Attention in Deep Learning

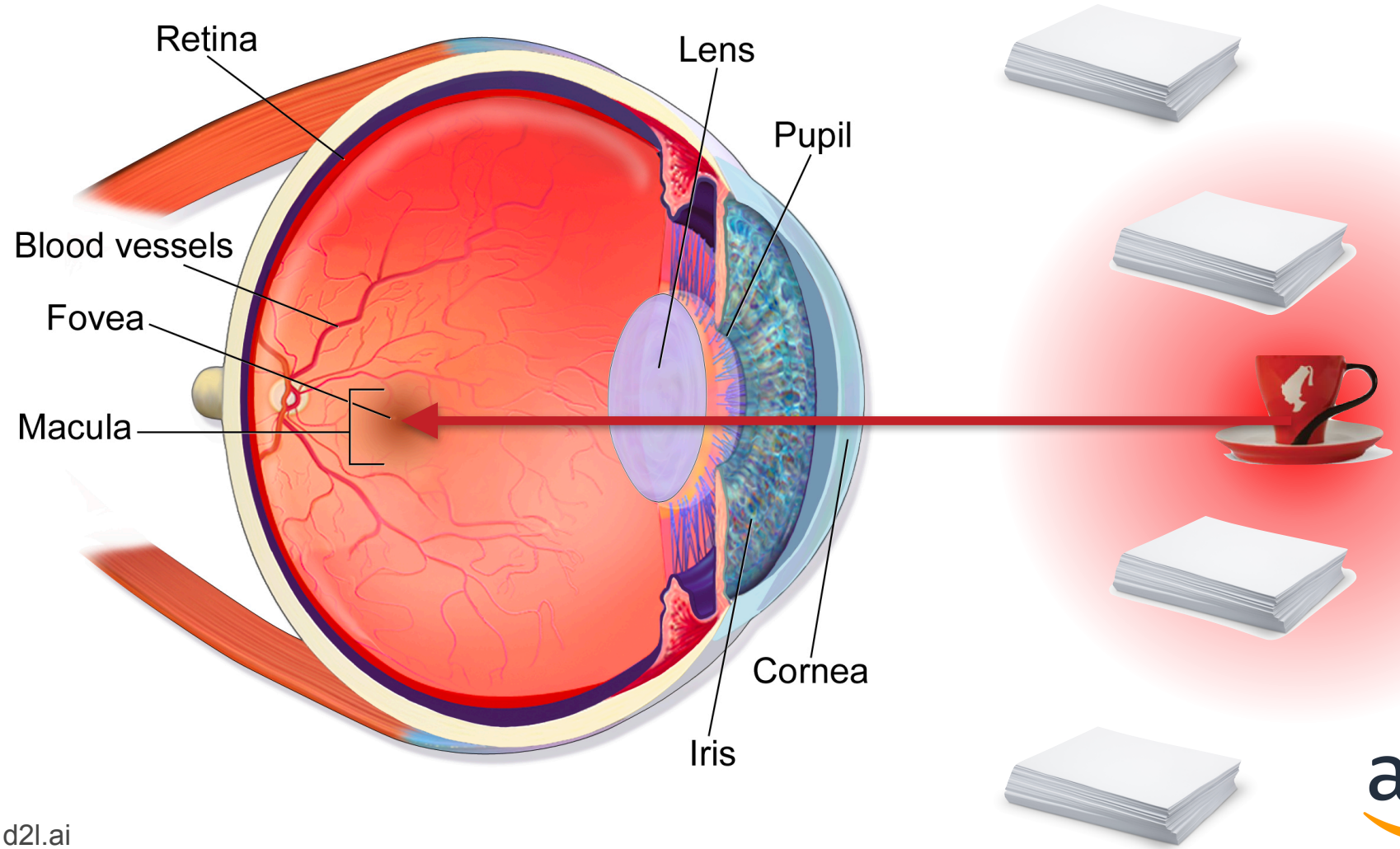
7. Attention Mechanism and Transformer

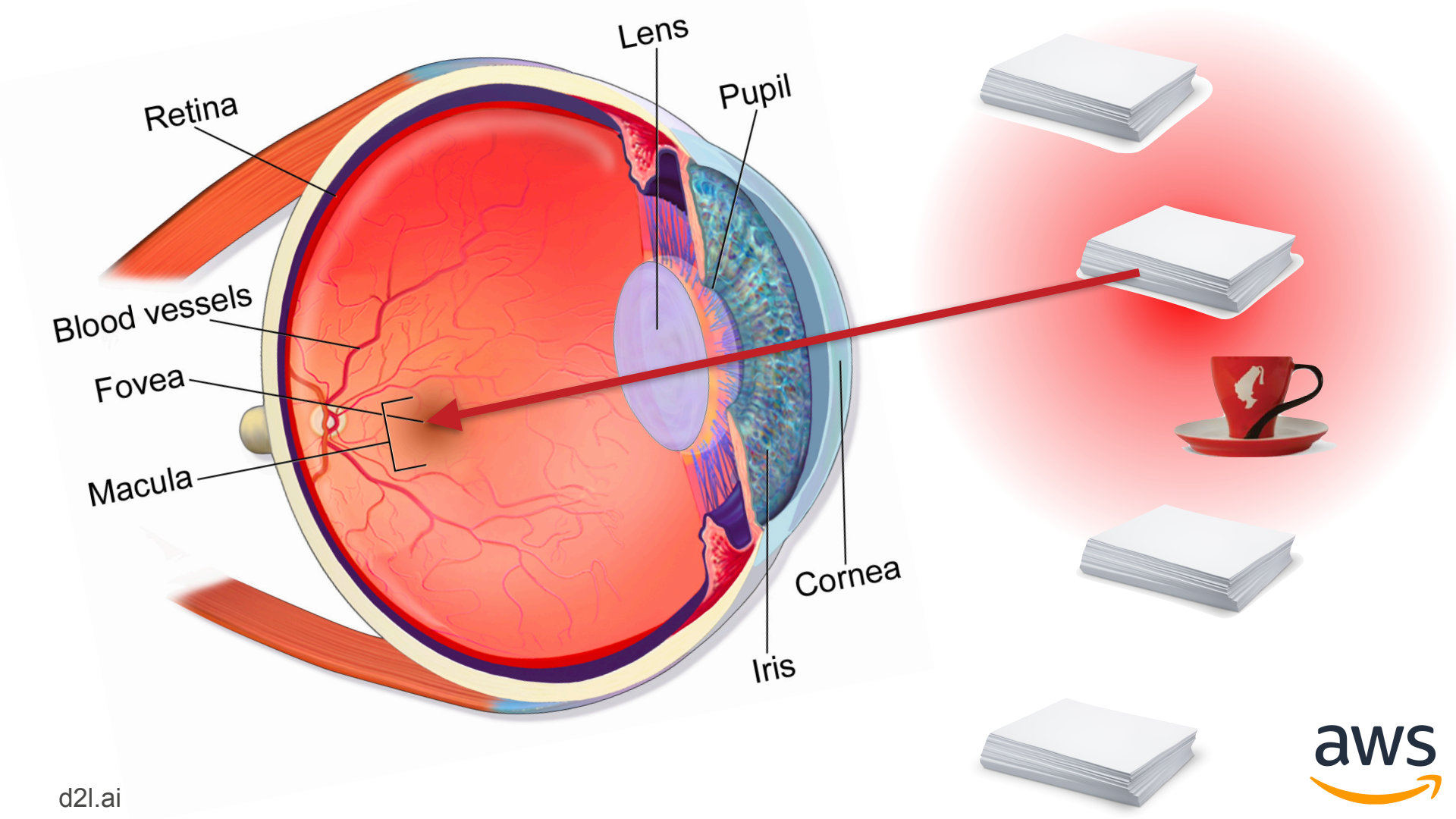
Haibin Lin and Leonard Lausen

gluon-nlp.mxnet.io

aws

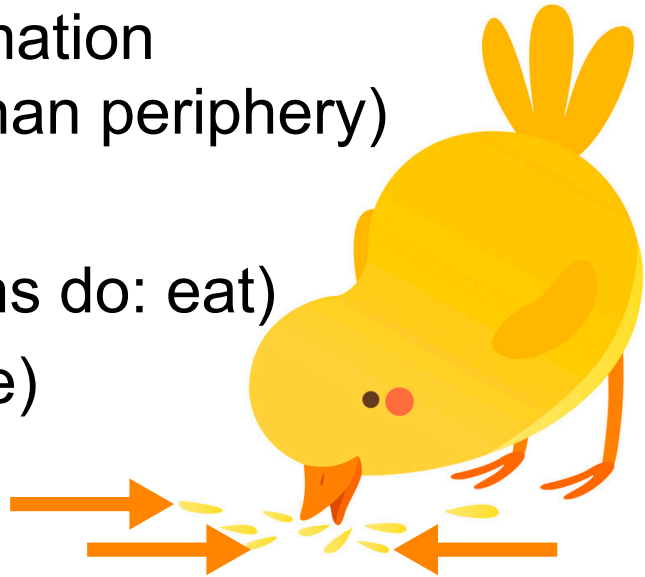
8:30-9:00	Continental Breakfast
9:00-9:45	Introduction and Setup
9:45-10:30	Neural Networks 101
10:30-10:45	Break
10:45-11:15	Machine Learning Basics
11:15-11:45	Context-free Representations for Language
11:45-12:15	Convolutional Neural Networks
12:15-13:15	Lunch Break
13:15-14:00	Recurrent Neural Networks
14:00-14:45	Attention Mechanism and Transformer
14:45-15:00	Coffee Break
15:00-16:15	Contextual Representations for Language
16:15-17:00	Language Generation


Retina
Lens
Pupil
Blood vessels
Fovea
Macula
Cornea
Iris


Lens
Retina
Pupil
Blood vessels
Fovea
Macula
Cornea
Iris

Attention in Animals

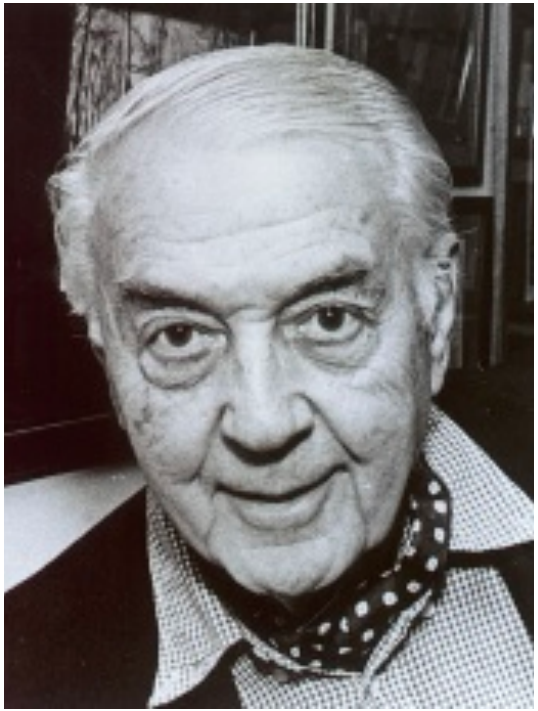
- **Resource saving**
  - Only need **sensors** where relevant bits are (e.g. fovea vs. peripheral vision)
  - Only **compute** relevant bits of information (e.g. fovea has many more ‘pixels’ than periphery)
- **Variable state manipulation**
  - Manipulate environment (for all grains do: eat)
  - Learn modular subroutines (not state)
- **In machine learning - nonparametric**



Outline

1. **Watson Nadaraya Estimator**
2. **Pooling**
   - Single objects - Pooling to attention pooling
   - Hierarchical structures - Hierarchical attention networks
3. **Iterative Pooling and Generation**
   Neural machine translation
4. **Multiple Attention Heads**
   - Transformers / BERT
   - Lightweight, structured, sparse
5. **Resources**

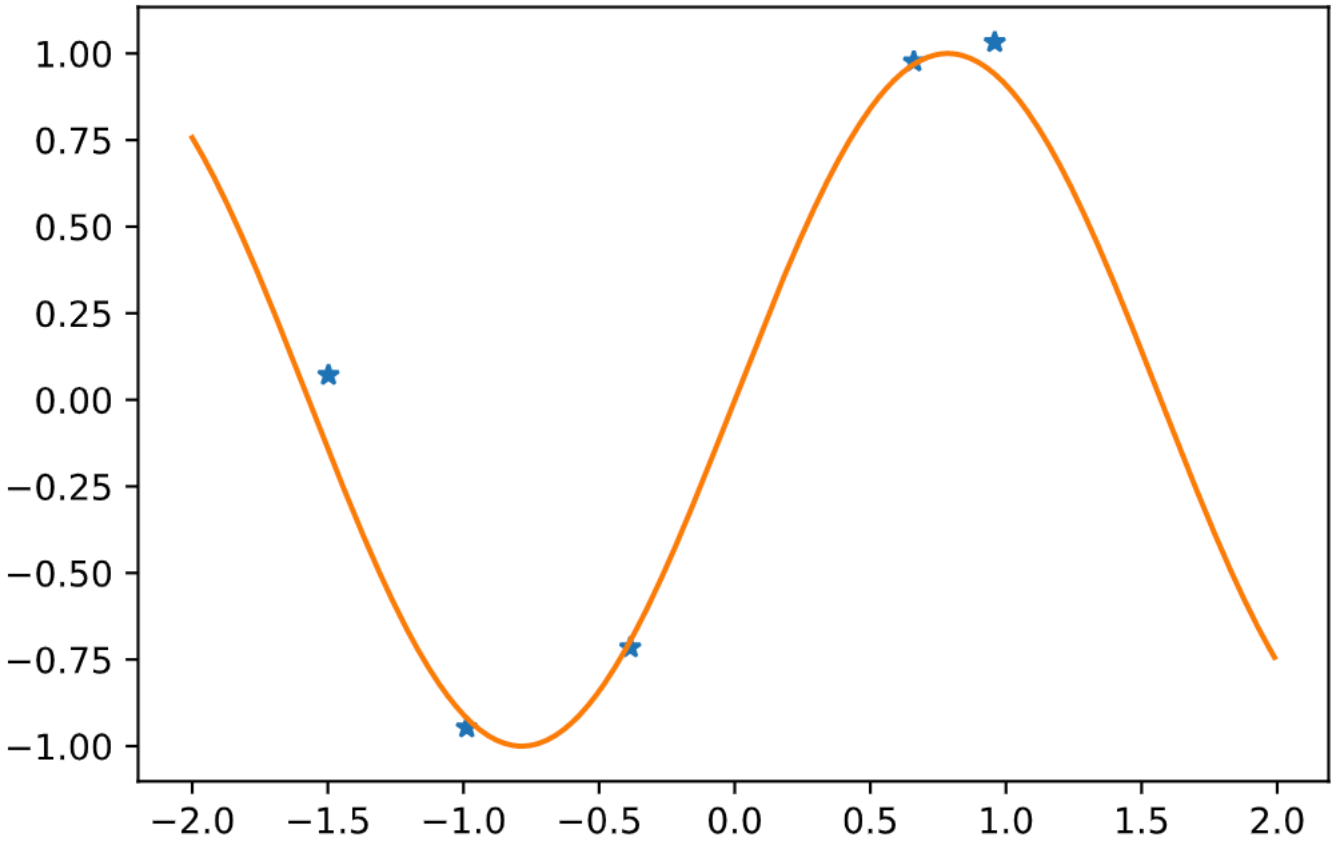
1. Watson Nadaraya Estimator ‘64



Geoffrey Watson

Elizbar Nadaraya

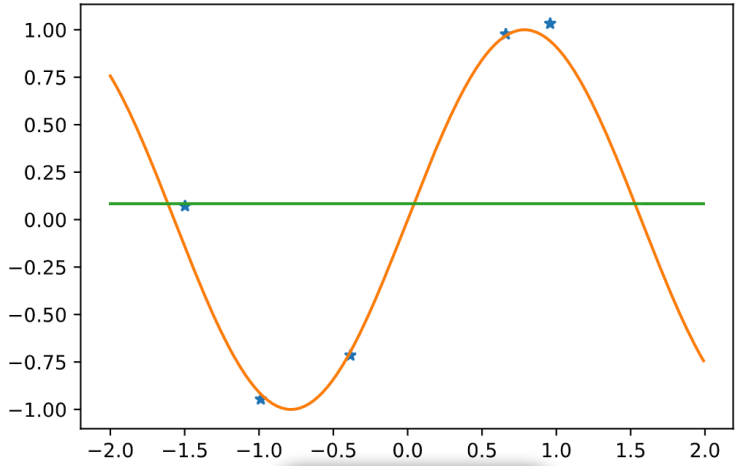
Regression Problem



Solving the regression problem

- Data $\{x_1, \ldots x_m\}$ and labels $\{y_1, \ldots y_m\}$
- Estimate label y at new location x
- **The world's dumbest estimator**
  Average over all labels

$$y = \frac{1}{m} \sum_{i=1}^{m} y_i$$



- **Better idea (Watson, Nadaraya, 1964)**
  Weigh the labels according to location

$$y = \sum_{i=1}^{m} \alpha(x, x_i) y_i$$

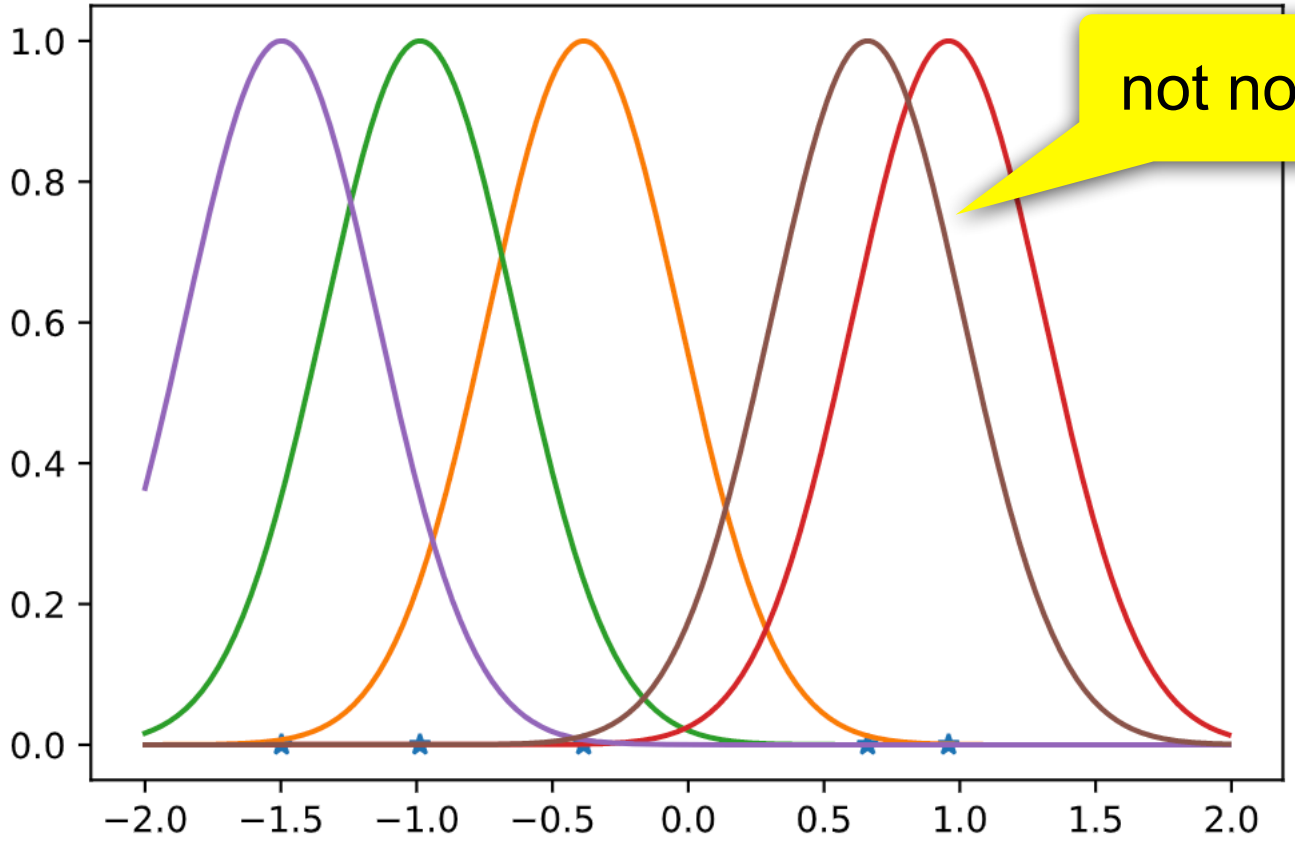
key

$\alpha(x, x_i) y_i$

query

value

Weighing the locations (e.g. with Gaussians)

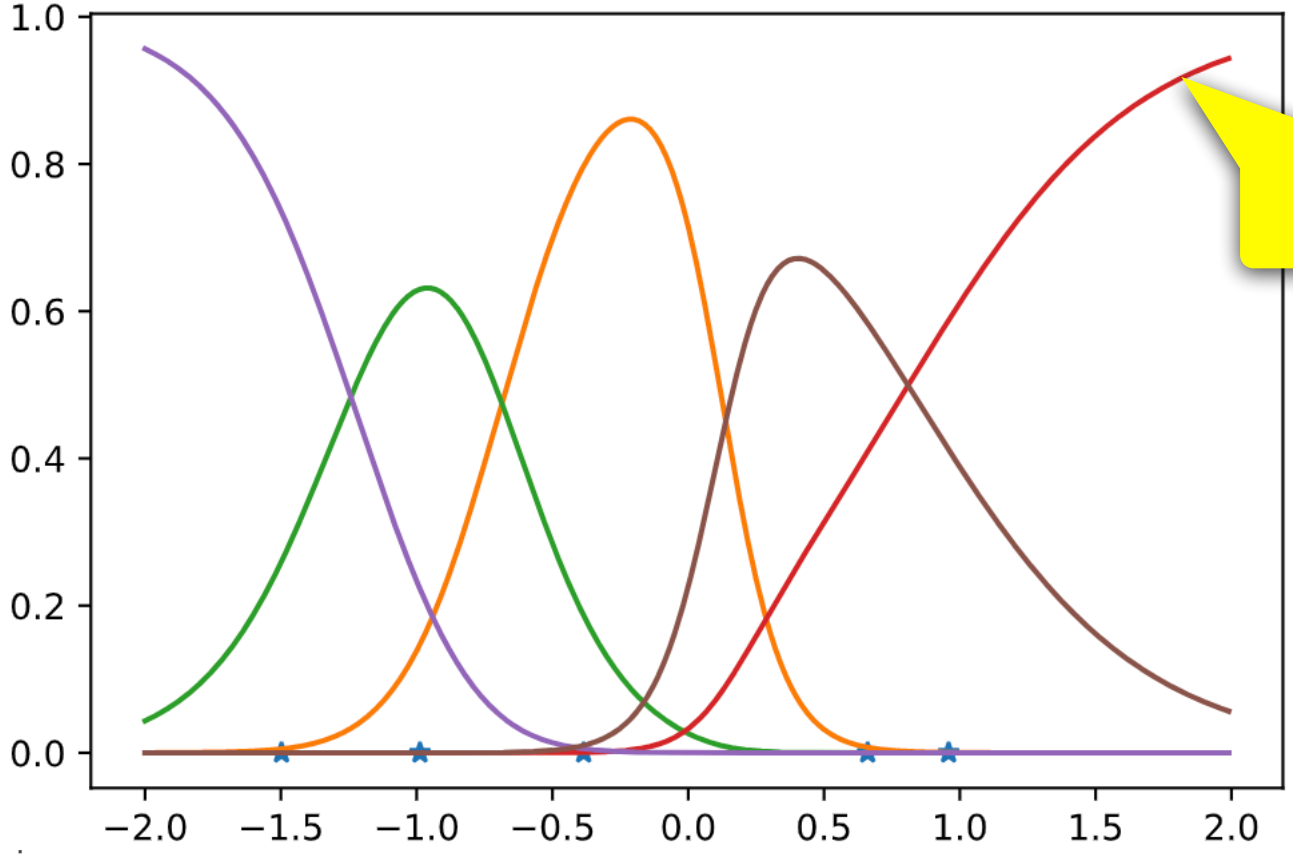


not normalized

$$\alpha(x, x_i) \propto k(x_i, x)$$

aws

Weighing the locations (e.g. with Gaussians)

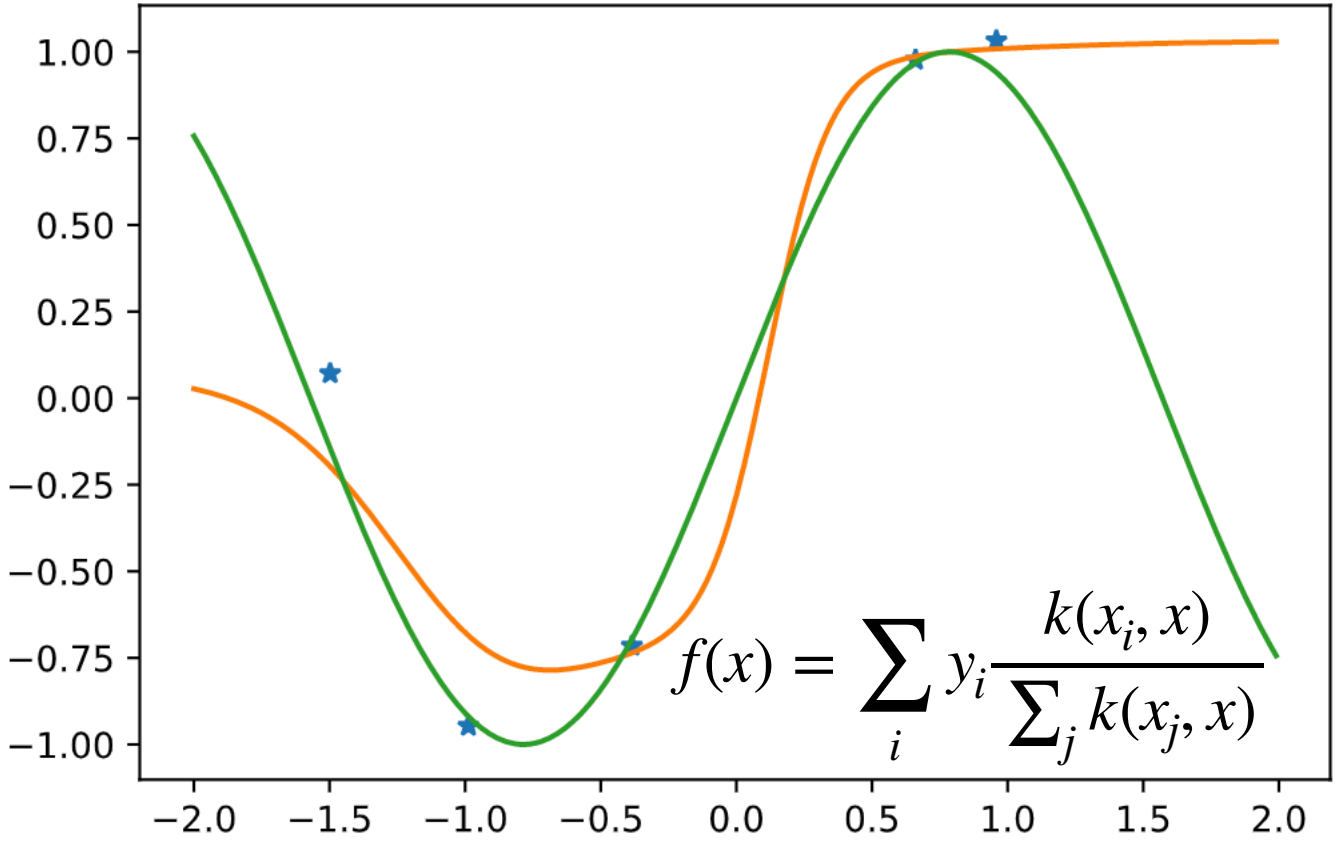


normalized

$$\alpha_i(x) = \frac{k(x_i, x)}{\sum_j k(x_j, x)}$$

aws

Weighted regression estimate



Why bother with a 55 year old algorithm?

- **Consistency**
  Given enough data this algorithm converges to the optimal solution (can your deep net do this?)
- **Simplicity**
  No free parameters - information is in the data not weights (or very few if we try to learn the weighting function)


2. Pooling

Deep Sets (Zaheer et al. 2017)

- Deep (Networks on) Sets $X = \{x_1, \ldots x_n\}$
  - Need permutation invariance for elements in set (e.g. LSTM doesn't work to ingest elements)
  - Theorem - all functions are of the form*

$$f(X) = \rho\left(\sum_{x \in X} \phi(x)\right)$$

*or some combination thereof

- Applications - point clouds, set extension, red shift for galaxies, text retrieval, tagging, etc.

Deep Sets (Zaheer et al. 2017)

Outliers in sets - learn function $f(X)$ on set such that

$$f(\{x\} \cup X) \geq f\left(\{x'\} \cup X\right) + \Delta\left(x, x'\right)$$

Deep Sets with Attention aka Multi-Instance Learning (Ilse, Tomczak, Welling, '18)

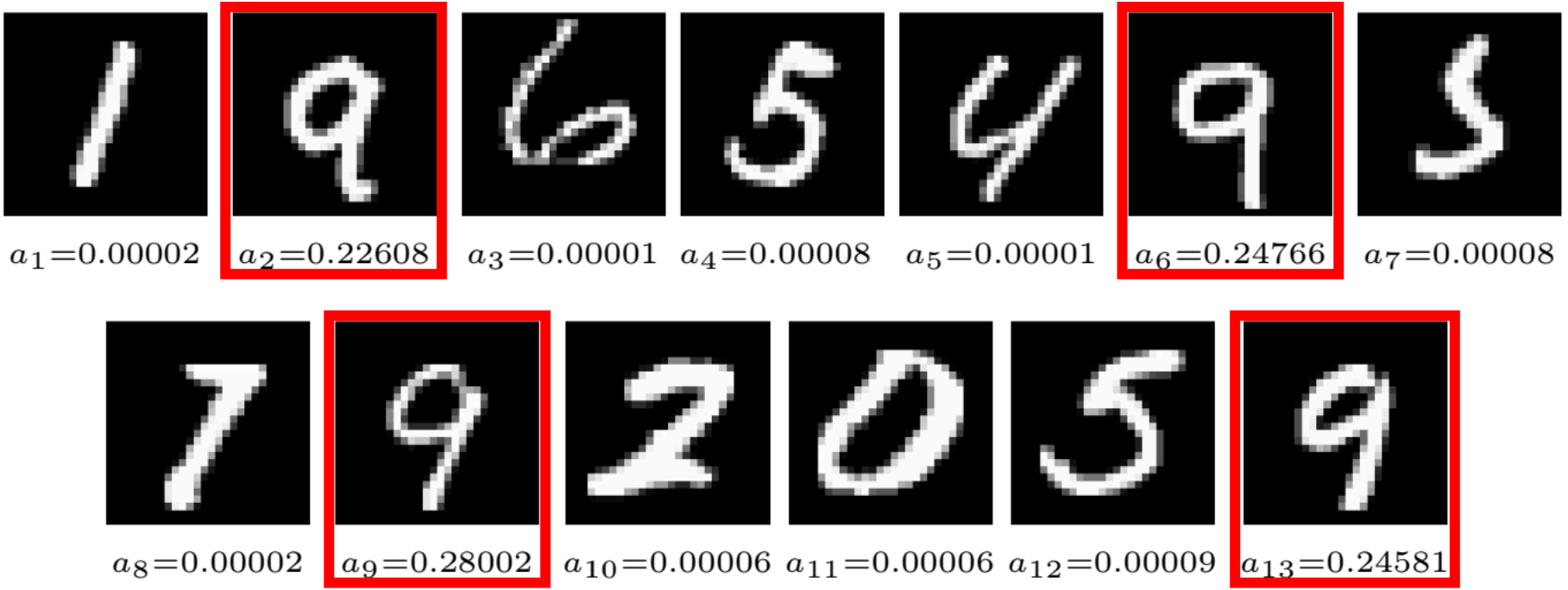
- Multiple Instance Problem
  Set contains one (or more) elements with desirable property (drug discovery, keychain). Identify those sets.
- Deep Sets have trouble focusing, hence weigh it

$$f(X) = \rho\left(\sum_{x \in X} \phi(x)\right) \longrightarrow f(X) = \rho\left(\sum_{x \in X} \alpha(w, x)\phi(x)\right)$$

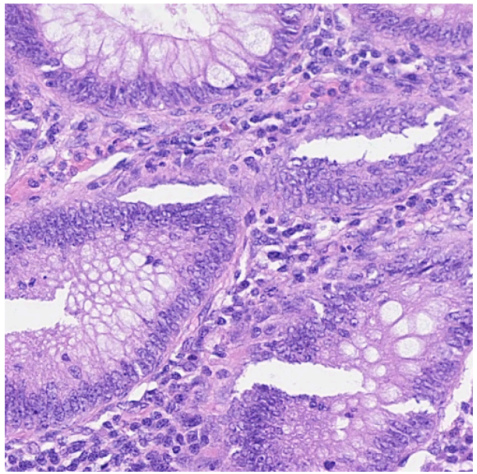
- Attention function e.g. $\alpha(w, x) \propto \exp\left(w^\top \tanh Vx\right)$

Deep Sets with Attention aka Multi-Instance Learning (Ilse, Tomczak, Welling, ’18)

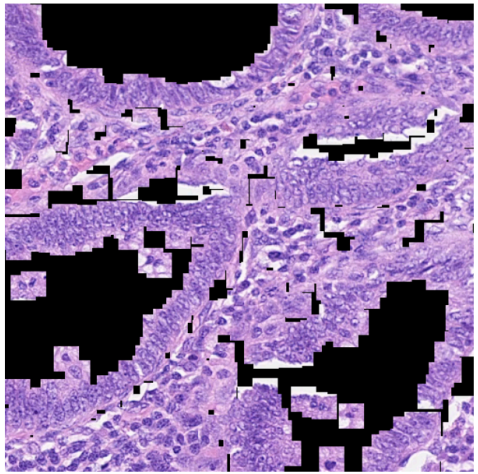
Identifying sets that contain the digit ‘9’



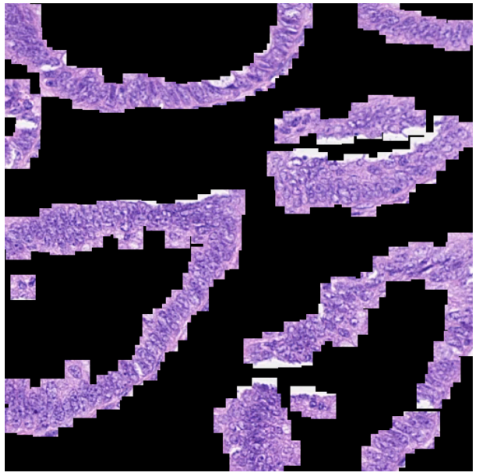
Deep Sets with Attention aka Multi-Instance Learning (Ilse, Tomczak, Welling, ’18)



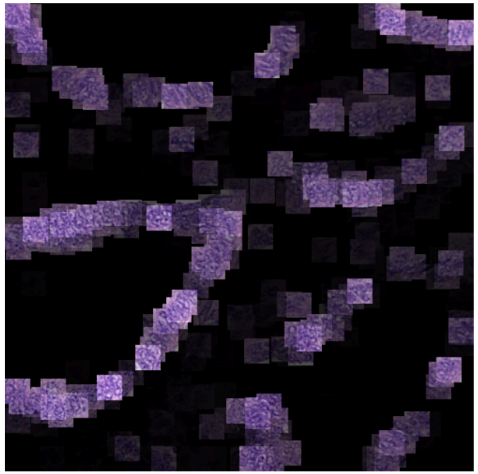
tissue sample



windowed cell nuclei



cancerous cells



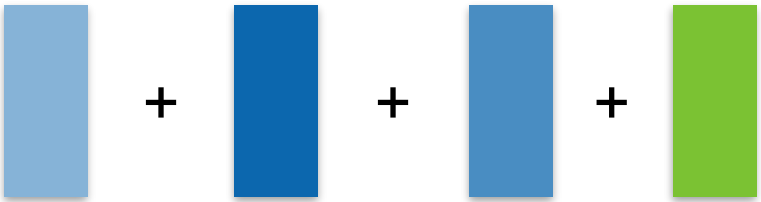
attention weights

Bag of words (Salton & McGill, 1986)
Word2Vec (Mikolov et al., 2013)

- Embed each word in sentence (word2vec, binary, GRU …)
- Add them all up
- Classify

$$f(X) = \rho\left(\sum_{i=1}^{n} \phi(x_i)\right)$$

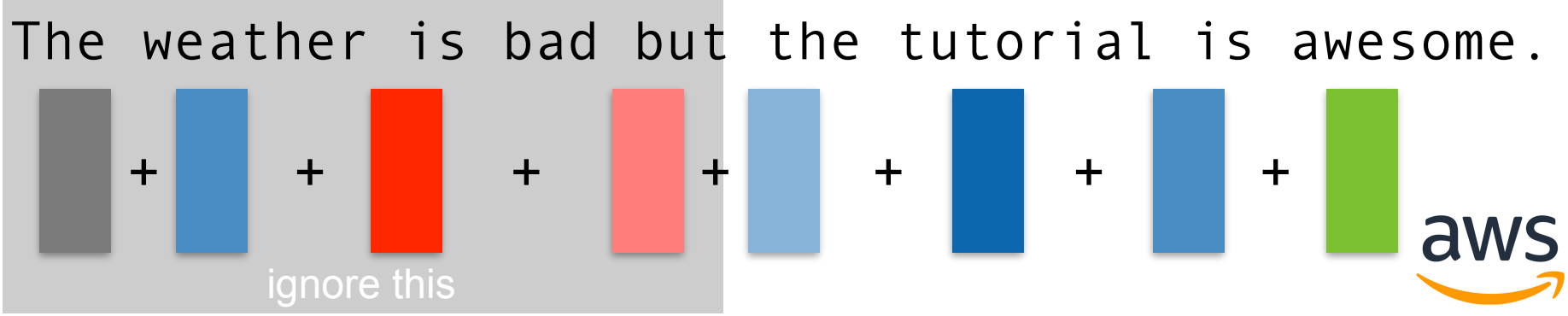
```
The tutorial is awesome.
```



Bag of words (Salton & McGill, 1986)
Word2Vec (Mikolov et al., 2013)

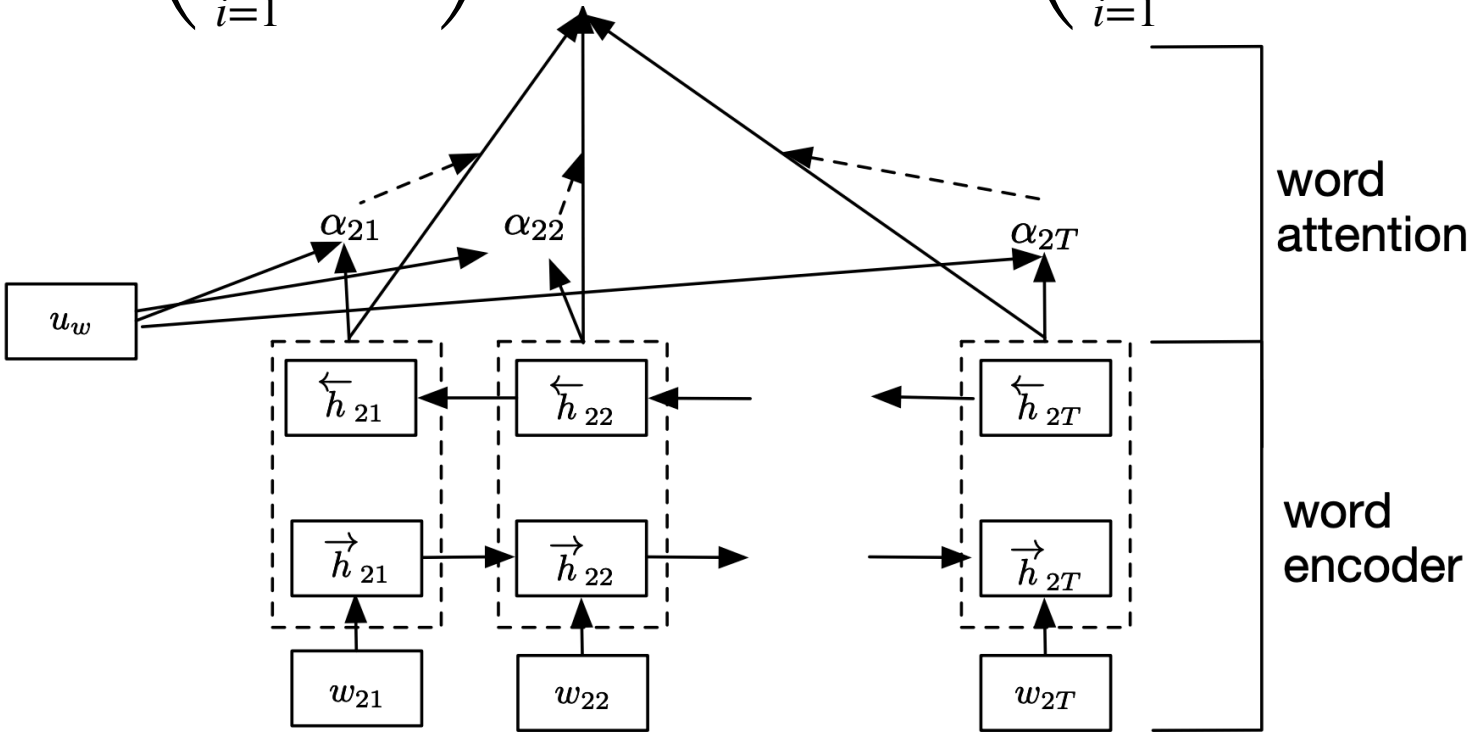
- Embed each word in sentence (word2vec, binary, GRU …)
- Add them all up
- Classify

$$f(X) = \rho\left(\sum_{i=1}^{n} \phi(w_i)\right)$$



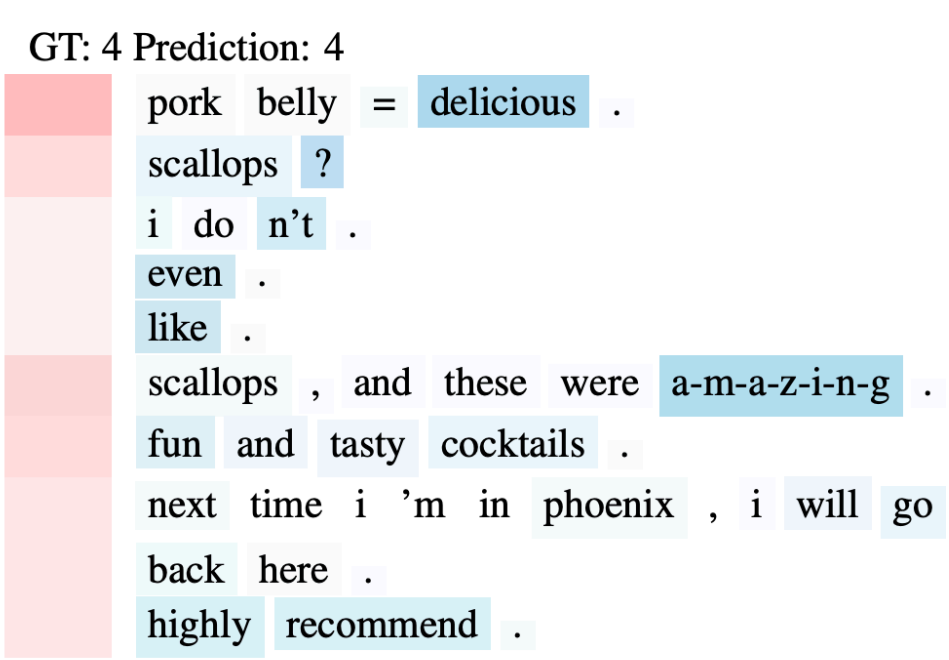
Attention weighting for documents (Wang et al, ’16)

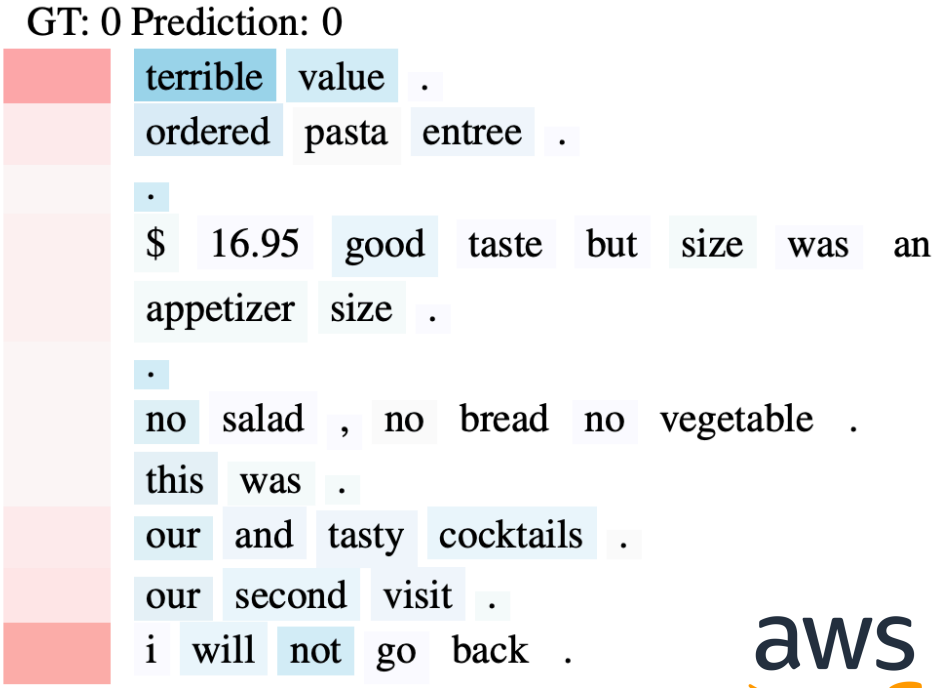
$$f(X) = \rho\left(\sum_{i=1}^{n} \phi(w_i)\right) \rightarrow f(X) = \rho\left(\sum_{i=1}^{n} \alpha(w_i, X)\phi(w_i)\right)$$



Hierarchical attention weighting (Yang et al. ’17)

Some sentences are more important than others …





Hierarchical attention

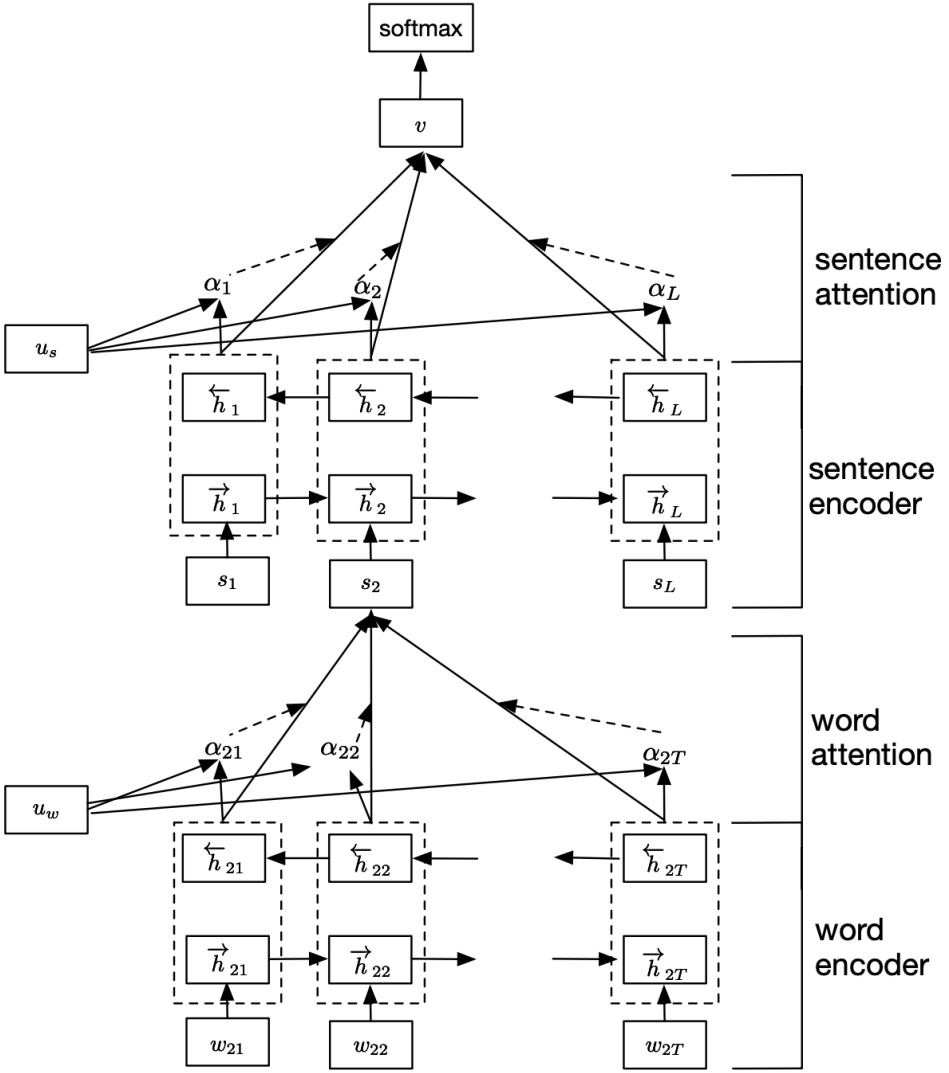
- Word level

$$f(s_i) = \rho\left(\sum_{j=1}^{n_i} \alpha(w_{ij}, s_i)\phi(w_{ij})\right)$$

- Sentence level

$$g(d) = \rho\left(\sum_{i=1}^{n} \alpha(s_i, d)\phi(f(s_i))\right)$$

- Embeddings e.g. via GRU



Attention Summary

- Pooling

$$f(X) = \rho\left(\sum_{x \in X} \phi(x)\right)$$

- Attention pooling

$$f(X) = \rho\left(\sum_{x \in X} \alpha(x, w)\phi(x)\right)$$

Query w can depend on context

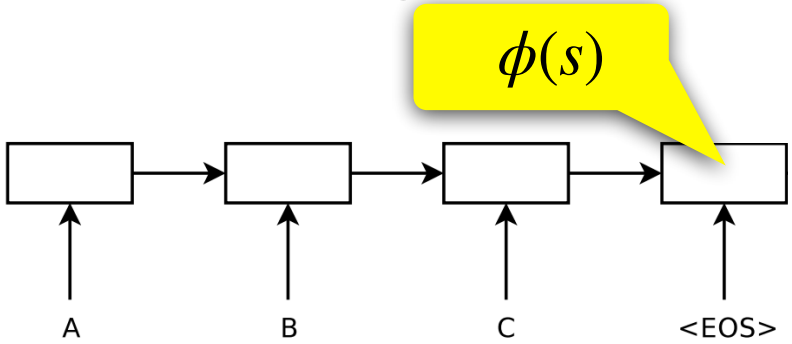
- Attention function (normalized to unit weight) such as

$$\alpha(x, X) \propto \exp\left(w^\top \tanh Ux\right)$$


3. Iterative
Output

Seq2Seq for Machine Translation, Sutskever, Vinyals, Le '14

- Encode source sequence s via LSTM to representation $\phi(s)$
- Decode to target sequence one character at a time

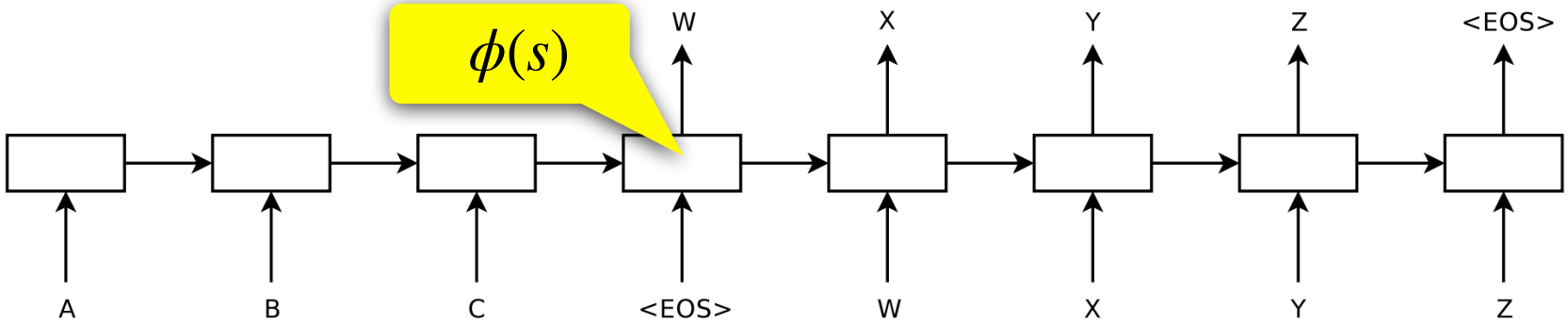


- 'The table is round.' - 'Der Tisch ist rund.'
- 'The table is very beautiful with r , blah blah blah blah' - 'Error …'

Representation not rich enough

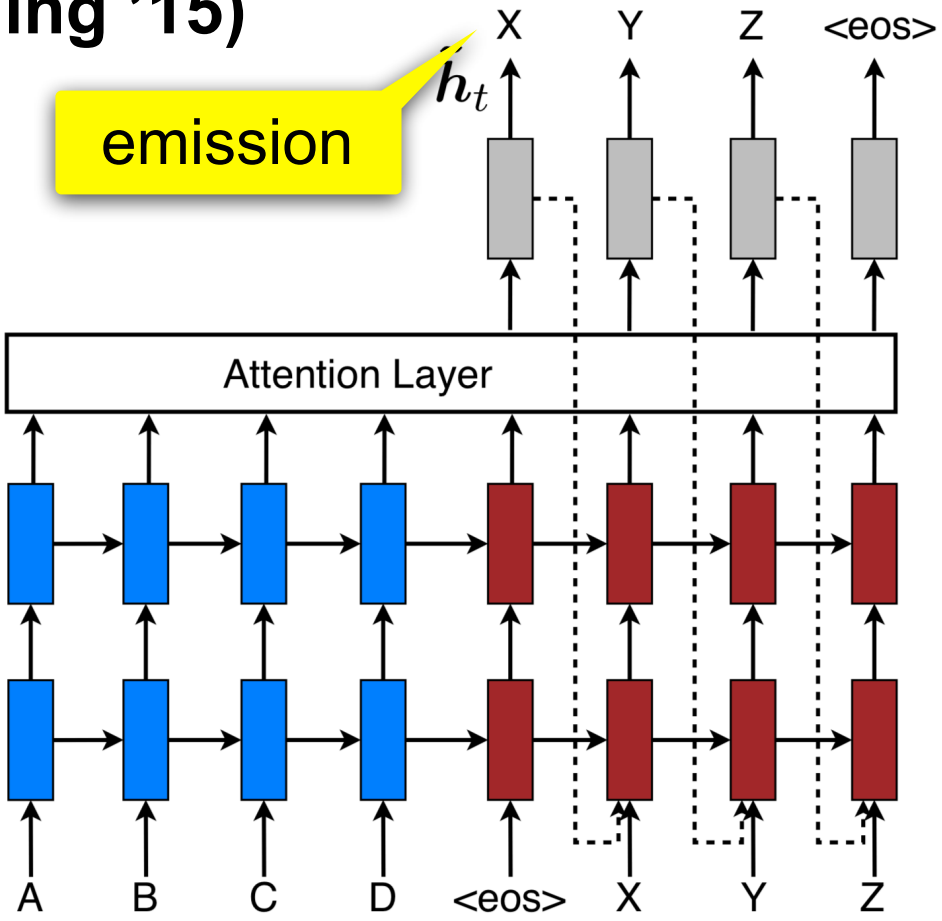
Seq2Seq for Machine Translation, Sutskever, Vinyals, Le '14

- Encode source sequence s via LSTM to latent representation $\phi(s)$
- Decode to target sequence one character at a time



- Need memory for long sequences
- Attention to iterate over source
  (we can look up source at any time after all)

Seq2Seq with attention (Bahdanau, Cho, Bengio '14) (Pham, Luong, Manning '15)



Seq2Seq with attention (Bahdanau, Cho, Bengio '14) (Pham, Luong, Manning '15)

$$\alpha_{ij} \propto \exp(a(\tilde{h}_{i-1}, h_j))$$

query

key & value

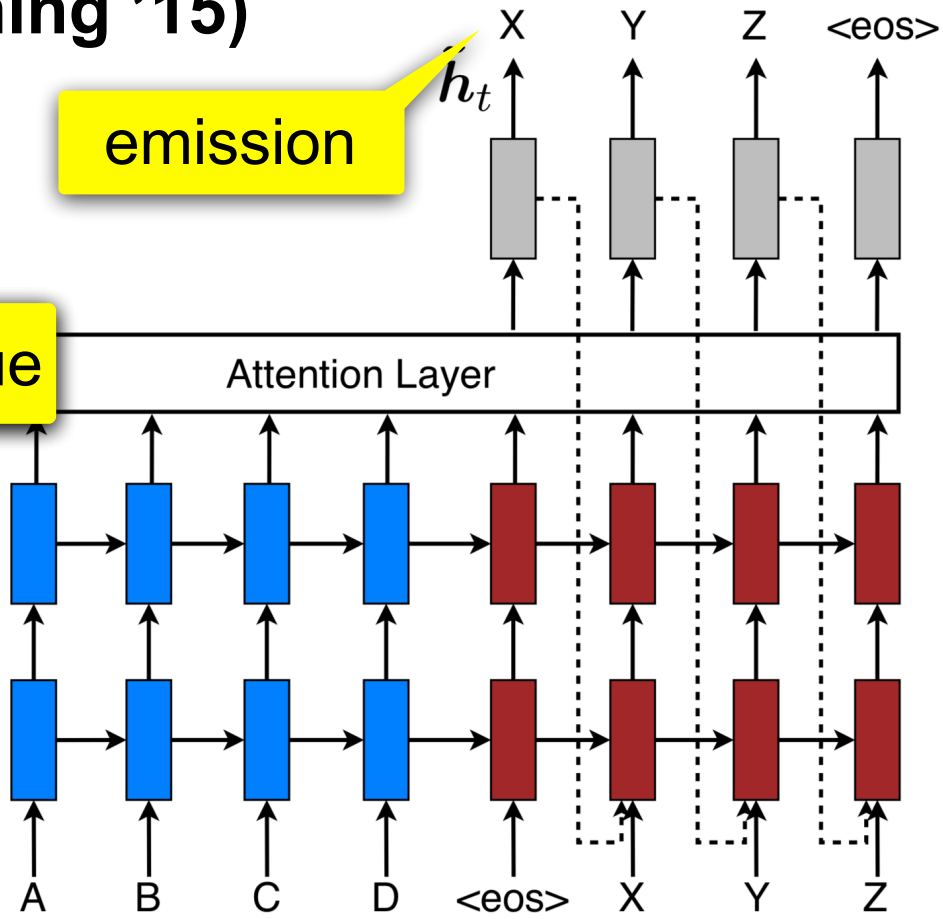
$$c_i = \sum_{j=1}^{n} \alpha_{ij} h_j$$

$$\tilde{h}_i = f(s_{i-1}, y_{i-1}, c_i)$$

y_i

emission

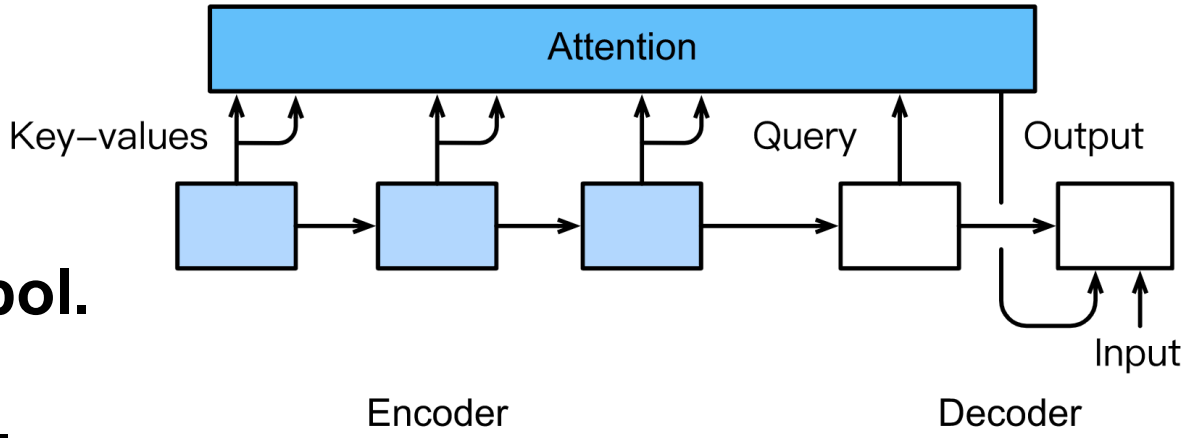
h_t



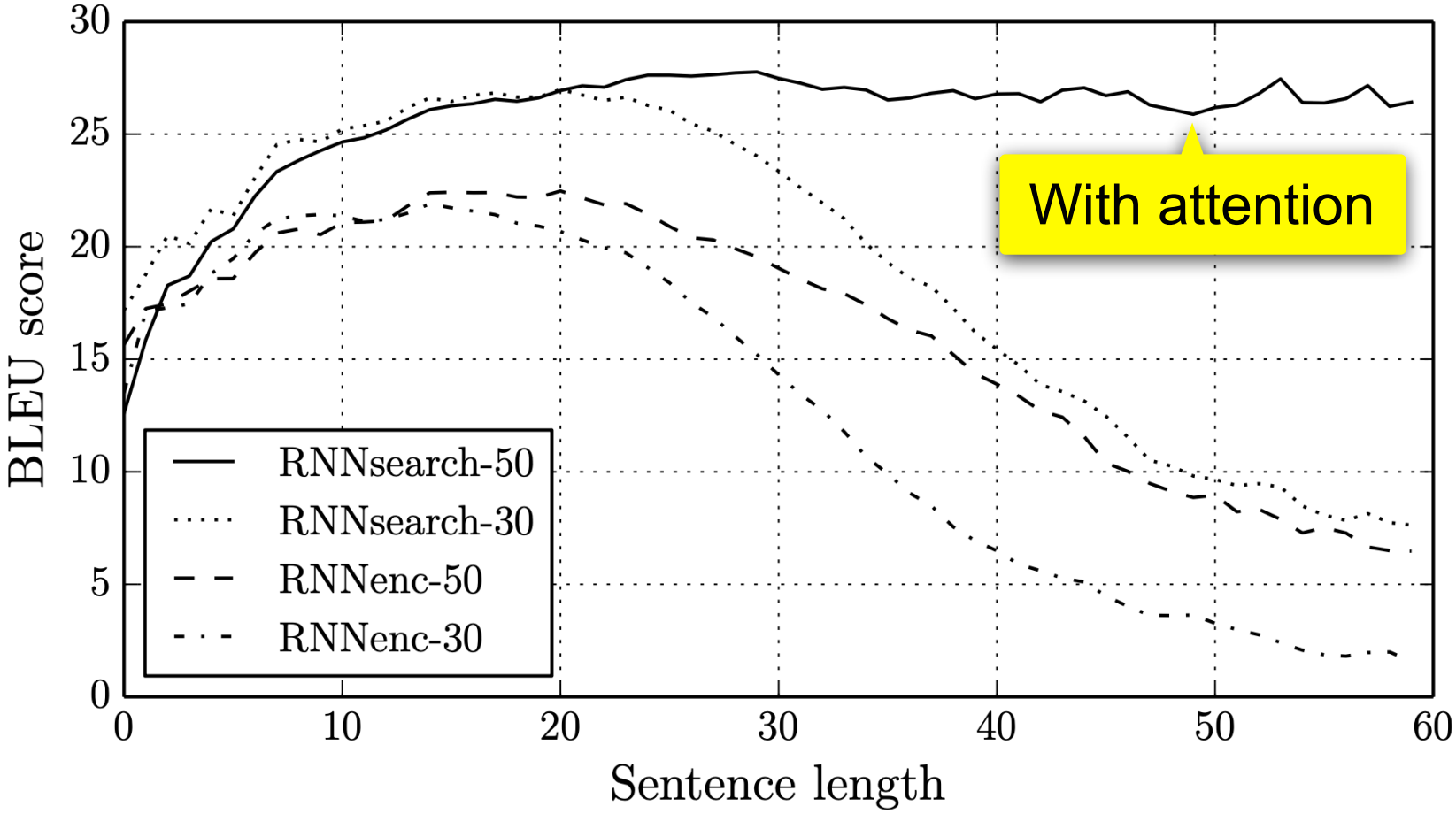
aws

Seq2Seq with attention (Bahdanau, Cho, Bengio '14) (Pham, Luong, Manning '15)

- Iterative attention model
  - Compute (next) attention weights
  - Aggregate next state
  - Emit next symbol
- Repeat
- **Memory networks emit only one symbol.**
- **NMT with attention emits many symbols.**



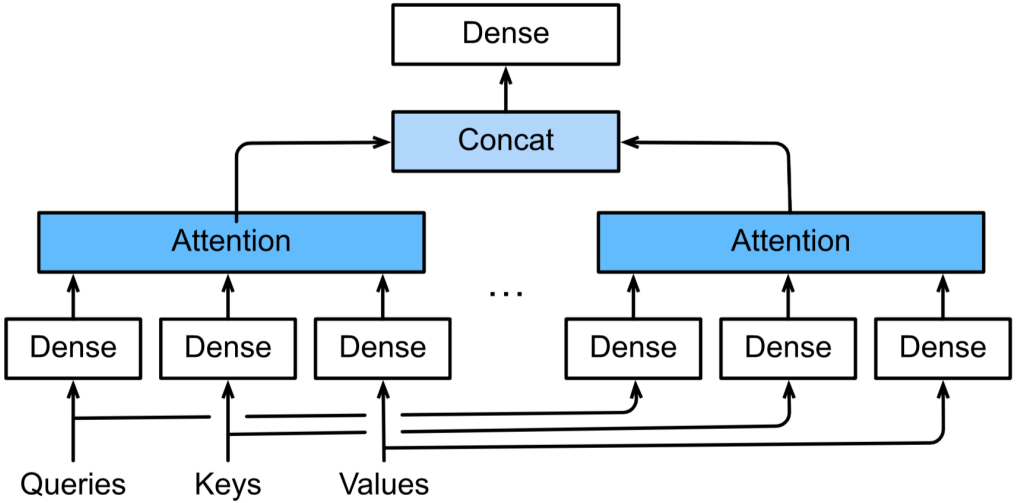
Seq2Seq with attention (Bahdanau, Cho, Bengio '14)



5. Multiple Heads

Multi-head attention (Vaswani et al., '17)

Q: query
K: key
V: value



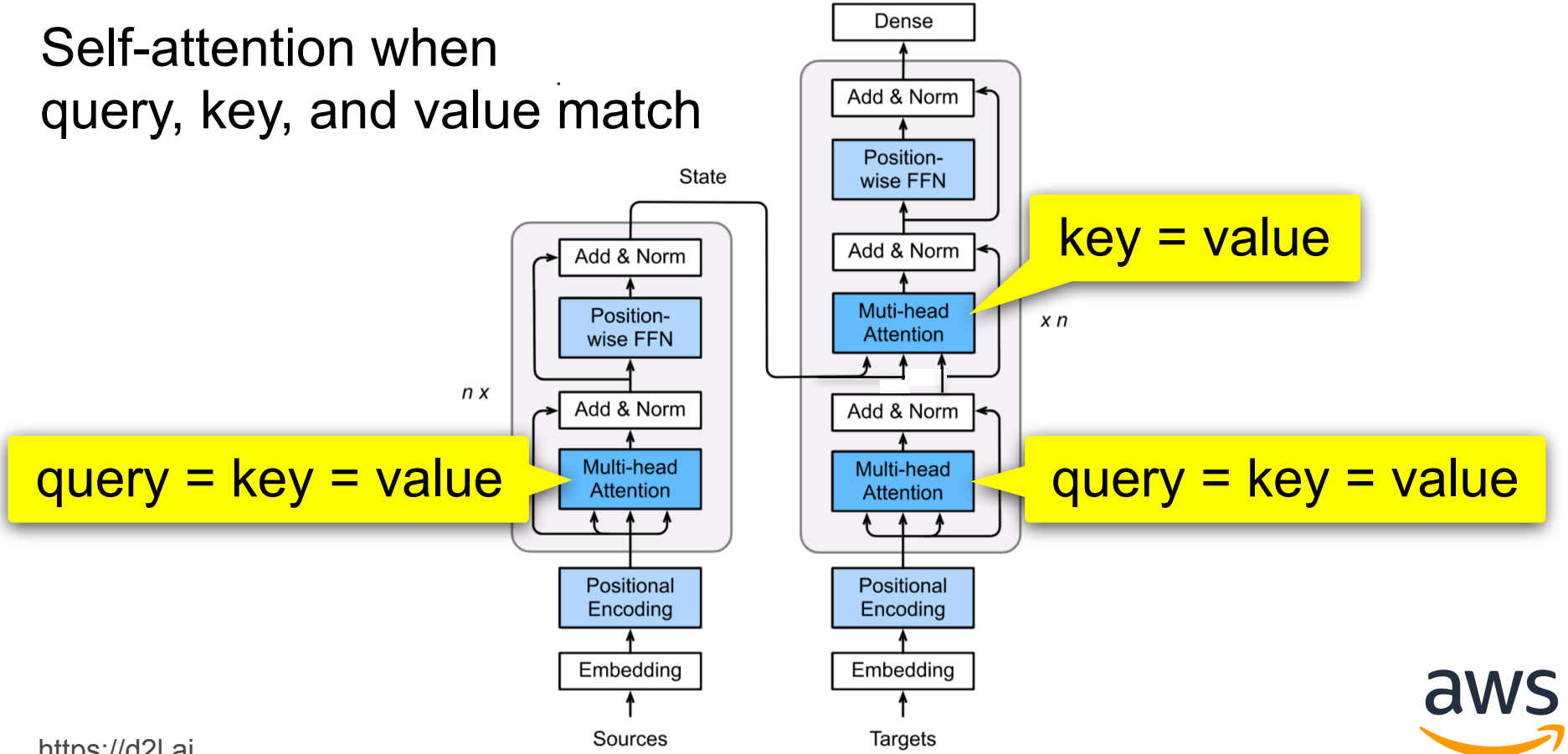
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots \text{head}_h)W^O$$

$$\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

Transformer with multi-head attention (Vaswani et al., ‘17)

Self-attention when
query, key, and value match



References

Zaheer, Manzil, et al. "Deep sets." *Advances in neural information processing systems*. 2017.

Ilse, Maximilian, Jakub M. Tomczak, and Max Welling. "Attention-based deep multiple instance learning." *arXiv preprint arXiv: 1802.04712* (2018).

Salton, Gerard, and Michael J. McGill. "Introduction to modern information retrieval." (1986).

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

Wang, Yequan, Minlie Huang, and Li Zhao. "Attention-based LSTM for aspect-level sentiment classification." *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016.

Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.

Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Veličković, Petar, et al. "Graph attention networks." *arXiv preprint arXiv:1710.10903* (2017).

Sukhbaatar, Sainbayar, Jason Weston, and Rob Fergus. "End-to-end memory networks." *Advances in neural information processing systems*. 2015.

Yang, Zichao, et al. "Stacked attention networks for image question answering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

Tay et al. "Lightweight and Efficient Neural Natural Language Processing with Quaternion Networks", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).

Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks." *Advances in Neural Information Processing Systems*. 2015.

Graves, Alex, Greg Wayne, and Ivo Danihelka. "Neural turing machines." *arXiv preprint arXiv:1410.5401* (2014).

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Zhang et al. Co-occurrent Features in Semantic Segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv: 1810.04805* (2018).

Radford, Alec, et al. "Improving language understanding by generative pre-training." *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf* (2018).

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI Blog* 1.8 (2019).

Al-Rfou, Rami, et al. "Character-level language modeling with deeper self-attention." *arXiv preprint arXiv:1808.04444* (2018).

Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." *arXiv preprint arXiv: 1901.02860* (2019).

Child, Rewon, et al. "Generating long sequences with sparse transformers." *arXiv preprint arXiv:1904.10509* (2019).