CS 4824/ECE 4424: Attention

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
Attention

- **Key idea**: highlight important parts of the inputs

- Mechanism for alignment in machine translation, image captioning, etc.

- Attention in machine translation: align each output word with relevant input words by computing a softmax of the inputs
Attention

- Attention in Computer Vision
  - 2014: Attention used to highlight important parts of an image that contribute to a desired output

- Attention in NLP
  - 2015: machine translation
  - 2017: Language modeling with Transformer networks
Sequence Modeling

- **Challenges with RNNs**
  - Long range dependencies
  - Gradient vanishing (and explosion)
  - Large # of training steps
  - Recurrence prevents parallel computation

- **Transformer Networks**
  - Facilitate long range dependencies
  - No gradient vanishing (and explosion)
  - Fewer training steps
  - No recurrence that facilitate parallel computation

vs
Attention Mechanism

- Mimics the retrieval of a **value** $v_i$ for a **query** $q$ based on a **key** $k_i$ in database

- Retrieval:

$$attention(q, k, v) = \sum_i similarity(q, k_i) \times v_i$$