Alternative Architectures

Philipp Koehn

Alternative Architectures

- We introduced one translation model
  - attentional seq2seq model
  - core organizing feature: recurrent neural networks
- Other core neural architectures
  - convolutional neural networks
  - attention
- But first: look at various components of neural architectures

components

Components of Neural Networks

- Neural networks originally inspired by the brain
  - a neuron receives signals from other neurons
  - if sufficiently activated, it sends signals
  - feed-forward layers are roughly based on this
- Computation graph
  - any function possible, as long as it is partially differentiable
  - not limited by appeals to biological validity
- *Deep learning* maybe a better name

Feed-Forward Layer

- Classic neural network component

- Given an input vector x, matrix multiplication M with adding a bias vector b

$$Mx + b$$

- Adding a non-linear activation function

$$y = \text{activation}(Mx + b)$$

- Notation

$$y = FF_{\text{activation}}(x) = a(Mx + b)$$

Feed-Forward Layer

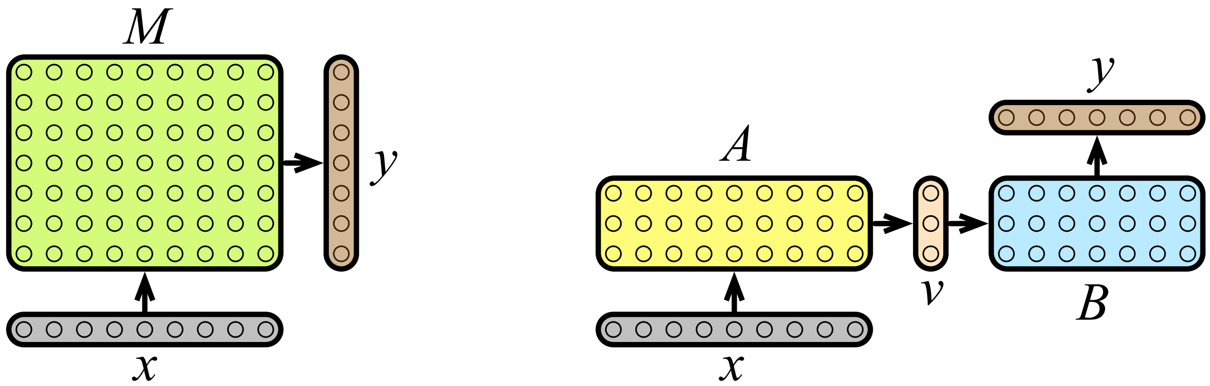
- Historic neural network designs: several feed-forward layers
  - input layer
  - hidden layers
  - output layer
- Powerful tools for a wide range of machine learning problems
- Matrix multiplication also called **affine transforms**
  - appeals to its geometrical properties
  - straight lines in input still straight lines in output

Factored Decomposition

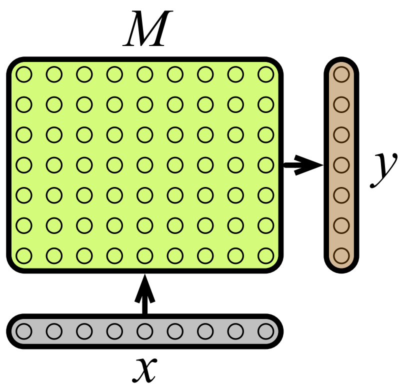
- One challenge: very large input and output vectors
- Number of parameters in matrix $M = |x| \times |y|$

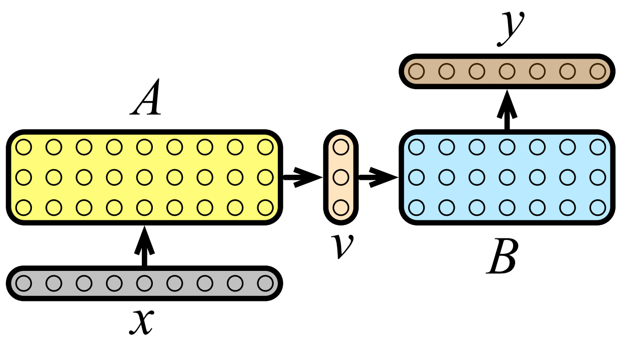
$\Rightarrow$ Need to reduce size of matrix

- Solution: first reduce to smaller representation



Factored Decomposition: Math





- Intuition
  - given highly dimension vector x
  - first map to into lower dimensional vector v (matrix A)
  - then map to output vector y (matrix B)

$$v = Ax$$
$$y = Bv = BAx$$

- Example
  - $|x|$ = 20,000, $|y|$ = 50,000 $\rightarrow$ M = 1,000,000,000
  - $|v|$ = 100 $\rightarrow$ A = 20,000 $\times$ 100 = 2,000,000, B = 100 $\times$ 50,000 = 5,000,000
  - reduction from 1,000,000,000 to 7,000,000

Factored Decomposition: Interpretation

- Vector v is a bottleneck feature
- Forced to captures salient features
- One example: word embeddings

processing sequences

Recurrent Neural Networks

- Already described recurrent neural networks at length
  - propagate state s
  - over time steps t
  - receiving an input x_t at each turn

$$s_t = f(s_{t-1}, x_t)$$

(state may computed may as a feed-forward layer)

- More successful
  - gated recurrent units (GRU)
  - long short-term memory cells (LSTM)

- Good fit for sequences, like words in a sentence
  - humans also receive word by word
  - most recent words most relevant
  - $\rightarrow$ closer to current state

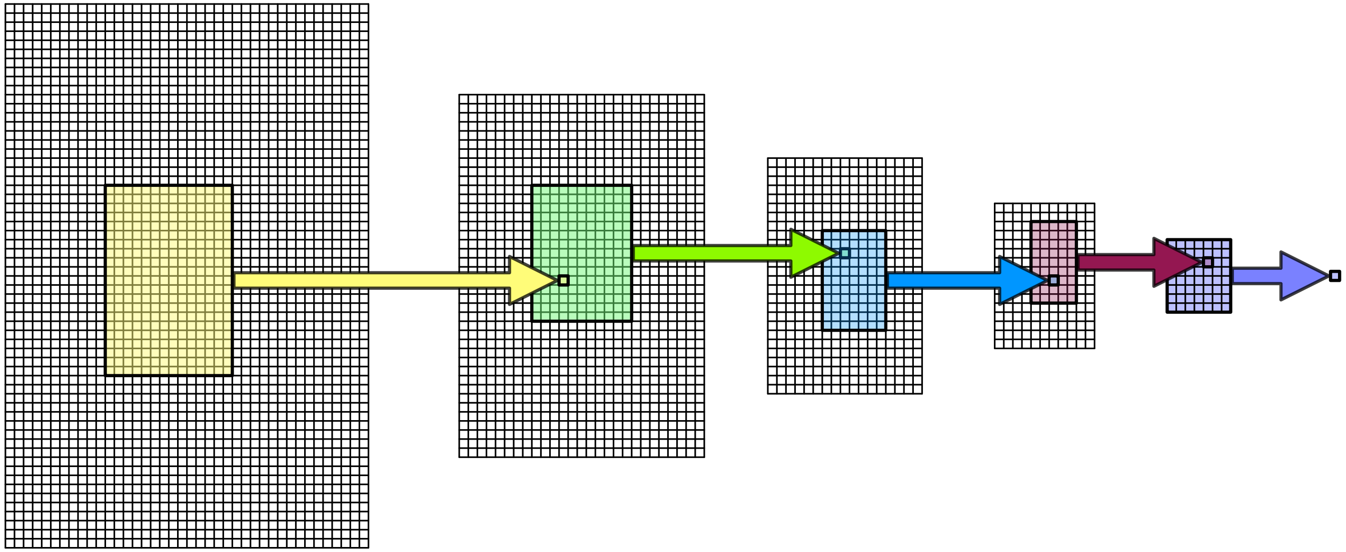
- But computational problematic: very long computation chains

Alternative Sequence Processing

- Convolutional neural networks
- Attention

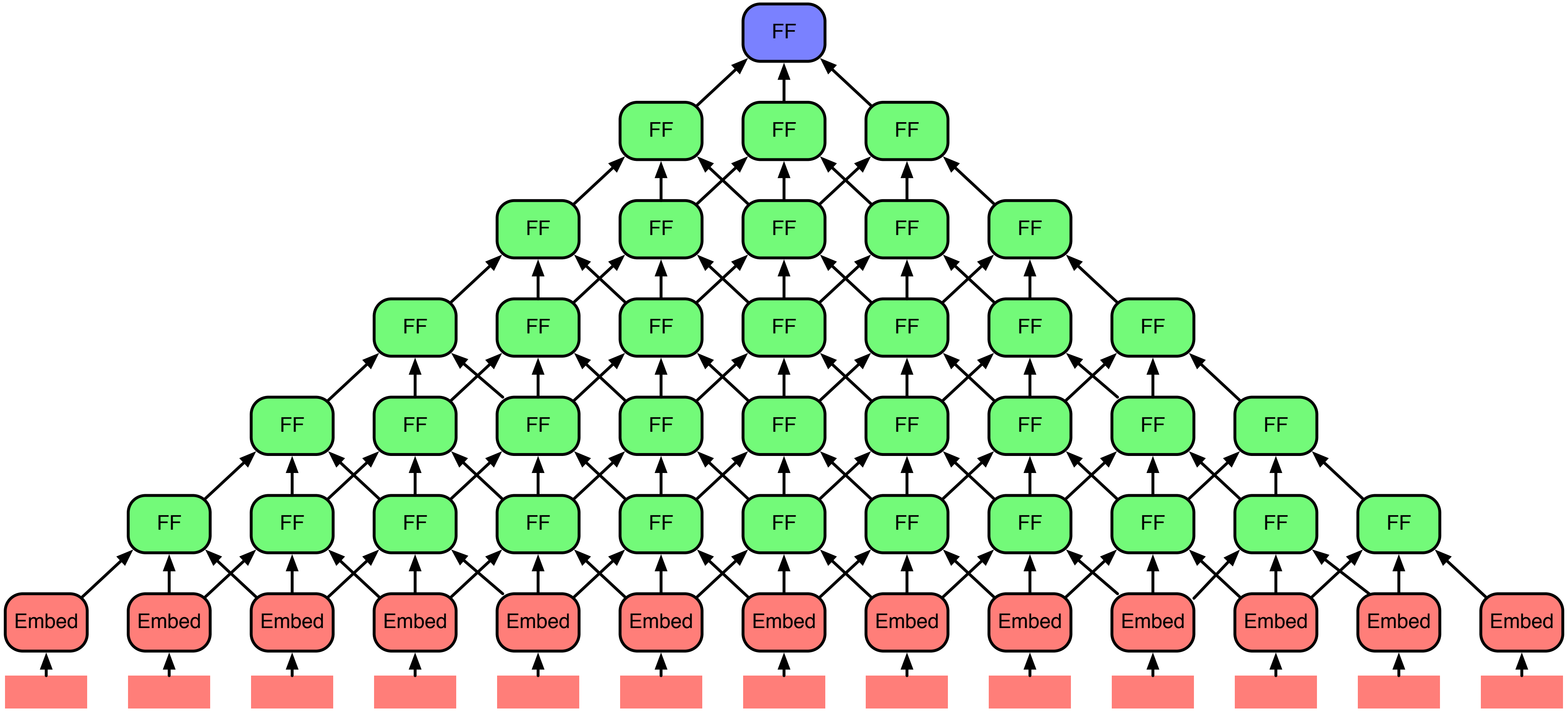
convolutional neural networks

Convolutional Neural Networks (CNN)



- Popular in image processing
- Regions of an image are reduced into increasingly smaller representation
  - matrix spanning part of image reduced to single value
  - overlapping regions

CNNs for Language



- Map words into fixed-sized sentence representation

Hierarchical Structure and Language

- Syntactic and semantic theories of language
  - language is recursive
  - central: verb
  - dependents: subject, objects, adjuncts
  - their dependents: adjectives, determiners
  - also nested: relative clauses
- How to compute sentence embeddings active research topic

Convolutional Neural Networks

- Key step
  - take a high dimensional input representation
  - map to lower dimensional representation

- Several repetitions of this step

- Examples
  - map 50×50 pixel area into scalar value
  - combine 3 or more neighboring words into a single vector

- Machine translation
  - encode input sentence into single vector
  - decode this vector into a sentence in the output language

attention

Attention

- Machine translation is a structured prediction task
  - output is not a single label
  - output structure needs to be built, word by word

- Relevant information for each word prediction varies

- Human translators pay attention to different parts of the input sentence when translating

$\Rightarrow$ Attention mechanism

Computing Attention

- Attention mechanism in neural translation model (Bahdanau et al., 2015)
  - previous hidden state s_{i-1}
  - input word embedding h_j
  - trainable parameters b, W_a, U_a, v_a

$$a(s_{i-1}, h_j) = v_a^T \tanh(W_a s_{i-1} + U_a h_j + b)$$

- Other ways to compute attention
  - Dot product: $a(s_{i-1}, h_j) = s_{i-1}^T h_j$
  - Scaled dot product: $a(s_{i-1}, h_j) = \frac{1}{\sqrt{|h_j|}} s_{i-1}^T h_j$
  - General: $a(s_{i-1}, h_j) = s_{i-1}^T W_a h_j$
  - Local: $a(s_{i-1}) = W_a s_{i-1}$

Attention of Luong et al. (2015)

- Luong et al. (2015) demonstrate good results with the dot product

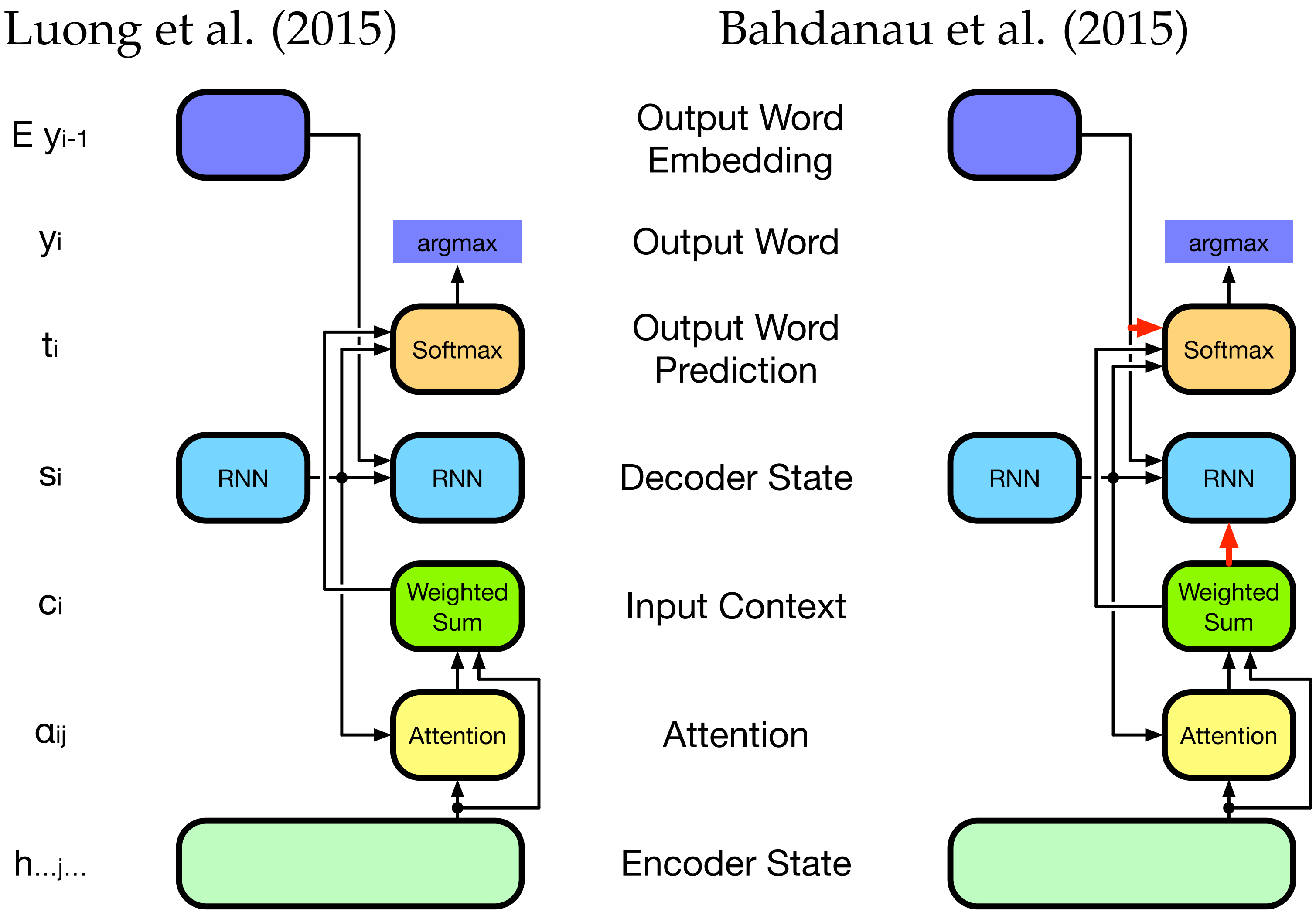
$$a(s_{i-1}, h_j) = s_{i-1}^T h_j$$

- No trainable parameters

- Additional changes

- Currently more popular

Attention of Luong et al. (2015)



Attention of Luong et al. (2015)

Luong et al. (2015)

Attention

$\alpha_{ij} = \text{softmax FF}(s_{i-1}, h_j)$

Input context $c_i = \sum_j \alpha_{ij} h_j$

Output word

$p(y_t|y_{<t}, x) =$

$\text{softmax}\big(W\ \text{FF}_{\text{tanh}}(s_{i-1}, c_i)\big)$

Decoder state

$s_i = \text{FF}_{\text{tanh}}(s_{i-1},\ Ey_{i-1})$

Bahdanau et al. (2015)

Attention

$\alpha_{ij} = \text{softmax FF}(s_{i-1}, h_j)$

Input context $c_i = \sum_j \alpha_{ij} h_j$

Output word

$p(y_t|y_{<t}, x) =$

$\text{softmax}\big(W\ \text{FF}_{\text{tanh}}(s_{i-1}, Ey_{i-1}, c_i)\big)$

Decoder state

$s_i = \text{FF}_{\text{tanh}}(s_{i-1},\ Ey_{i-1}, c_i)$

Multi-Head Attention

- Add redundancy
  - say, 16 attention weights
  - each based on its own parameters
- Formally, for each head k compute an associated between
  - decoder state s_{i-1} at time step i
  - encoder state h_j for the jth input word
  - using the softmax of some parameterized function a^k

$$\alpha_{ij}^k = \text{softmax } a^k(s_{i-1}, h_j)$$

- Average the attention weights

$$\alpha_{ij} = \frac{1}{k} \sum_k \alpha_{ij}^k$$

- Multi-head attention is a form of ensembling

Fine-Grained Attention

- Why just use a single scalar value to weight entire vectors?
  - learn weights for each element
  - computation of attention values returns vector instead of scalar
- Architecturally, still a feed-forward neural network (or any of variants)

$$a(s_{i-1}, h_j) = \mathrm{FF}^k(s_{i-1}, h_j)$$

- Softmax is now applied over each dimension d

$$\alpha_{ij}^d = \frac{\exp a^d(s_{i-1}, h_j)}{\sum_k a^d(s_{i-1}, h_k)}$$

- Input context is now computed by a element-wise multiplication

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

Self Attention

- Finally, a very different take at attention
- Motivation so far: need for alignment between input words and output words
- Now: refine representation of input words in the encoder
  - representation of an input word mostly depends on itself
  - but also informed by the surrounding context
  - previously: recurrent neural networks (considers left or right context)
  - now: attention mechanism
- Self attention:

  Which of the surrounding words is most relevant to refine representation?

Self Attention

- Formal definition (based on sequence of vectors h_j, packed into matrix H

$$\text{self-attention}(H) = \text{softmax}\Big(\frac{HH^T}{\sqrt{|h|}}\Big)H$$

- Association between every word representation h_j any other context word h_k
  - computed by dot product
  - results in a vector of raw association values

$$HH^T$$

- Scaled by the size of the word representation vectors $|h|$, and softmax

$$\text{softmax}\Big(\frac{HH^T}{\sqrt{|h|}}\Big)$$

- Resulting vector of normalized association values used to weigh context words

Self Attention

- More familiar math, using word representation vectors h_j

- Raw association $\frac{HH^T}{\sqrt{|h|}}$

$$a_{jk} = \frac{1}{|h|} h_j h_k^T$$

- Normalized association (softmax)

$$\alpha_{jk} = \frac{\exp(a_{jk})}{\sum_\kappa \exp(a_{j\kappa})}$$

- Weighted sum

$$\text{self-attention}(h_j) = \sum_k \alpha_{j\kappa} h_k$$

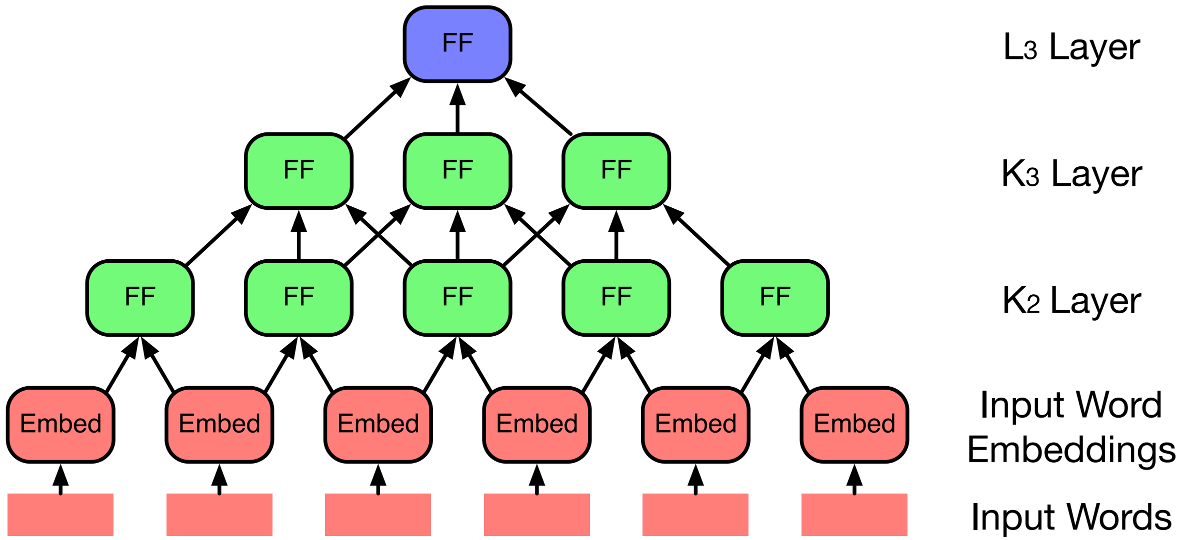
- More on this later (Transformer)

convolutional machine translation

Convolutional Machine Translation

- First end-to-end neural machine translation model of the modern era [Kalchbrenner and Blunsom, 2013]

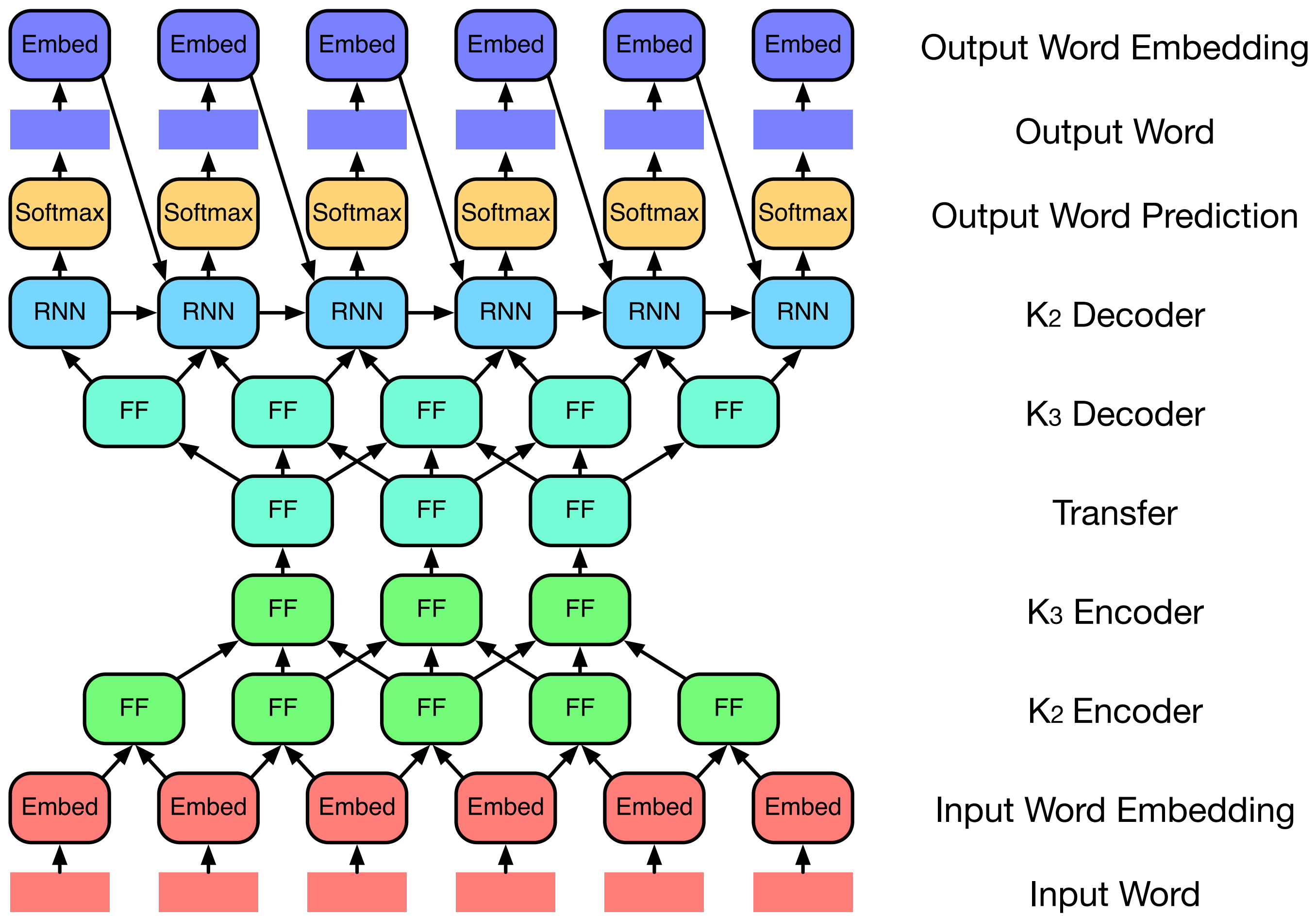
- Encoder



  - always two convolutional layers, with different size
  - here: K_2 and K_3

- Decoder similar

Refinement



- Convolutions do not result in a single sentence embedding but a sequence
- Decoder is also informed by a recurrent neural network

CNNs With Attention

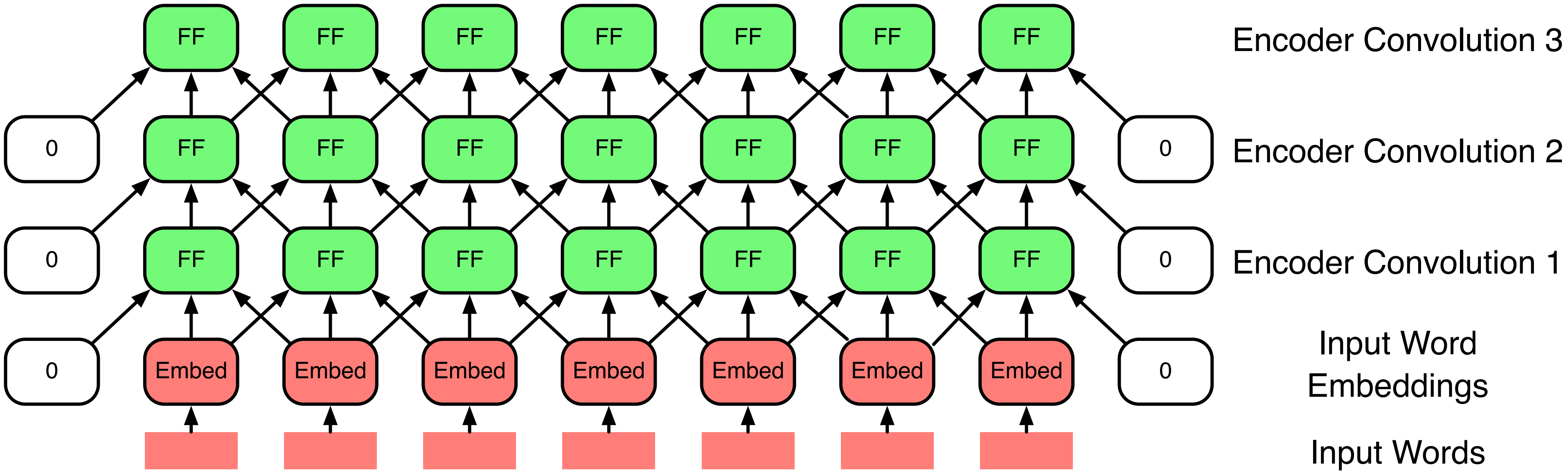
[Gehring et al. 2017]

- Combination of
  - convolutional neural networks
  - attention

- Sequence-to-sequence attention, mainly as before

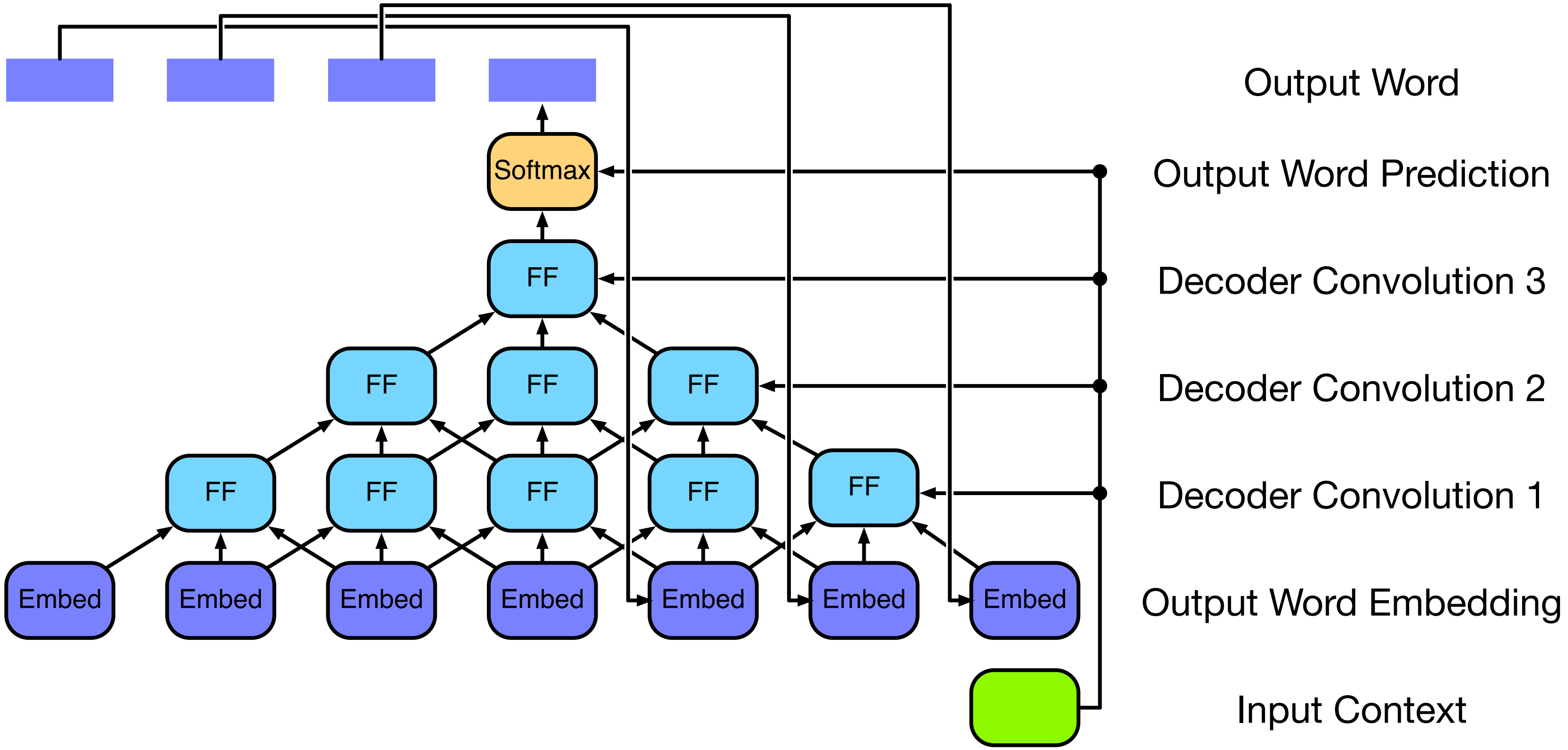
- Recurrent neural networks replaced by convolutional layers

Encoder



- Stacked encoder convolutions
- Not shortening representations
- But: faster processing due to more parallelism

Decoder



- Decoder state computed by convolutional layers over previous output words
- Each convolutional state also informed by the input context (using attention)

transformer

Self Attention: Transformer

- Self-attention in encoder
  - refine word representation based on relevant context words
  - relevance determined by self attention

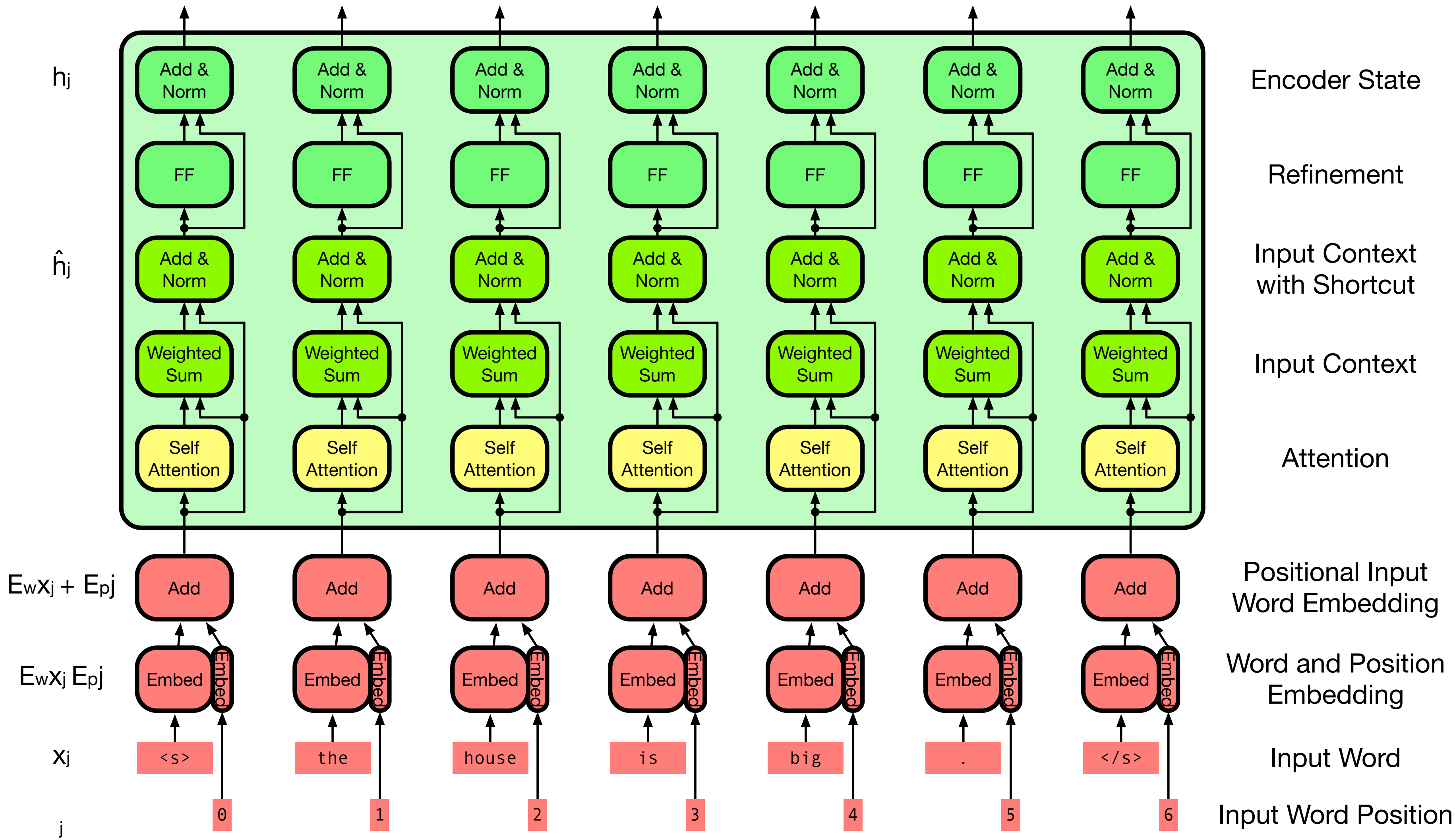
- Self-attention in decoder
  - refine output word predictions based on relevant previous output words
  - relevance determined by self attention

- Also regular attention to encoder states in decoder

- Currently most successful model

  (maybe only with self attention in decoder, but regular recurrent decoder)

Encoder



Sequence of self-attention layers

Self Attention Layer

- Given: input word representations h_j, packed into a matrix $H = (h_1, ..., h_j)$
- Self attention

$$\text{self-attention}(H) = \text{softmax}\Big(\frac{HH^T}{\sqrt{|h|}}\Big)H$$

- Shortcut connection

$$\text{self-attention}(h_j) + h_j$$

- Layer normalization

$$\hat{h}_j = \text{layer-normalization}(\text{self-attention}(h_j) + h_j)$$

- Feed-forward step with ReLU activation function

$$\text{relu}(W\hat{h}_j + b)$$

- Again, shortcut connection and layer normalization

$$\text{layer-normalization}(\text{relu}(W\hat{h}_j + b) + \hat{h}_j)$$

Stacked Self Attention Layers

- Stack several such layers (say, $D = 6$)

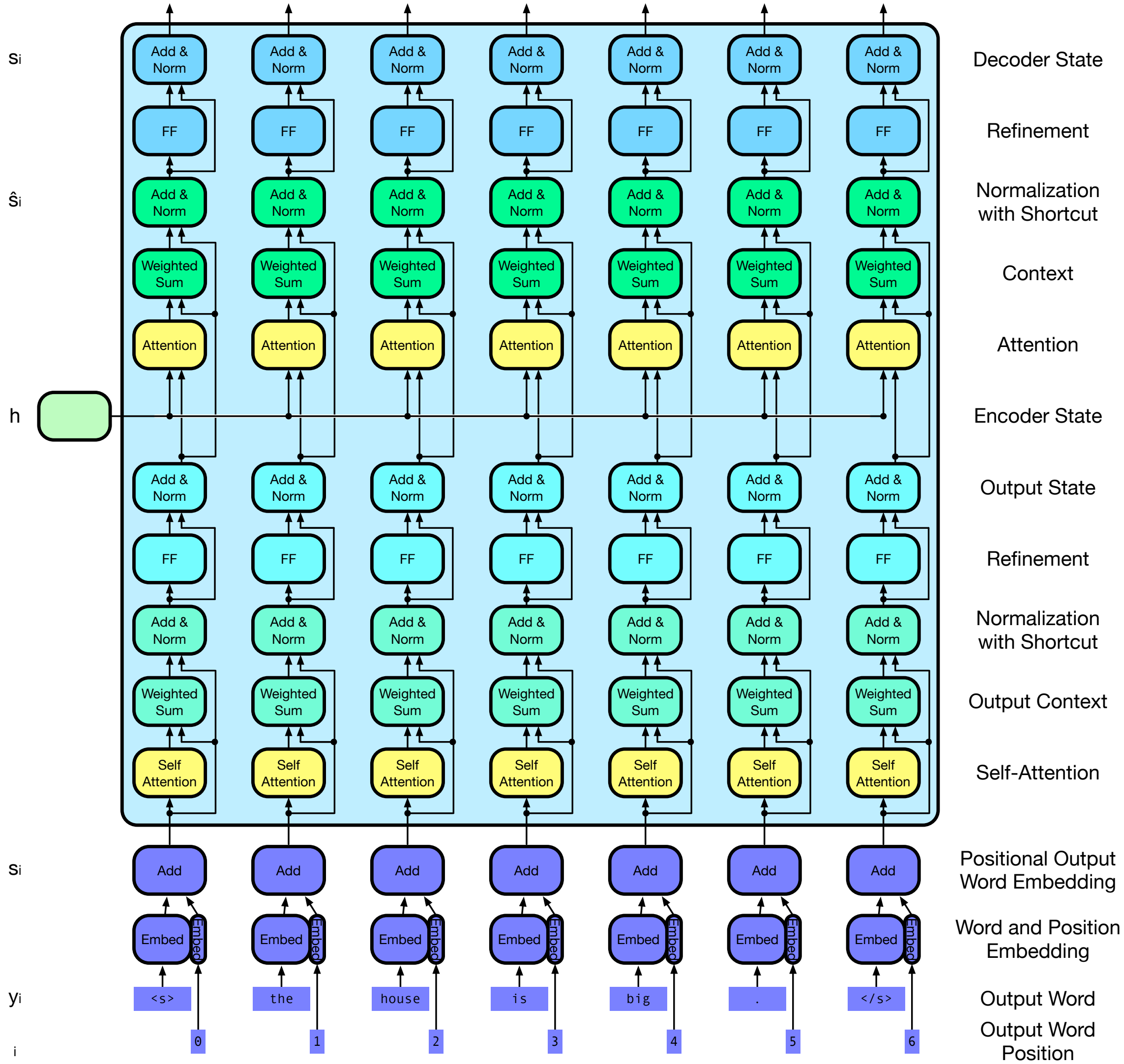
- Start with input word embedding

$$h_{0,j} = Ex_j$$

- Stacked layers

$$h_{d,j} = \text{self-attention-layer}(h_{d-1,j})$$

Decoder



Decoder computes attention-based representations of the output in several layers, initialized with the embeddings of the previous output words

Self-Attention in the Decoder

- Same idea as in the encoder
- Output words are initially encoded by word embeddings $s_i = Ey_i$.
- Self attention is computed over previous output words
  - association of a word s_i is limited to words s_k $(k \leq i)$
  - resulting representation $\tilde{s}_i$

$$\text{self-attention}(\tilde{S}) = \text{softmax}\Big(\frac{SS^T}{\sqrt{|h|}}\Big)S$$

Attention in the Decoder

- Original intuition of attention mechanism: focus on relevant input words
- Computed with dot product $\tilde{S}H^T$
- Compute attention between the decoder states $\tilde{S}$ and the final encoder states H

$$\text{attention}(\tilde{S}, H) = \text{softmax}\Big(\frac{\tilde{S}H^T}{\sqrt{|h|}}\Big)H$$

- Note: attention mechanism formally mirrors self-attention

Full Decoder