Attention is All You Need

Bill Watson

S&P Global

November 22, 2019

Recap: Encoder-Decoder Models

How can we improve this model?

What are Context Vectors?

$$c_i = \sum_j \alpha_{ij} \cdot h_j$$

(Simple) Attention Mechanisms

Mean Attention

Location Based: Laplace

$$\alpha_{ij} = f\left(j \mid i, b\right) = \frac{1}{2b} \exp\left(-\frac{|j - i|}{b}\right)$$

- Weighted by location in sequence
  - Center a Laplace at $\mu = i$, scale b
  - Weight is the probability of position j relative to i
- Penalizes elements away from the diagonal
- Heavier tail than a Gaussian

Location Based: Gaussian

$$\alpha_{ij} = f\left(j \mid i, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{()j - i)^2}{2\sigma^2}\right)$$

- Weighted by location in sequence
  - Center a Gaussian at $\mu = i$, scale σ
  - Weight is the probability of position j relative to i
- Smoother distribution closer to the diagonal
- Smaller tail than Laplace, so outliers are penalized more

Visualizing Location-Based Distributions

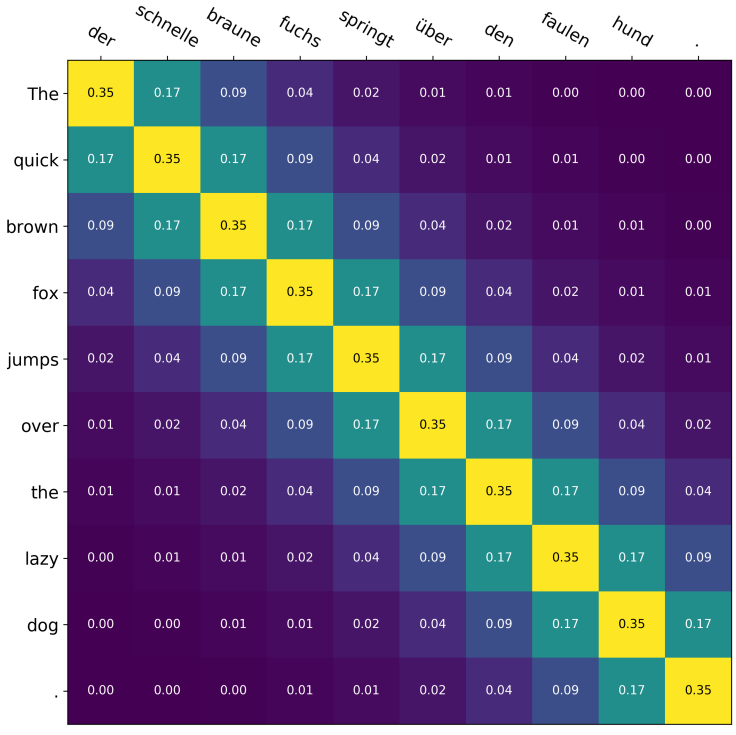

Figure: Laplace

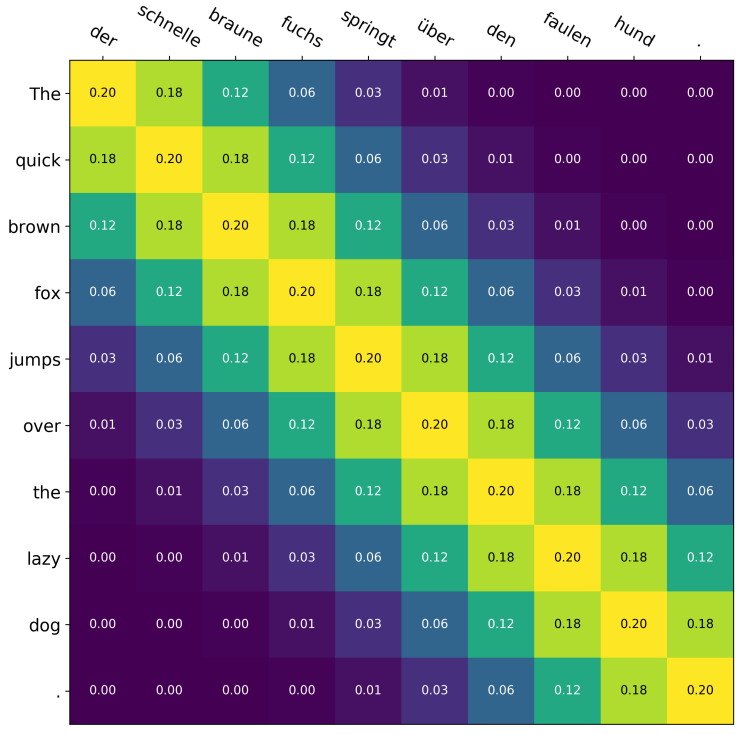

Figure: Gaussian

(Parameter-based) Attention Mechanisms

Bahdanau Attention

Luong Attention

(Advanced) Attention Mechanisms

Fine-Grained Attention

Multi-Headed Attention

Self Attention

Practical Considerations

Masking

Which one to use?

Mixing and Matching

Tools, References, and Further Reading

Refrences & Further Reading

- Machine Learning: A Probabilistic Perspective by Kevin Murphy