

MAC5725 - Linguística Computacional

Relatorio EP2

Andre Barbosa-nUSP 7971751

30 de outubro de 2020

1 Configurações dos Experimentos

Para todos os experimentos utilizados, foram mantidas a mesma arquitetura geral da rede, otimizador, e hiperparâmetros.

Em relação à rede, com exceção do desejado de variações de arquitetura (uso ou não de redes bidirecionais e presença/ausência de taxa de *dropout*), as demais configurações foram **exatamente as mesmas**, destacando-se pelas seguinte camadas, por ordem sequencial:

1. Um *input layer* que recebe um texto de entrada
2. Um *text vectorization layer*, que pré processa o texto de entrada, tokenizando e fazendo os pré processamentos necessários, como mapear as palavras para inteiros, eliminar excesso de espaços e remover acentos
3. Cada palavra é mapeada para uma matriz de Embeddings pré treinada. No caso, foi utilizada a matriz do NILC [Har+17] com 50 dimensões. É válido dizer que as palavras do vocabulário que estavam corrompidas de alguma forma foram excluídas.
4. A camada a se avaliar o experimento (LSTM unidirecional ou bidirecional)
5. Uma taxa de aprovação de dropout (um valor entre [0, 0.25 , 0.5])
6. Uma camada densa, mapeando para cada uma das classes (0 a 5)

Como critério de parada, utilizou-se o *callback* de *EarlyStopping* do *Keras* [Cho+15] com *patience=1*. Em outras palavras, se a *loss* de validação apresentar piora em **uma** avaliação, o treinamento é encerrado. Além disso, as escolhas para proporção de treino, teste e validação foram, respectivamente, 0.8, 0.1 e 0.1 e o tamanho de sentença máximo de sentença (*tammax*) foi de 256.

Além disso, em relação ao otimizador a função de *loss* aplicada foi a *Sparse Categorical Cross Entropy*, uma vez que as classes foram implementadas como a codificação de inteiro e não *dummy*. O tamanho de batch escolhido foi de 32 e o otimizador foi o Adam [KB17]. Por fim, o *learning rate* escolhido foi,

inicialmente de 0.001 , mas optou-se por utilizar uma queda do valor da taxa de aprendizado conforme o tempo passava. A taxa de queda foi de 0.5 a cada 50 passos dentro de uma época. A figura 1 apresenta a queda e os valores de learning rate com o avanço do treinamento.

2 Resultados e Discussão

Para apresentar os resultados, foram apresentadas **apenas** a comparação entre as **loss** entre treino e validação em todos os casos. As figuras estão na seção 3. Aqui, apenas descreveremos o que cada figura representa:

1. A figura 2 apresenta a comparação entre as loss de treinamento e validação quando a rede usada é uma LSTM **unidirecional sem** taxa de dropout. É interessante que a rede praticamente não aprendeu nada uma vez que a loss está bem alta nos dois casos e, além disso, é possível notar um overfitting rápido uma vez que a curva de loss cai apenas uma vez.
2. A figura 3 apresenta a comparação entre as loss de treinamento e validação quando a rede usada é uma LSTM **unidirecional** com taxa de dropout de **0.25**. É interessante que a rede praticamente não aprendeu nada uma vez que o early stopping acontece na primeira comparação entre as funções de custo de validação (época 10)
3. A figura 4 apresenta a comparação entre as loss de treinamento e validação quando a rede usada é uma LSTM **unidirecional** com taxa de dropout de **0.50**. Aqui, percebemos que a rede aprendeu bem, conseguindo ter uma boa queda de loss tanto em treinamento e validação, apresentando uma queda "suave" e precisando de 50 épocas para convergir.
4. A figura 5 apresenta a comparação entre as loss de treinamento e validação quando a rede usada é uma LSTM **bidirecional** com taxa de dropout de **0**. Aqui, percebemos que a rede aprendeu bem, chegando a uma convergência em menos de 30 épocas. Contudo, a diferença entre as taxas de validação e treinamento sugerem a presença de *sobreajuste*.
5. A figura 6 apresenta a comparação entre as loss de treinamento e validação quando a rede usada é uma LSTM **bidirecional** com taxa de dropout de **0.25**. Aqui, percebemos que a rede aprendeu bem, chegando a uma convergência em menos de 30 épocas, e a diferença entre a loss de treino e validação é um pouco menor, indicando uma certa regularização no processo de treinamento.
6. A figura 7 apresenta a comparação entre as loss de treinamento e validação quando a rede usada é uma LSTM **bidirecional** com taxa de dropout de **0.50**. Aqui, percebemos que a configuração é a melhor possível uma vez que a rede converge rápido e as curvas da função de *loss* são praticamente a mesma, indicando que essa taxa de *dropout* conseguiu lidar bem com o problema de overfitting.

Os resultados para cada um dos seis experimentos estão apresentados abaixo

Experimento	Acurácia no Teste
unidirecional com 0 de dropout	0.4344
unidirecional com 0.25 de dropout	0.3623
unidirecional com 0.5 de dropout	0.5926
bidirecional com 0 de dropout	0.5987
bidirecional com 0.25 de dropout	0.6016
bidirecional com 0.5 de dropout	0.5996

Tabela 1: Métricas de Teste para cada Experimento

É nítida a vantagem das redes bidirecionais em relação à unidirecional. Apesar do melhor resultado, em termos de acurácia, ter sido melhor no caso de bidirecional com 0.25 de dropout, ao comparar as curvas de função de custo das imagens 6 e 7 e levando em conta a diferença absolutas entre as acurácias nesse dois casos, podemos dizer que a rede bidirecional com 0.5 de dropout *generalizou* melhor

- [Cho+15] François Chollet et al. *Keras*. <https://keras.io>. 2015.
- [Har+17] Nathan Hartmann et al. *Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks*. 2017. arXiv: [1708.06025](#) [cs.CL].
- [KB17] Diederik P. Kingma e Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: [1412.6980](#) [cs.LG].

3 Lista de Figuras

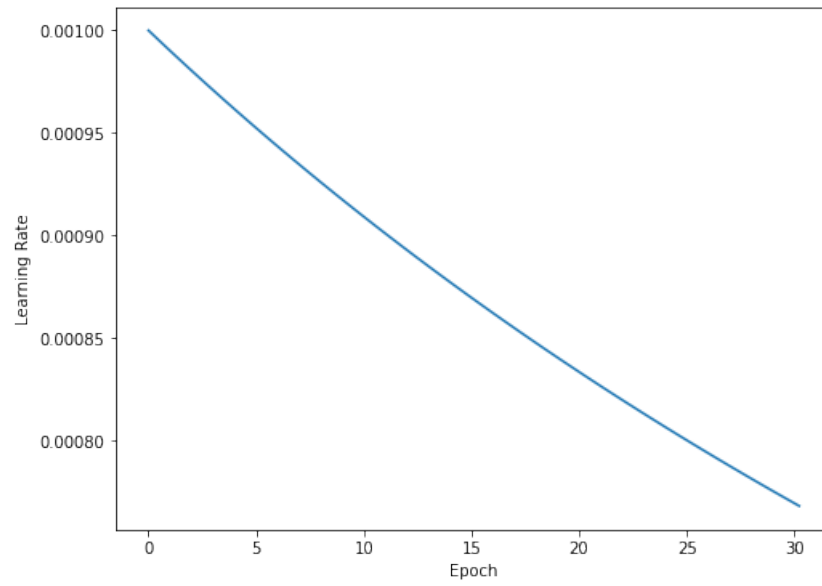


Figura 1: Taxa de queda do *Learning Rate*

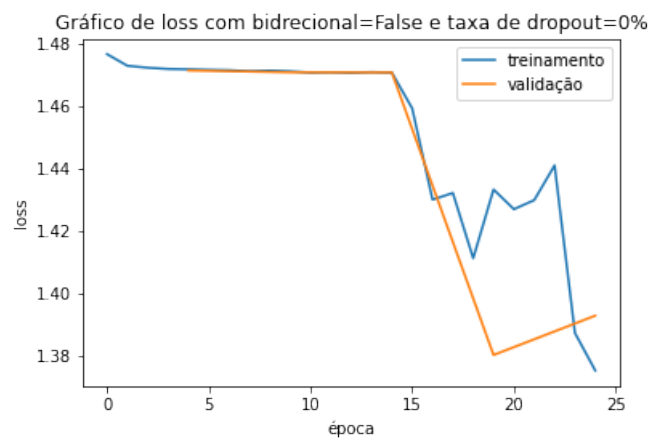


Figura 2: Curva de loss com uma LSTM unidirecional e sem dropout

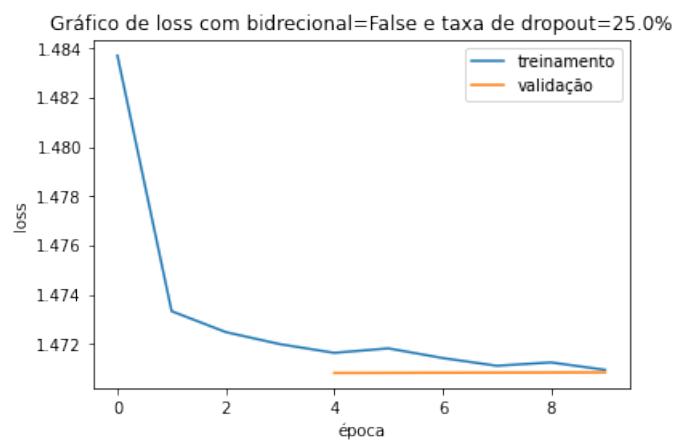


Figura 3: Curva de loss com uma LSTM unidirecional e 25% de taxa de dropout

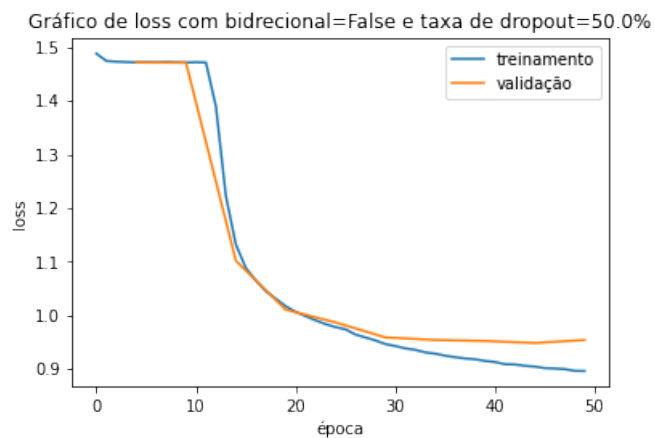


Figura 4: Curva de loss com uma LSTM unidirecional e 50% de taxa de dropout

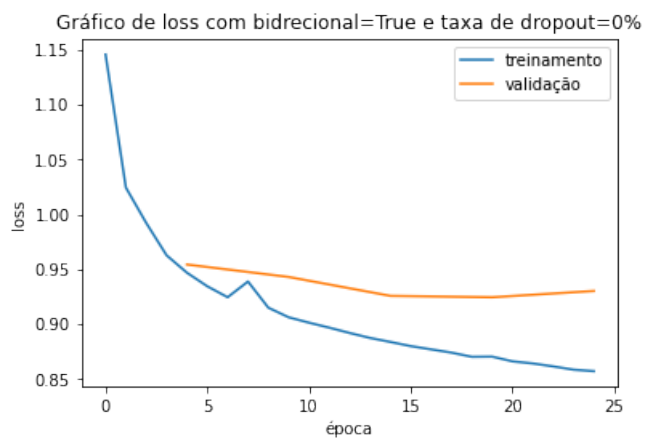


Figura 5: Curva de loss com uma LSTM bidirecional e 0% de taxa de dropout

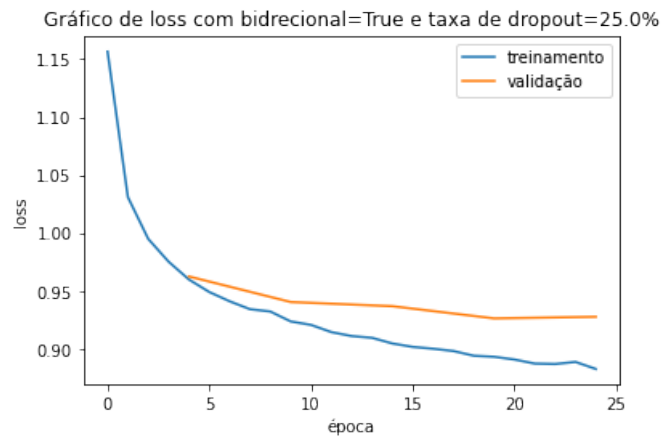


Figura 6: Curva de loss com uma LSTM bidrecional e 25% de taxa de dropout

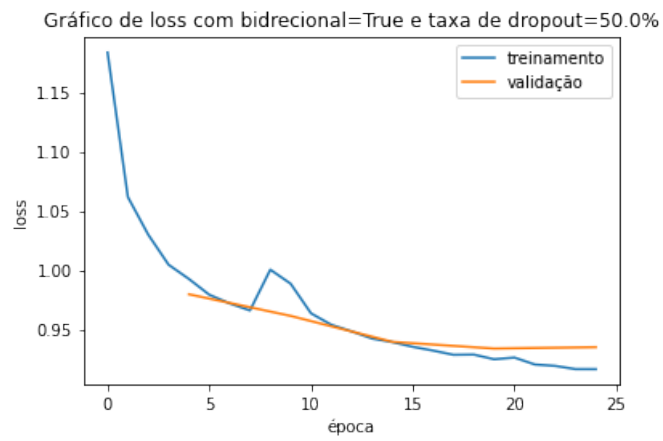


Figura 7: Curva de loss com uma LSTM bidrecional e 50% de taxa de dropout