos LSTMs para aprender e codificar o estilo de sequências de caracteres dos documentos federalistas (“The Federalist Papers”)
- Determinamos o autor de um documento determinando qual autor escreveu mais sequências em um documento disputado de Artigos Federalistas