

RNN and Attention

Author: Anton Zhitomirsky

Contents

1 RNN Basics	3
1.1 Sequential Data	3
1.2 Advantages	3
1.3 Simple Architecture	3
1.3.1 Mathematical Form	4
1.4 Training RNNs	4
1.4.1 Back-propagation through time (BPTT)	4
1.4.2 Gradient vanishing/explosion issues	5
2 Long Short-Term Memory (LSTM)	7
2.1 Forward Pass	7
2.2 Backwards Pass	8
2.2.1 *Gradient computation	8
2.3 How LSTMs alleviate training difficulties of RNNs	10
2.4 Alternatives — Gated Recurrent Unit (GRU)	10
2.4.1 LSTM vs GRU	11
2.5 Stacking LSTMs	11
2.6 Bidirectional LSTMs	11
3 Sequence-to-sequence models	12
3.1 Sequence decoder inputs during training/testing	13
3.2 *Generative models for sequences — Sequence VAE	15
3.3 State-space models	16
3.3.1 Hidden Markov Model	17
3.3.2 Training	18
4 Attention & Transformers	19
4.1 Motivation	19
4.2 NMT model	19

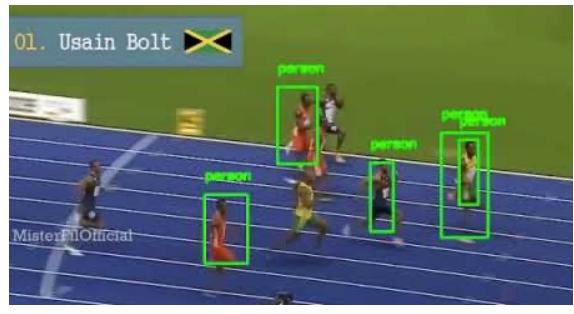
4.3	Attention in Transformers — Single-head attention	21
4.3.1	Hard-attention	22
4.3.2	Soft-attention	22
4.3.3	Msked-attention	22
4.3.4	Complexity analysis	23
4.4	Attention in Transformers — Multi-head attention	23
4.5	Transformer Architecture	24
5	Advances and Applications	28
5.1	Attention Applications	28
5.2	Transformer in NLP	29
5.3	Multi-head Attention in other applications	30
5.4	Efficient Transformers	32
5.5	Memorisation in Transformers	34
A	RNN Notes	35
A.1	Tricks to fix the gradient vanishing/explosion issue	35
B	Transformer Notes	35
B.1	Position encoding	35
B.2	Layer normalization	36
B.3	Point-wise feed-forward network	36

1 RNN Basics

1.1 Sequential Data

Object tracking in videos:

- Data sequence: (x_1, \dots, x_T)
 - x_t : video frame at time t
- Label sequence: (y_1, \dots, y_T)
 - y_t : object identifier, bounding box coordinates, ... at time t
- Goal: learn a mapping $(x_1, \dots, x_T) \rightarrow (y_1, \dots, y_T)$

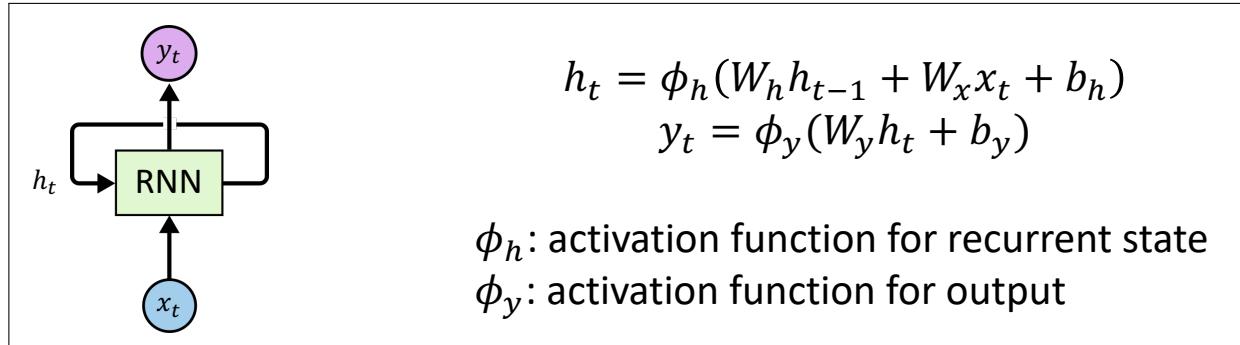


This kind of application works better than a supervised learning context (With labels x and y) because this bounding box information has high sequential time dependencies

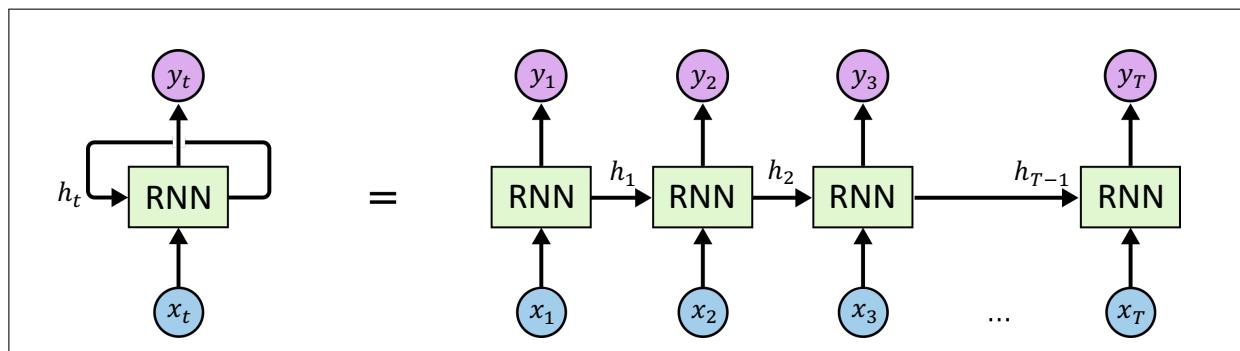
1.2 Advantages

- Model dependencies within the sequence
- Can handle inputs/outputs of different lengths

1.3 Simple Architecture



A recurrent state is maintained throughout the network to maintain information that the network has seen in the past.



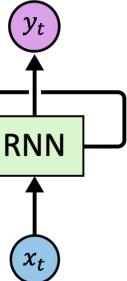
These architectures can be unrolled through time

1.3.1 Mathematical Form

Assume we want to build a neural network to process a data sequence $\mathbf{x}_{1:T} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$. Recurrent neural networks (RNNs) are neural networks suited for processing sequential data, which, if well trained, can model dependencies within a sequence of arbitrary length.

We also assume a supervised learning task which aims to learn the mapping from inputs $\mathbf{x}_{1:T}$ to outputs $\mathbf{y}_{1:T} = (\mathbf{y}_1, \dots, \mathbf{y}_T)$. Then a simple RNN computes the following mapping for $t = 1, \dots, T$:

$$\begin{aligned} \mathbf{h}_t &= \phi_h(W_h \mathbf{h}_{t-1} + W_x \mathbf{x}_t + \mathbf{b}_h), \\ \mathbf{y}_t &= \phi_y(W_y \mathbf{h}_t + \mathbf{b}_y). \end{aligned} \quad (1) \quad (2)$$



Here the network parameters are $\boldsymbol{\theta} = \{W_h, W_x, W_y, \mathbf{b}_h, \mathbf{b}_y\}$, ϕ_h and ϕ_y are the non-linear activation functions for the hidden state \mathbf{h}_t and the output y_t , respectively. For $t = 1$ the convention is to set $\mathbf{h}_0 = \mathbf{0}$ so that $\mathbf{h}_1 = \phi_h(W_x \mathbf{x}_1 + \mathbf{b}_h)$; alternatively \mathbf{h}_0 can also be added to $\boldsymbol{\theta}$ as a learnable parameter.

1.4 Training RNNs

1.4.1 Back-propagation through time (BPTT)

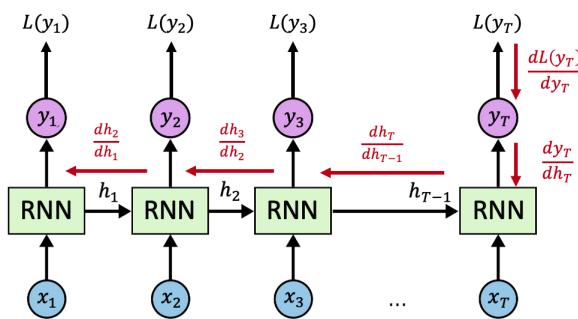


Figure 1: Visualising Back-propagation through time (BPTT) without truncation. The black arrows show forward pass computations, while the red arrows show the gradient back-propagation in order to compute $\nabla_{W_h} \mathcal{L}(y_t)$.

We want to minimise the loss of $\mathcal{L}(\boldsymbol{\theta}) = L_{total}(\boldsymbol{\theta}) = \sum_{t=1}^T L(y_t)$ using gradient descent. This means we need to compute the gradient of the loss with respect to the parameters of the model.

A loss function is required for training an RNN. Assuming the following loss function to minimise:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^T \mathcal{L}(\mathbf{y}_t). \quad (3)$$

This is a common form for the loss function in many sequential modelling tasks such as video/audio sequence reconstruction. The derivative of the loss function w.r.t. $\boldsymbol{\theta}$ is $\frac{d}{d\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^T \frac{d}{d\boldsymbol{\theta}} \mathcal{L}(\mathbf{y}_t)$, therefore it remains to compute $\frac{d}{d\boldsymbol{\theta}} \mathcal{L}(\mathbf{y}_t)$ for $\boldsymbol{\theta} = \{W_h, W_x, W_y, \mathbf{b}_h, \mathbf{b}_y\}$.

- Derivative of $\mathcal{L}(\mathbf{y}_t)$ w.r.t. W_y and \mathbf{b}_y :

$$\frac{d\mathcal{L}(\mathbf{y}_t)}{dW_y} = \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{y}_t} \frac{d\mathbf{y}_t}{dW_y}, \quad \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{b}_y} = \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{y}_t} \frac{d\mathbf{y}_t}{d\mathbf{b}_y}.$$

- Derivative of $\mathcal{L}(\mathbf{y}_t)$ w.r.t. W_x and \mathbf{b}_h :

$$\frac{d\mathcal{L}(\mathbf{y}_t)}{dW_x} = \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{y}_t} \frac{dy_t}{d\mathbf{h}_t} \frac{d\mathbf{h}_t}{dW_x}, \quad \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{b}_h} = \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{y}_t} \frac{dy_t}{d\mathbf{h}_t} \frac{d\mathbf{h}_t}{d\mathbf{b}_h}.$$

Here W_x and \mathbf{b}_h contributes to \mathbf{h}_t in two ways: both direct and indirect contributions, where the latter is through \mathbf{h}_{t-1} . This means:

$$\frac{d\mathbf{h}_t}{dW_x} = \frac{\partial \mathbf{h}_t}{\partial W_x} + \frac{d\mathbf{h}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{h}_{t-1}}{dW_x}, \quad \frac{d\mathbf{h}_t}{d\mathbf{b}_h} = \frac{\partial \mathbf{h}_t}{\partial \mathbf{b}_h} + \frac{d\mathbf{h}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{h}_{t-1}}{d\mathbf{b}_h}.$$

Derivations of these derivatives requires the usage of chain rule which will be explained next.

- Derivative of $\mathcal{L}(\mathbf{y}_t)$ w.r.t. W_h : by chain rule, we have

$$\frac{d\mathcal{L}(\mathbf{y}_t)}{dW_h} = \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{h}_t} \frac{d\mathbf{h}_t}{dW_h}$$

where $\frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{h}_t} = \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{y}_t} \frac{d\mathbf{y}_t}{d\mathbf{h}_t}$. Importantly, here the entries in the Jacobian $\frac{d\mathbf{h}_t}{dW_h}$ contains the *total gradient* of $\mathbf{h}_t[i]$ w.r.t. $W_h[m, n]$. It remains to compute $\frac{d\mathbf{h}_t}{dW_h}$ and notice that \mathbf{h}_t depends on \mathbf{h}_{t-1} which also depends on W_h :

$$\frac{d\mathbf{h}_t}{dW_h} = \frac{\partial \mathbf{h}_t}{\partial W_h} + \frac{d\mathbf{h}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{h}_{t-1}}{dW_h}. \quad (4)$$

Here the entries in $\frac{\partial \mathbf{h}_t}{\partial W_h}$ contains *partial gradient* only (by treating \mathbf{h}_{t-1} as a constant w.r.t. W_h , note that \mathbf{h}_t depends on \mathbf{h}_{t-1}). By expanding the $\frac{d\mathbf{h}_{t-1}}{dW_h}$ term further, we have:

$$\begin{aligned} \frac{d\mathbf{h}_t}{dW_h} &= \frac{\partial \mathbf{h}_t}{\partial W_h} + \frac{d\mathbf{h}_t}{d\mathbf{h}_{t-1}} \frac{\partial \mathbf{h}_{t-1}}{\partial W_h} + \frac{d\mathbf{h}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{h}_{t-1}}{d\mathbf{h}_{t-2}} \frac{d\mathbf{h}_{t-2}}{dW_h} = \dots \\ &= \sum_{\tau=1}^t \left(\prod_{l=\tau}^{t-1} \frac{d\mathbf{h}_{l+1}}{d\mathbf{h}_l} \right) \frac{\partial \mathbf{h}_\tau}{\partial W_h}, \end{aligned} \quad (5)$$

with the convention that when $\tau = t$, $\prod_{l=t}^{t-1} \frac{d\mathbf{h}_{l+1}}{d\mathbf{h}_l} = 1$. This means the chain rule of the gradients needs to be computed in an reversed order from time $t = T$ to time $t = 1$, hence the name *Back-propagation through time (BPTT)*. A visualisation of BPTT is provided in Figure 1. Truncation with length L might be applied to this back-propagation procedure, and with *truncated BPTT* the gradient is computed as

$$\text{truncate}\left[\frac{d\mathbf{h}_t}{dW_h}\right] = \sum_{\tau=\max(1,t-L)}^t \left(\prod_{l=\tau}^{t-1} \frac{d\mathbf{h}_{l+1}}{d\mathbf{h}_l} \right) \frac{\partial \mathbf{h}_\tau}{\partial W_h}.$$

1.4.2 Gradient vanishing/explosion issues

Depending on the weight matrix and the activation function the product of the partial graident can vanish or explode as the number of time steps increases.

Simple RNNs are often said to suffer from gradient vanishing or gradient explosion issues. To understand this, notice that

$$\frac{d\mathbf{h}_{l+1}}{d\mathbf{h}_l}^\top = \phi'_h(W_h \mathbf{h}_l + W_x \mathbf{x}_{l+1} + \mathbf{b}_h) \odot W_h, \quad (6)$$

Here $\phi'_h(W_h \mathbf{h}_l + W_x \mathbf{x}_{l+1} + \mathbf{b}_h)$ denotes a vector containing element-wise derivatives, and we reload the element-wise product operator for vector $\mathbf{a} \in \mathbb{R}^{d \times 1}$ and $\mathbf{B} \in \mathbb{R}^{d \times d'}$ as the “broadcasting element-wise product” $\mathbf{a} \odot \mathbf{B} := [\underbrace{\mathbf{a}, \dots, \mathbf{a}}_{\text{repeat } d' \text{ times}}] \odot \mathbf{B}$. This means $\prod_{l=\tau}^{t-1} \frac{d\mathbf{h}_{l+1}}{d\mathbf{h}_l}$ contains products of $t - \tau$ copies of W_h and the derivative $\phi'_h(\cdot)$ at time steps $l = \tau, \dots, t - 1$.

Here, the circled dot operator is typically used to represent element-wise multiplication

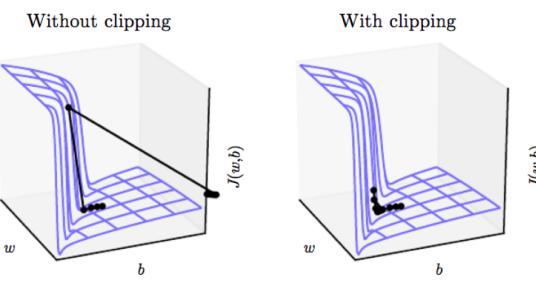
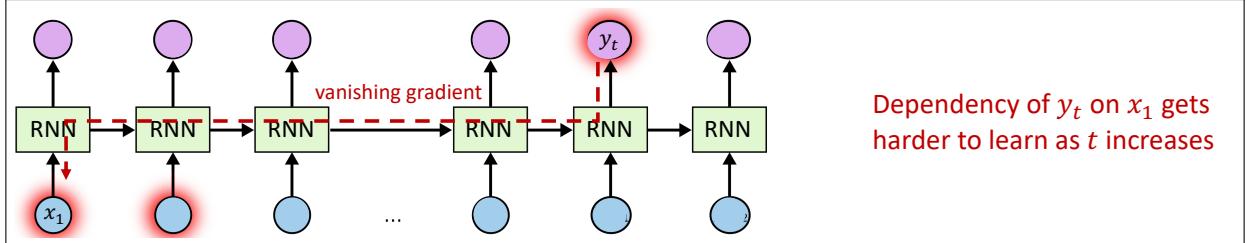


Figure 2: Visualising the gradient step with/out gradient clipping. Source: [Goodfellow et al. \[2016\]](#).

Now consider a simple case where $\phi_h(\cdot)$ is an identity mapping so that $\phi'_h(\cdot) = 1$. We further assume the hidden states have scalar values, i.e. $\dim(\mathbf{h}_t) = 1$. Then we have $\prod_{l=\tau}^{t-1} \frac{d\mathbf{h}_{l+1}}{d\mathbf{h}_l} = (W_h^{t-\tau})^\top$ which can vanish or explode when $t - \tau$ is large, depending on whether $W_h < 1$ or not. In the general case when W_h is a matrix, depending on whether the largest singular value (i.e. maximum of the absolute values of the largest and smallest eigenvalues) of W_h is smaller or larger than 1, the spectral norm of $\prod_{l=\tau}^{t-1} \frac{d\mathbf{h}_{l+1}}{d\mathbf{h}_l} = (W_h^{t-\tau})^\top$ will vanish or explode when $t - \tau$ increases. When $\phi_h(\cdot)$ is selected as the sigmoid function or the hyperbolic tangent function, gradient vanishing problem can still happen. Take hyperbolic tangent function as an example. When entries in \mathbf{h}_t is close to ± 1 , then $\phi'_h(\cdot) \approx 0$, i.e. the derivative is saturated. Multiplying several of such saturated derivatives together also leads to the gradient vanishing problem.

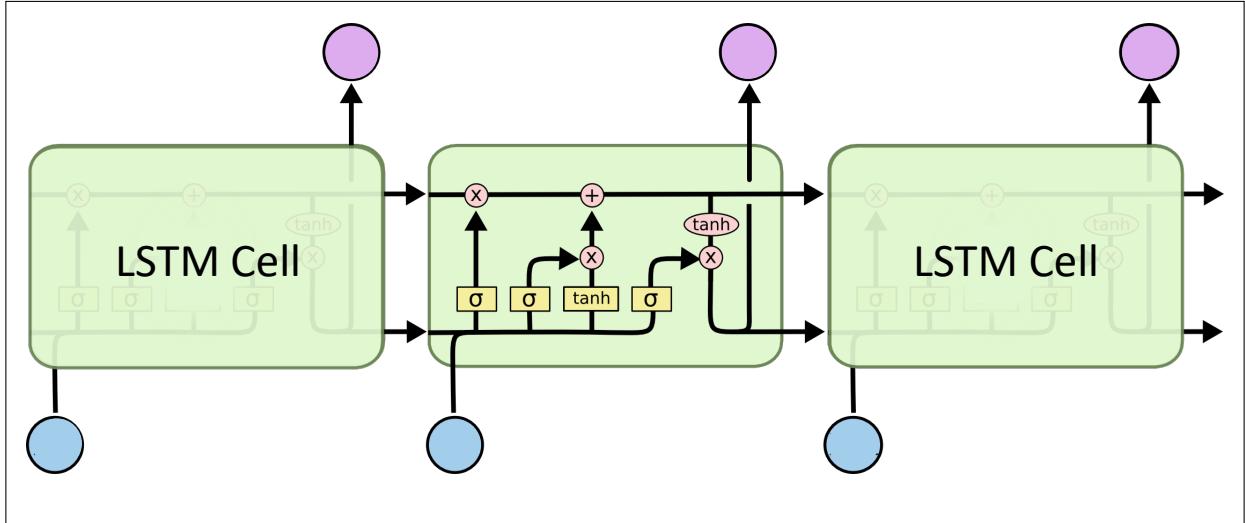
Therefore, long term dependencies are getting harder to learn. With vanishing gradient, Since the gradient for W_h is the sum of back-prop gradient with different window length, this gradient will be dominated especially when $t \gg \tau$; the learning signal will be dominated by the short term dependencies.



Tricks for how to fix the gradient vanishing or explosion issue are detailed in the Appendix A.1.

2 Long Short-Term Memory (LSTM)

2.1 Forward Pass



Addresses the problem of gradient descent problem. They perform better at learning long-term dependencies than simple RNNs

The Long Short-term Memory was proposed with the motivation of addressing the gradient vanishing/explosion problem. It introduces memory cell states and gates to control the error flows, in detail the computation goes as follows (with $\sigma(\cdot)$ as sigmoid function):

$$f_t : \text{forget gate} \quad f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (7)$$

$$i_t : \text{input gate} \quad i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8)$$

$$o_t : \text{output gate} \quad o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

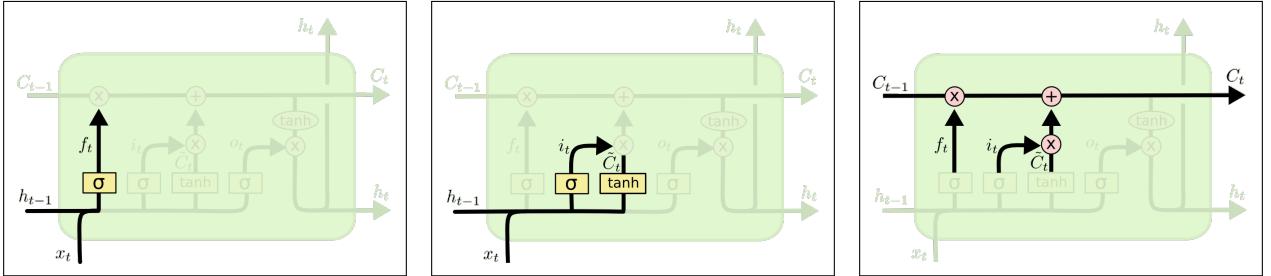
$$x_t : \text{input} \quad \tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (10)$$

$$c_t : \text{memory cell state} \quad c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (11)$$

$$h_t : \text{hidden state} \quad h_t = o_t \odot \tanh(c_t) \quad (12)$$

The parameters of an LSTM are therefore $\theta = \{W_f, W_i, W_o, W_c, b_f, b_i, b_o, b_c\}$. Again by convention, h_0 and c_0 are either set to zero vectors or added to the learnable parameters. We note that if the initial cell state c_0 is initialised to zero, then we have the elements in c_t and h_t bounded within $(-1, 1)$. The output y_t can be produced by a 1-layer neural network similar to the simple RNN case: $y_t = \phi_y(W_y h_t + b_y)$.

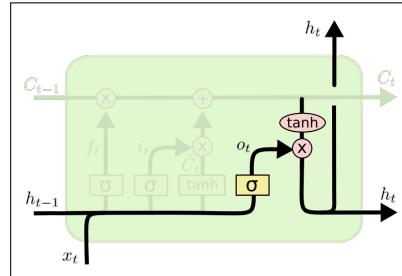
Here, commas (e.g.) $[h_{t-1}, x_t]$ mean concatenations



(a) Forget gate: used to control the cell state update. It depends on the previous hidden state h_{t-1} and current input x_t . This is a one layer network with sigmoid activation function. Equation (7) above.

(b) The second step is to propose a candidate cell state update. It depends on the previous candidate update value \tilde{C}_{t-1} to the cell state. In this case an input gate is also computed to form the cell state value as well. Equations (8) and (10) above.

(c) Cell state update. f_t controls the maintenance of the previous cell states. Similarly, the input gate controls whether the candidate update will be accepted or not. Equation (11) above.



(d) Finally, compute the updates for the hidden states h_t . Here h_t is used to output a cell state c_t . But here, an output gate is also being used. This means sometimes the hidden state will be 0 given the output gate values. Equations (9) and (12) above.

The Prediction of y_t can proceed in a similar way as in simple RNNs: $y_t = \Phi_y(W_y h_t + b_y)$ by transforming the hidden state h_t with one neural network layer.

2.2 Backwards Pass

In this case, the recurrent state we care about is the cell state c_t . Back-prop through time applies to computing the gradient with respect to W_c the weight matrix for computing the candidate cell state.

2.2.1 *Gradient computation

How to derive $\frac{dc_t}{dc_{t-1}}$: Notice c_t depends on c_{t-1} in 4 paths:

direct dependence

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

Depends on $h_{t-1} = o_{t-1} \odot c_{t-1}$ (indirect dependence)

"To give an idea of how this gradient is derived, we can look at the recurrent update equation for the cell state c_t . Here, c_t depends on c_{t-1} as shown in the red box. However, there are also indirect paths via $f_t \wedge i_t \wedge c_t$ because they all depend of h_{t-1} which depends on c_{t-1} . This gives us an idea of how to compute the total gradient."

Readers are encouraged to derive themselves the gradient of $\mathcal{L}(\mathbf{y}_t)$ with respect to θ . Specifically for the recurrent weight matrix W_c , computing the derivative $\frac{d\mathcal{L}(\mathbf{y}_t)}{dW_c}$ requires the following terms:

$$\frac{d\mathcal{L}(\mathbf{y}_t)}{dW_c} = \frac{d\mathcal{L}(\mathbf{y}_t)}{d\mathbf{h}_t} \frac{d\mathbf{h}_t}{dW_c} \quad (13)$$

$$\frac{d\mathbf{h}_t}{dW_c} = \mathbf{o}_t \odot \frac{dtanh(\mathbf{c}_t)}{dW_c} + \tanh(\mathbf{c}_t) \odot \frac{d\mathbf{o}_t}{dW_c} \quad (14)$$

$$\frac{d\mathbf{o}_t}{dW_c} = \frac{d\mathbf{o}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{h}_{t-1}}{dW_c} \quad (15)$$

$$\frac{d\mathbf{c}_t}{dW_c} = \mathbf{f}_t \odot \frac{d\mathbf{c}_{t-1}}{dW_c} + \mathbf{c}_{t-1} \odot \frac{d\mathbf{f}_t}{dW_c} + \mathbf{i}_t \odot \frac{d\tilde{\mathbf{c}}_t}{dW_c} + \tilde{\mathbf{c}}_t \odot \frac{d\mathbf{i}_t}{dW_c}. \quad (16)$$

As \mathbf{h}_{t-1} also depends on \mathbf{c}_{t-1} and \mathbf{o}_{t-1} , it means that

$$\frac{d\mathbf{f}_t}{dW_c} = \frac{d\mathbf{f}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{h}_{t-1}}{dW_c} = \frac{d\mathbf{f}_t}{d\mathbf{h}_{t-1}} \left(\mathbf{o}_{t-1} \odot \frac{dtanh(\mathbf{c}_{t-1})}{dW_c} + \tanh(\mathbf{c}_{t-1}) \odot \frac{d\mathbf{o}_{t-1}}{dW_c} \right) \quad (17)$$

$$\frac{d\mathbf{i}_t}{dW_c} = \frac{d\mathbf{i}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{h}_{t-1}}{dW_c} = \frac{d\mathbf{i}_t}{d\mathbf{h}_{t-1}} \left(\mathbf{o}_{t-1} \odot \frac{dtanh(\mathbf{c}_{t-1})}{dW_c} + \tanh(\mathbf{c}_{t-1}) \odot \frac{d\mathbf{o}_{t-1}}{dW_c} \right) \quad (18)$$

$$\frac{d\tilde{\mathbf{c}}_t}{dW_c} = \frac{\partial \tilde{\mathbf{c}}_t}{\partial W_c} + \frac{d\tilde{\mathbf{c}}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{h}_{t-1}}{dW_c} = \frac{\partial \tilde{\mathbf{c}}_t}{\partial W_c} + \frac{d\tilde{\mathbf{c}}_t}{d\mathbf{h}_{t-1}} \left(\mathbf{o}_{t-1} \odot \frac{dtanh(\mathbf{c}_{t-1})}{dW_c} + \tanh(\mathbf{c}_{t-1}) \odot \frac{d\mathbf{o}_{t-1}}{dW_c} \right). \quad (19)$$

Note again the difference between $\frac{d\tilde{\mathbf{c}}_t}{dW_c}$ and $\frac{\partial \tilde{\mathbf{c}}_t}{\partial W_c}$. The former Jacobian $\frac{\partial \tilde{\mathbf{c}}_t}{\partial W_c}$ has its entries as the *total gradient* of $\tilde{\mathbf{c}}_t[i]$ w.r.t. $W_c[m, n]$, while the latter partial gradient $\frac{\partial \tilde{\mathbf{c}}_t}{\partial W_c}$ has its entries as the *partial gradient* of $\tilde{\mathbf{c}}_t[i]$ w.r.t. $W_c[m, n]$ (by treating \mathbf{h}_{t-1} as a constant w.r.t. W_c , note that $\tilde{\mathbf{c}}_t$ depends on \mathbf{h}_{t-1} as well). Combining the derivations, we have (notice that $\frac{dtanh(\mathbf{c}_{t-1})}{dW_c} = \frac{dtanh(\mathbf{c}_{t-1})}{dc_{t-1}} \frac{dc_{t-1}}{dW_c}$):

$$\frac{d\mathbf{o}_t}{dW_c} = \frac{d\mathbf{o}_t}{d\mathbf{h}_{t-1}} \left(\mathbf{o}_{t-1} \odot \frac{dtanh(\mathbf{c}_{t-1})}{dW_c} + \tanh(\mathbf{c}_{t-1}) \odot \frac{d\mathbf{o}_{t-1}}{dW_c} \right) \quad (20)$$

$$\frac{d\mathbf{c}_t}{dW_c} = \underbrace{\left(\mathbf{f}_t + \mathbf{o}_{t-1} \odot \frac{dtanh(\mathbf{c}_{t-1})}{dc_{t-1}} \odot \frac{dc_{t-1}}{d\mathbf{h}_{t-1}} \right) \frac{dc_{t-1}}{dW_c}}_{= \frac{dc_t}{dc_{t-1}}} + \tanh(\mathbf{c}_{t-1}) \odot \frac{d\mathbf{c}_t}{d\mathbf{h}_{t-1}} \frac{d\mathbf{o}_{t-1}}{dW_c} + \mathbf{i}_t \odot \frac{\partial \tilde{\mathbf{c}}_t}{\partial W_c} \quad (21)$$

$$\frac{dc_t}{dh_{t-1}} = \mathbf{c}_{t-1} \odot \frac{df_t}{dh_{t-1}} + \tilde{\mathbf{c}}_t \odot \frac{di_t}{dh_{t-1}} + \mathbf{i}_t \odot \frac{d\tilde{\mathbf{c}}_t}{dh_{t-1}}. \quad (22)$$

This means for computing $\frac{dc_t}{dW_c}$ it requires computing

$$\prod_{l=\tau}^{t-1} \frac{dc_{l+1}}{dc_l} = \prod_{l=\tau}^{t-1} [\mathbf{f}_{l+1} + \mathbf{o}_l \odot \frac{dtanh(\mathbf{c}_l)}{dc_l} \odot \frac{dc_{l+1}}{dh_l}]$$

for all $\tau = 1, \dots, t$. There is no guarantee that this term will not vanish or explode, however the usage of forget gates makes the issue less severe. To see this, notice that by expanding the product term above, we have that it contains terms proportional to $\mathbf{f}_{i+1} \odot \prod_{l=\tau}^i \mathbf{o}_l \odot \frac{dc_{l+1}}{dh_l}$ for $i = \tau + 1, \dots, t - 1$. So if in the forward pass the network sets $\mathbf{f}_{i+1} \rightarrow 0$ (i.e. forgetting the previous cell state), then this will also likely to bring $\mathbf{f}_{i+1} \odot \prod_{l=\tau}^i \mathbf{o}_l \odot \frac{dc_{l+1}}{dh_l} \approx 0$, which is helpful to cope with the gradient explosion problem. On the other hand, $\frac{dc_t}{dW_c}$ also contains terms proportional to $\prod_{l=\tau+1}^i \mathbf{f}_l \odot \mathbf{o}_\tau \odot \frac{dc_{\tau+1}}{dh_\tau}$ for $i = \tau + 1, \dots, t - 1$. This means if the network sets $\mathbf{f}_l \rightarrow 1$ for $l = \tau + 1, \dots, i$ (i.e. maintaining the cell state until at least time $t = i$), then it will be likely that $\prod_{l=\tau+1}^i \mathbf{f}_l \odot \mathbf{o}_\tau \odot \frac{dc_{\tau+1}}{dh_\tau} \approx \mathbf{o}_\tau \odot \frac{dc_{\tau+1}}{dh_\tau}$, which helps prevent the gradient info at time τ from vanishing when $\mathbf{o}_\tau \rightarrow 1$, and thus be helpful to learn longer term dependencies. The gradients $\frac{dc_t}{dW_c}$ and $\frac{d\mathbf{o}_t}{dW_c}$ also require computing products of $\frac{d\mathbf{o}_{i+1}}{dh_i}$ and $\frac{dc_{i+1}}{dh_i}$ terms, and analogous analysis can be done for those product terms. It is worth emphasising again that LSTM does NOT solve the gradient vanishing/explosion problem completely, however empirical evidences have shown that it is easier for LSTMs to learn longer term dependencies when compared with the simple RNN.

2.3 How LSTMs alleviate training difficulties of RNNs

BPTT in LSTMs:

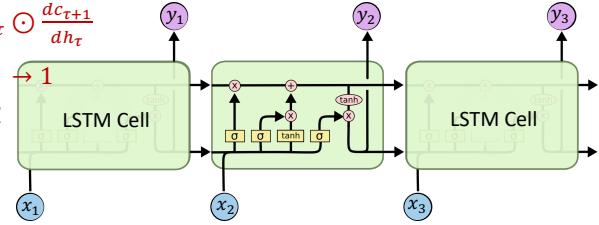
$$\text{Requires computing } \prod_{l=1}^{t-1} \frac{dc_{l+1}}{dc_l} = \prod_{l=1}^{t-1} (f_{l+1} + o_l \odot \frac{dtanh(c_l)}{dc_l} \odot \frac{dc_{l+1}}{dh_l})$$

Alleviating gradient explosion:

$$\begin{aligned} \text{The gradient contains terms proportional to } & f_{i+1} \prod_{l=1}^{i-1} o_l \odot \frac{dc_{l+1}}{dh_l} \\ & \approx 0 \text{ when } f_{i+1} \approx 0 \end{aligned}$$

Alleviating gradient vanishing:

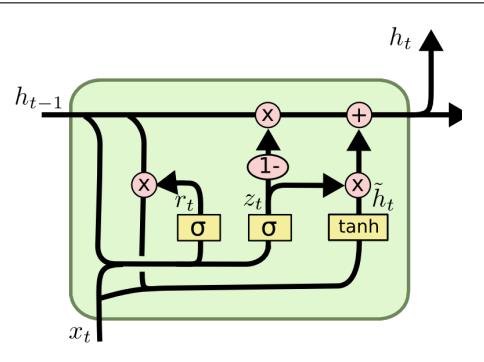
$$\begin{aligned} \text{The gradient contains terms proportional to } & \prod_{l=\tau+1}^i f_l \odot o_\tau \odot \frac{dc_{\tau+1}}{dh_\tau} \\ & \approx o_\tau \odot \frac{dc_{\tau+1}}{dh_\tau} \text{ when } f_l \rightarrow 1 \\ & \text{for } l = \tau + 1, \dots, i \end{aligned}$$



Hochreiter and Schmidhuber (1997). Long Short-Term Memory. Neuro Computation.
Figure adapted from <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

The usage of forget gates is useful for gradient explosion problems, because the network can learn a forget gate value of 0. Similarly, the product of gradients after expansion contains terms like product of f_t terms times other terms. If the forget gates are open, then this term becomes the gradient of $c_{\tau+1}$ with respect to h_τ . If τ is small and i is big, then errors can also be back-propogated and doesn't vanish. This is useful for learning long-term dependencies.

2.4 Alternatives — Gated Recurrent Unit (GRU)



Simplified versions of the LSTM have been proposed. Compared to LSTMs, GRU only uses switching gates z_t and reset gates r_t . you can still see that GRU still implements gating mechanisms to control the updates fo rthe recurrent states.

The Gated Recurrent Unit (GRU) [Cho et al., 2014] improves the simple RNN with the gating mechanism as well. Compared with LSTM, GRU removes the input/output gates and the cell state, but still maintains the forgetting mechanism in some form:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (23)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (24)$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \quad (25)$$

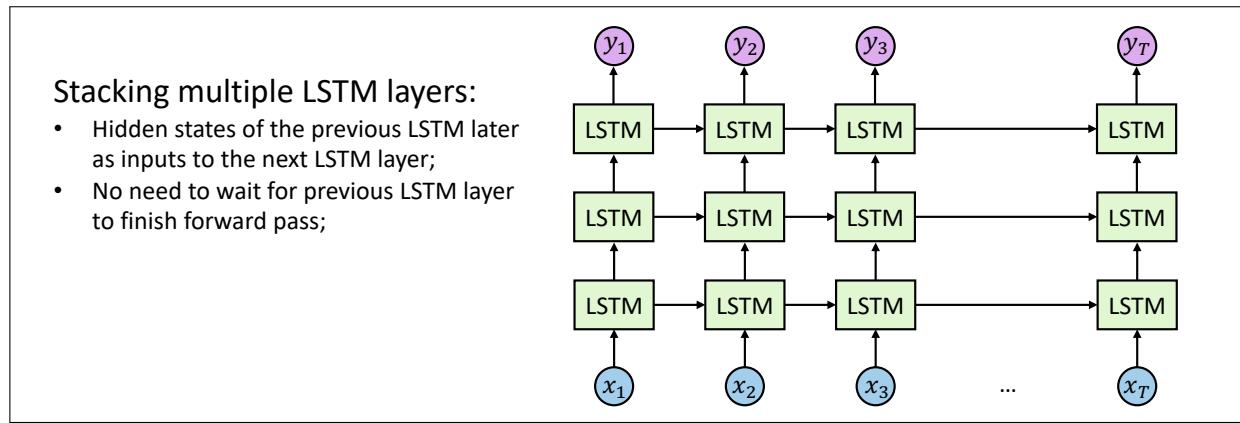
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (26)$$

The network parameters are then $\theta = \{W_z, W_r, W_h, b_z, b_r, b_h\}$. We see here z_t acts as the update gate which determines the incorporation of the current info to the hidden states, and r_t is named the reset gate which also impact on the maintenance of the historical information.

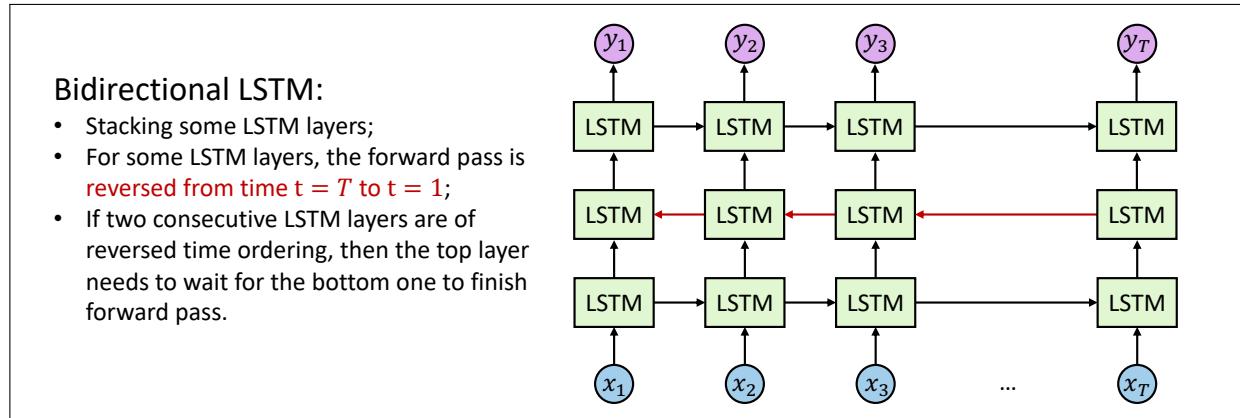
2.4.1 LSTM vs GRU

- Other gated RNN variants exists, but LSTM and GRU are the most widely-used
- GRU is quicker to compute and has fewer parameters
- No conclusive evidence for LSTM > GRU or vice versa
- LSTM is a good default choice (especially if your data has particularly long dependencies, or you have lots of training data)
- Switch to GRU if you want more efficient compute & less overfitting

2.5 Stacking LSTMs



2.6 Bidirectional LSTMs



Has shown to improve performance

3 Sequence-to-sequence models

Given dataset of input-output sequence pairs $(\mathbf{x}_{1:T}, \mathbf{y}_{1:L})$, the goal of sequence prediction is to build a model $p_{\theta}(\mathbf{y}_{1:L}|\mathbf{x}_{1:T})$ to fit the data. Note here that the \mathbf{x} , \mathbf{y} sequences might have different lengths, and the input/output length T and L can vary across input-output pairs. So to handle sequence outputs of arbitrary length, we define an *auto-regressive model*

$$p_{\theta}(\mathbf{y}_{1:L}|\mathbf{x}_{1:T}) = \prod_{l=1}^L p_{\theta}(\mathbf{y}_l|\mathbf{y}_{<l}, \mathbf{v}), \quad \mathbf{v} = enc(\mathbf{x}_{1:T}). \quad (27)$$

Here $p_{\theta}(\mathbf{y}_l|\mathbf{y}_{<l}, \mathbf{v})$ is defined by a sequence decoder, e.g. an LSTM:

$$p_{\theta}(\mathbf{y}_l|\mathbf{y}_{<l}, \mathbf{v}) = p_{\theta}(\mathbf{y}_l|\mathbf{h}_l^d, \mathbf{c}_l^d), \quad \mathbf{h}_l^d, \mathbf{c}_l^d = LSTM_{\theta}^{dec}(\mathbf{y}_{<l}), \quad (28)$$

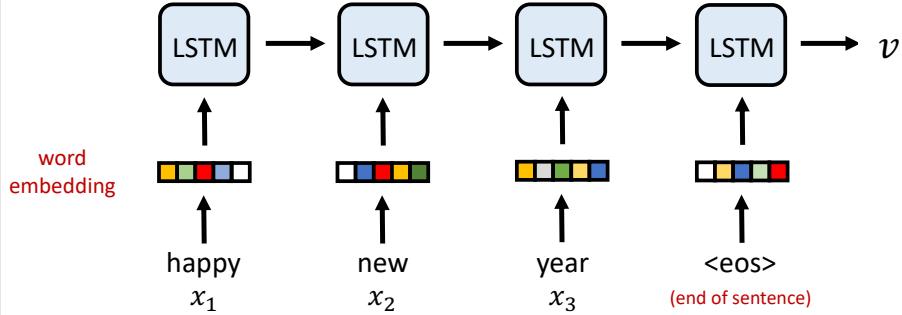
and the decoder LSTM has its internal recurrent states $\mathbf{h}_0^d, \mathbf{c}_0^d$ initialised using the input sequence representation $\mathbf{v} = enc(\mathbf{x}_{1:T})$. The encoder also uses an LSTM, meaning that

$$\mathbf{v} = enc(\mathbf{x}_{1:T}) = NN_{\theta}(\mathbf{h}_T^e, \mathbf{c}_T^e), \quad \mathbf{h}_T^e, \mathbf{c}_T^e = LSTM_{\theta}^{enc}(\mathbf{x}_{1:T}). \quad (29)$$

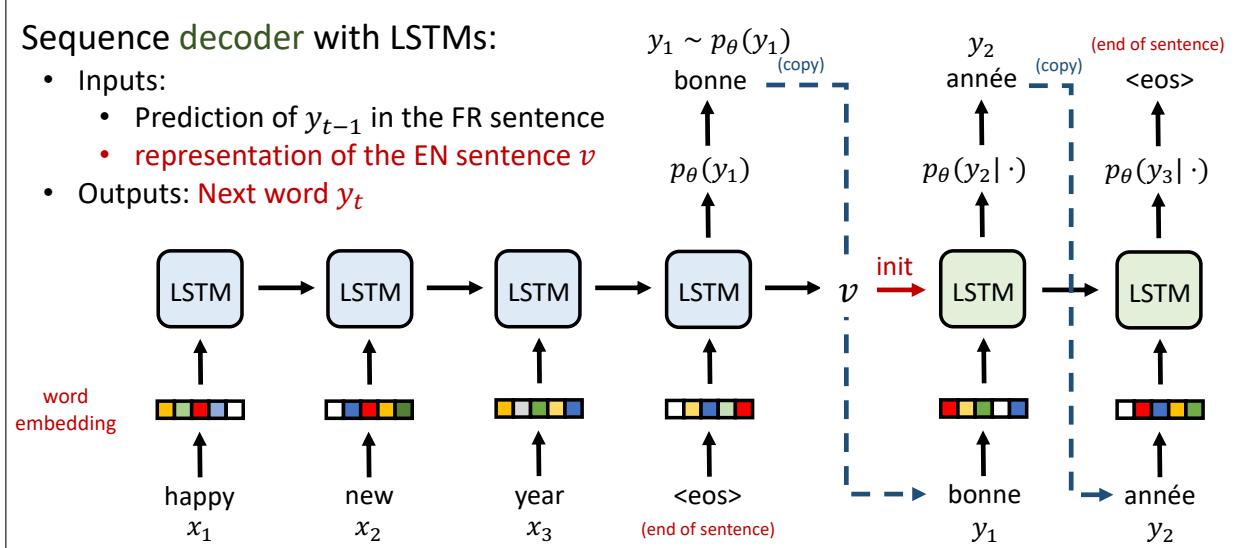
Here, the auto-regression model means that the output y sequence is generated in a sequential way, and the current location y_l depends on the y values at all the previous steps.

Sequence encoder with LSTMs:

- Inputs: word embeddings of the words $x_{1:T}$
- Outputs:
 - representation of the EN sentence v



First, we need a sequence encoder to summarise the information presented in the X sequence. We know that X sequence can have different lengths, but RNNs can handle this situation. Firstly, we map words into word embeddings before we feed them into neural networks. Here, one-hot encoding is inefficient (sparse and many dimensions). Then, use an LSTM or Stacked LSTM to process each of the word embedding vectors. We don't generate outputs at each time-step t here, instead at the end we use the hidden state h_t to fully describe the input sequence X . Note, we also need to have an end of sentence token so it may terminate (good cuz not fixed length).



Now, given the vector representation v , how do we generate the output sequences? Since v is related closely to the last LSTM state h_t in a sequence to sequence model, it can be used to produce the probability of the first word y_1 in the y output sequence. Then the first word will be sampled from this probability vector.

Given y_1 the sequence decoder proceeds to predict the next word in the y sequence. The sequence decoder is also an LSTM. At each time step for prediction, the decoder LSTM takes in the word embeddings of the predicted word in the last step then runs the LSTM equation to update the internal state. Now, the decoder LSTM will predict the probability vector for the output word y_t at each time step. The representation vector v of the input sentence is used to initialise the hidden and cell states of the decoder LSTM. It will run until the end-of-sentence token.

Regarding the prediction for the first output y_1 , either $p_\theta(y_1|x_{1:T})$ can be produced using the last recurrent states of the encoder LSTM (i.e. $p_\theta(y_1|x_{1:T}) = p_\theta(y_1|h_T^e, c_T^e)$), or we can add in a “start of sentence” token as y_0 and compute the probability vector for y_1 using the LSTM decoder: $p_\theta(y_1|x_{1:T}) = p_\theta(y_1|y_0, v)$. This model is named *Sequence-to-sequence* model or *Seq2Seq* model in short, which is proposed by Sutskever et al. [2014].

3.1 Sequence decoder inputs during training/testing

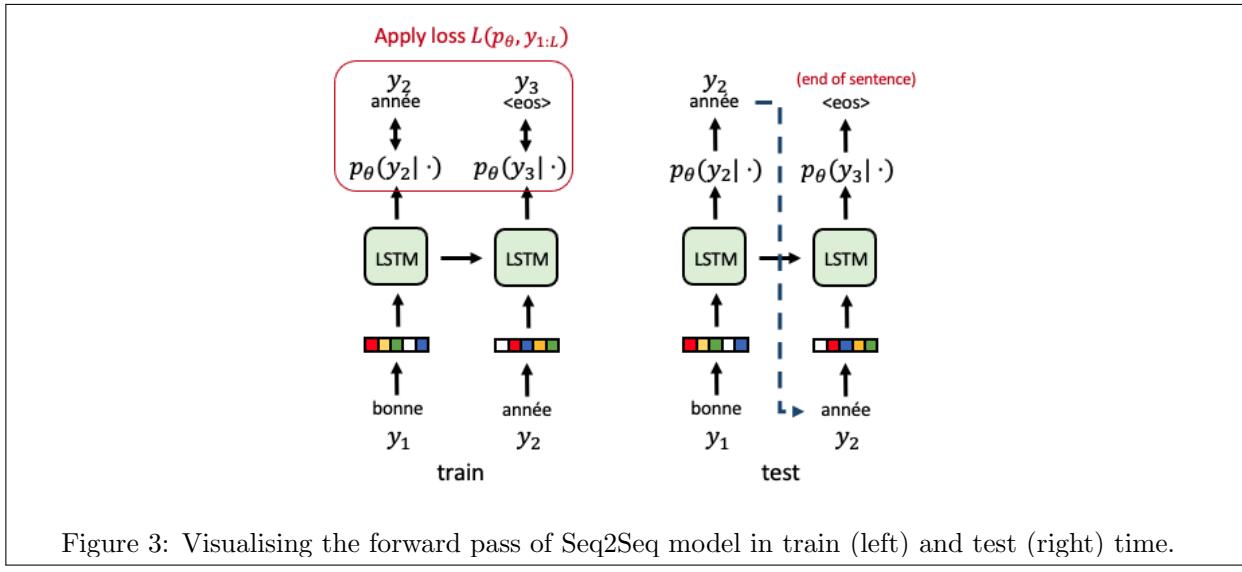


Figure 3: Visualising the forward pass of Seq2Seq model in train (left) and test (right) time.

In training, since maximum likelihood estimation of the network parameters require calculating the conditional probability distribution from the x, y data pairs. This means that the decoder needs to use the input labels y rather than using the predicted y predicted in the previous steps. The probability vectors produced by the decoders in training are only used to compute the training objective (cross-entropy loss) to train the network during back-prop. But during the test, we should use the previous predicted words as the inputs at the current steps.

The decoder LSTM forward pass in training and test times are different, which is visualised in Figure 3. In training, since maximum likelihood training requires evaluating $p_{\theta}(y_l|y_{<l}, v)$ for the output sequences $y_{1:L}$ from the dataset, this means the inputs to the decoder LSTM are words in the data label sequence, and the output of the LSTM is the probability vector for the current word, which is then used to compute the MLE objective (i.e. negative cross-entropy). In test time, however, there is no ground truth output sequence provided, therefore the input to the decoder LSTM at step l is the *predicted* word $y_{l-1} \sim p_{\theta}(y_{l-1}|y_{<l-1}, v)$. In practice prediction is done by e.g. beam search rather than naive sequential sampling [Sutskever et al., 2014].

In NLP applications such as machine translation, both $x_{1:T}$ and $y_{1:L}$ are sequence of words which cannot be directly processed by neural networks. Instead each x_t (y_t) needs to be mapped to a real-value vector before feeding it to the encoder (decoder) LSTM. A naive approach is to use one-hot encoding: assume that the input sentence is in English and we have a sorted English vocabulary of size V , then

$$x_t \rightarrow (\underbrace{0, 0, \dots, 1, \dots, 0}_{k-1}), \quad \text{if } x_t \text{ is the } k\text{th word in the vocabulary.}$$

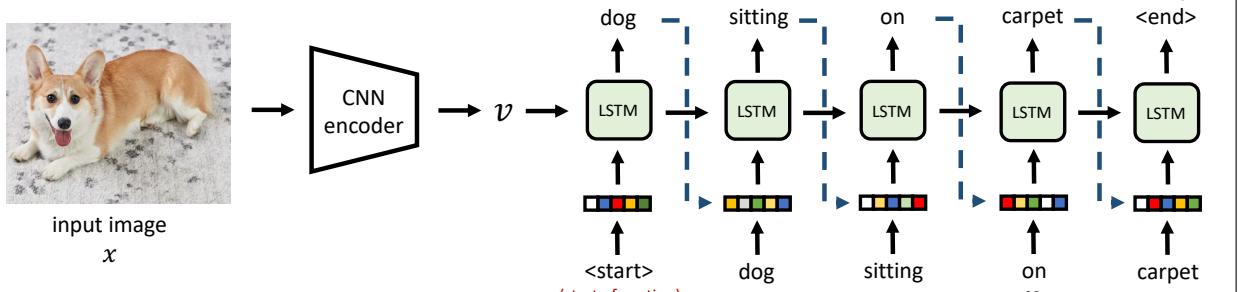
This is clearly inefficient since English vocabulary has tens of thousands of words. Instead, it is recommended to map the words to their *word embeddings* using word2vec [Mikolov et al., 2013a] or GloVe [Pennington et al., 2014], which has much smaller dimensions. But more importantly, semantics are preserved to some extent in these word embeddings, e.g. it has been shown that vector calculus results like “emb(king) - emb(male) + emb(female) = emb(queen)” hold for word2vec embeddings.

In many NLP applications the decoder output is a probability vector $p_{\theta}(y_l|y_{<l}, v)$ which specifies the predictive probability of each of the words in the vocabulary. When the vocabulary is large (which is often so), applying softmax to obtain the probability vector can be very challenging. Interested readers can check e.g. hierarchical softmax [Mikolov et al., 2013a] and negative sampling [Mikolov et al., 2013b] for solutions to mitigate this issue.

This sort of encoding decoding idea can be extended to other applications, such as image captioning.

Image captioning with CNN encoder + LSTM decoder:

- CNN encoder extract representation v of image x
- LSTM decoder generate caption conditioned on v



The goal is to describe an image with a sentence. A popular approach is to ENCODE a feature representation of image x and use this to initialise the decoder LSTM and run the decoder LSTM to produce a sentence.

3.2 *Generative models for sequences — Sequence VAE

To generate sequential data such as video, text and audio, one needs to build a generative model $p_\theta(x_{1:T})$ and train it with e.g. (approximate) maximum likelihood. In the following we discuss two types of latent variable models that are often used in sequence generation tasks.

Similar to VAEs for image generation, one can define a latent variable model with a global latent variable for sequence generation [Fabius and van Amersfoort, 2014; Bowman et al., 2016]:

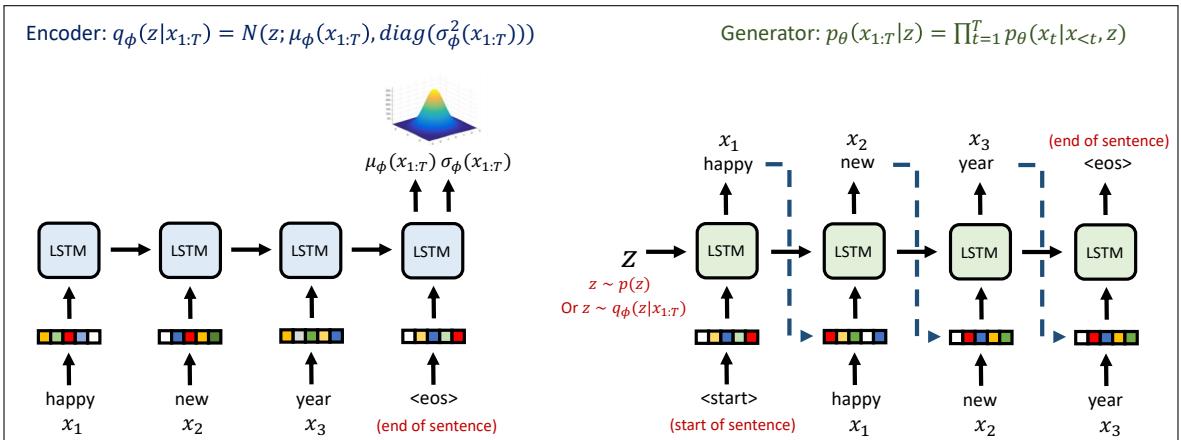
$$p_\theta(\mathbf{x}_{1:T}) = \int p_\theta(\mathbf{x}_{1:T}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \quad (30)$$

If variational lower-bound is used for training, then this also requires an approximate posterior $q_\phi(\mathbf{z}|\mathbf{x}_{1:T})$ to be optimised:

$$\phi^*, \theta^* = \arg \max \mathcal{L}(\phi, \theta), \quad \mathcal{L}(\phi, \theta) = \mathbb{E}_{p_{\text{data}}(\mathbf{x}_{1:T})} [\underbrace{\log p_\theta(\mathbf{x}_{1:T}|\mathbf{z}) - \text{KL}[q_\phi(\mathbf{z}|\mathbf{x}_{1:T})||p(\mathbf{z})]}_{:=\mathcal{L}(\mathbf{x}_{1:T}, \phi, \theta)}]$$

Contains the training objective function. Here, $\log p_\theta(x_{1:T}|\mathbf{z}) = \sum_{t=1}^T \log p_\theta(x_t|x_{<t}, \mathbf{z})$

We may also use a β coefficient in front of the KL term, to control the deregularisation effect. It has shown to be quite important to achieve a good fit to data, and high quality generation results.



The solution is similar to the image captioning task. We can build an auto-regressive model for the conditional distribution of x_1 to x_T given z . This auto-regressive model is parameterised by an LSTM, which takes in the previous x words to produce the current output x_t . The latent variable z can be used to initialise the decoder.

We can use a VAE approach to train these sequence generation models. We need to build an encoder (use LSTM) but here, at the end, the encoder needs to produce distributional parameters of the q distribution like mean and variance. This is similar to the image in VAE case, except now the input and output are sequences from a static image.

Now it remains to define the encoder and decoder distributions, such that they can process sequences of any length. This can be achieved using e.g. LSTMs to define an auto-regressive decoder:

$$p_\theta(\mathbf{x}_{1:T}|z) = \prod_{t=1}^T p_\theta(x_t|\mathbf{x}_{<t}, z), \quad p_\theta(x_1|\mathbf{x}_{<1}, z) = p_\theta(x_1|z). \quad (32)$$

with the distributional parameters of $p_\theta(x_t|\mathbf{x}_{<t}, z)$ defined by $LSTM_\theta(\mathbf{x}_{<t})$ which has its recurrent states $\mathbf{h}_0, \mathbf{c}_0$ initialised using z . For the encoder, LSTMs can also be used to process the input:

$$q_\phi(z|\mathbf{x}_{1:T}) = \mathcal{N}(z; \mu_\phi(\mathbf{x}_{1:T}), \text{diag}(\sigma_\phi^2(\mathbf{x}_{1:T}))), \quad \mu_\phi(\mathbf{x}_{1:T}), \log \sigma_\phi(\mathbf{x}_{1:T}) = LSTM_\phi(\mathbf{x}_{1:T}). \quad (33)$$

3.3 State-space models

- Stochastic dynamic model for the latent state:

$$p_{\theta}(z_{1:T}) = p_{\theta}(z_1) \prod_{t=2}^T p_{\theta}(z_t | z_{t-1})$$

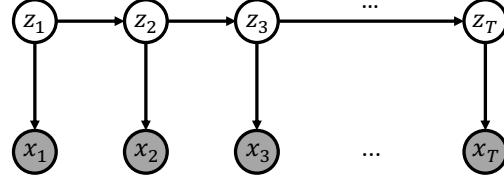
- Emission model:

$$p_{\theta}(x_t | z_t)$$

- Joint distribution:

$$p_{\theta}(x_{1:T}, z_{1:T}) = \prod_{t=1}^T p_{\theta}(z_t | z_{t-1}) p_{\theta}(x_t | z_t)$$

(with the convention that $p_{\theta}(z_1 | z_0) := p_{\theta}(z_1)$)



In a state-space model, it has a stochastic dynamic model for the latent state. In other words, the latent variable z is split, with each z_t associated with each state time step, depending on previous latent states. It also contains an emission model, where the assumption is that the generation of the current x_t depends on the z_t only.

State-space models assume that for every observation \mathbf{x}_t at time t , there is a latent variable \mathbf{z}_t that generates it, and the sequence dynamic model is defined in the latent space rather than in the observation space. In detail, a *prior dynamic model* is assumed on the transitions of the latent states \mathbf{z}_t , often in an auto-regressive way:

$$p_{\theta}(\mathbf{z}_{1:T}) = \prod_{t=1}^T p_{\theta}(\mathbf{z}_t | \mathbf{z}_{<t}), \quad p_{\theta}(\mathbf{z}_1 | \mathbf{z}_{<1}) = p_{\theta}(\mathbf{z}_1). \quad (34)$$

The observation \mathbf{x}_t at time t is assumed to be conditionally dependent on \mathbf{z}_t only, and this conditional distribution is also called the *emission model*:

$$p_{\theta}(\mathbf{x}_t | \mathbf{z}_{1:T}) = p_{\theta}(\mathbf{x}_t | \mathbf{z}_t). \quad (35)$$

Combining both definitions, we have the sequence generative model defined as

$$p_{\theta}(\mathbf{x}_{1:T}) = \int \prod_{t=1}^T p_{\theta}(\mathbf{x}_t | \mathbf{z}_t) p_{\theta}(\mathbf{z}_t | \mathbf{z}_{<t}) d\mathbf{z}_{1:T}. \quad (36)$$

A variational lower-bound objective for training this state-space model require an approximate posterior distribution $q_{\phi}(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})$:

$$\begin{aligned} \mathcal{L}(\phi, \theta) &= \mathbb{E}_{p_{\text{data}}(\mathbf{x}_{1:T})} [\mathbb{E}_{q_{\phi}(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} [\log p_{\theta}(\mathbf{x}_{1:T} | \mathbf{z}_{1:T})] - \text{KL}[q_{\phi}(\mathbf{z}_{1:T} | \mathbf{x}_{1:T}) || p_{\theta}(\mathbf{z}_{1:T})]] \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x}_{1:T})} \left[\mathbb{E}_{q_{\phi}(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} \left[\sum_{t=1}^T \log p_{\theta}(\mathbf{x}_t | \mathbf{z}_t) \right] - \text{KL}[q_{\phi}(\mathbf{z}_{1:T} | \mathbf{x}_{1:T}) || p_{\theta}(\mathbf{z}_{1:T})] \right]. \end{aligned} \quad (37)$$

3.3.1 Hidden Markov Model

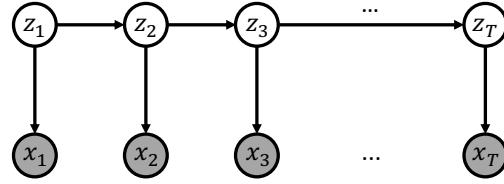
Example: Hidden Markov Model (HMM)

- Stochastic **linear** dynamic model for the latent state:

$$z_t = Az_{t-1} + B\epsilon_t, \epsilon_t \sim N(0, I)$$

- **Linear Gaussian** emission model:

$$x_t = Cz_t + D\psi_t, \psi_t \sim N(0, I)$$



This model assumes a stochastic linear dynamic model from the latent space, meaning that the current latent state x_t is computed by a linear transition of the previous state z_{t-1} plus some gaussian noise.

Also, it has a linear Gaussian emission model. So that the observation at time T is also a linear transformation of the correlated state z_t plus gaussian noise.

This is linear and stochastic; add a non-linearity model to the transition model of z and make the dynamic model a simple and stochastic RNN. Yet, this is difficult to train (see section regarding RNNs)

State-space models + non-linear dynamics:

- Stochastic dynamic model **parameterized by RNNs**:

$$p_\theta(z_{1:T}) = \prod_{t=1}^T p_\theta(z_t|z_{<t})$$

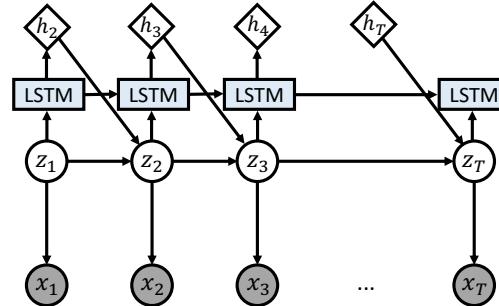
$$p_\theta(z_t|z_{<t}) \neq p_\theta(z_t|z_{t-1}) \quad (\text{LSTM makes all historical states relevant})$$

- **Non-linear** emission model:

$$p_\theta(y_t|z_t): x_t = \mu_\theta^x(z_t) + \sigma_\theta^x(z_t)\psi_t, \psi_t \sim N(0, I)$$

- Joint distribution:

$$p_\theta(x_{1:T}, z_{1:T}) = \prod_{t=1}^T p_\theta(z_t|z_{<t})p_\theta(x_t|z_t)$$



A typical solution for parameterising the stochastic dynamic model with LSTMs. The idea is to treat the previous latent space z_{t-1} as input to the LSTM at time t and ask the LSTM to produce the distribution parameters (mean and variance) for the current latent state z_t .

Now the dynamic model is a combination of deterministic and stochastic models. We can make it non-linear. See slide.

The joint distribution contains the conditional distributions of the latent dynamic model and the emission model. For the emission model it is similar to writing the conditional distribution in the image generative model case.

TODO WTF IS GOING ON??

3.3.2 Training

Training:

$$E_{p_{\text{data}}(\mathbf{x}_{1:T})}[\log p_{\theta}(\mathbf{x}_{1:T})] \geq E_{p_{\text{data}}(\mathbf{x}_{1:T})}[E_{q_{\phi}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}[\log p_{\theta}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})] - KL[q_{\phi}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})||p_{\theta}(\mathbf{z}_{1:T})]]$$

Prior parameters to be learned!

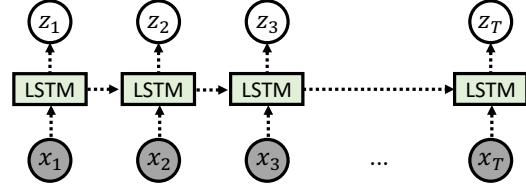
- Generative model: $p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^T p_{\theta}(\mathbf{z}_t|\mathbf{z}_{<t}) p_{\theta}(\mathbf{x}_t|\mathbf{z}_t)$

- Designing an LSTM-based encoder:

$$q_{\phi}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^T q_{\phi}(\mathbf{z}_t|\mathbf{x}_{\leq t})$$

$$q_{\phi}(\mathbf{z}_t|\mathbf{x}_{\leq t}) = N(\mathbf{z}_t; \mu_{\phi}^z(h_t^e), \text{diag}(\sigma_{\phi}^z(h_t^e)^2))$$

$$[h_t^e, c_t^e] = LSTM_{\phi}(\mathbf{x}_t, h_{t-1}^e, c_{t-1}^e)$$



TODO WTF IS GOING ON!!!!

Different from the image generation case, now the prior distribution $p_{\theta}(\mathbf{z}_{1:T})$ also has learnable parameters in θ , so in this case it is less appropriate to view this KL term as a “regulariser”.

The expanded expression for the variational lower-bound depends on the definition of the encoder distribution $q_{\theta}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$. The simplest solution is to use a factorised approximate posterior

$$q_{\phi}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^T q(\mathbf{z}_t|\mathbf{x}_{\leq t}), \quad (38)$$

and the variational lower-bound becomes

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{p_{\text{data}}(\mathbf{x}_{1:T})} \left[\sum_{t=1}^T \mathbb{E}_{q(\mathbf{z}_{\leq t}|\mathbf{x}_{\leq t})} \left[\underbrace{\mathbb{E}_{q_{\phi}(\mathbf{z}_t|\mathbf{x}_{\leq t})}[\log p_{\theta}(\mathbf{x}_t|\mathbf{z}_t)] - KL[q_{\phi}(\mathbf{z}_t|\mathbf{x}_{\leq t})||p_{\theta}(\mathbf{z}_t|\mathbf{z}_{<t})]}_{:=\mathcal{L}(\mathbf{x}_t, \phi, \theta, \mathbf{z}_{<t})} \right] \right]. \quad (39)$$

We see that the term $\mathcal{L}(\mathbf{x}_t, \phi, \theta, \mathbf{z}_{<t})$ resembles the VAE objective in the image generation case, except that the prior distribution $p_{\theta}(\mathbf{z}_t|\mathbf{z}_{<t})$ is conditioned on the previous latent states $\mathbf{z}_{<t}$ rather than a standard Gaussian, and the q distribution takes $\mathbf{x}_{\leq t}$ as the input rather than a single frame \mathbf{x}_t .

Neural networks can be used to construct the conditional distributions in the following way. The distributional parameters (e.g. mean and variance) of the emission model $p_{\theta}(\mathbf{x}_t|\mathbf{z}_t)$ can be defined by a neural network transformation of \mathbf{z}_t , similar to deep generative models for images. The prior dynamic model $p_{\theta}(\mathbf{z}_t|\mathbf{z}_{<t})$ can be defined as (e.g. with an LSTM)

$$p_{\theta}(\mathbf{z}_t|\mathbf{z}_{<t}) = p_{\theta}(\mathbf{z}_t|\mathbf{h}_t^p, \mathbf{c}_t^p), \quad \mathbf{h}_t^p, \mathbf{c}_t^p = LSTM_{\theta}(\mathbf{z}_{<t}). \quad (40)$$

This means the previous latent states $\mathbf{z}_{<t}$ are summarised by the LSTM internal recurrent states \mathbf{h}_t^p and \mathbf{c}_t^p , which are then transformed into the distributional parameters of $p_{\theta}(\mathbf{z}_t|\mathbf{z}_{<t})$. For the factorised encoder distribution, it can also be defined using an LSTM:

$$q_{\phi}(\mathbf{z}_t|\mathbf{x}_{\leq t}) = q_{\phi}(\mathbf{z}_t|\mathbf{h}_t^q, \mathbf{c}_t^q), \quad \mathbf{h}_t^q, \mathbf{c}_t^q = LSTM_{\phi}(\mathbf{x}_{\leq t}). \quad (41)$$

4 Attention & Transformers

4.1 Motivation

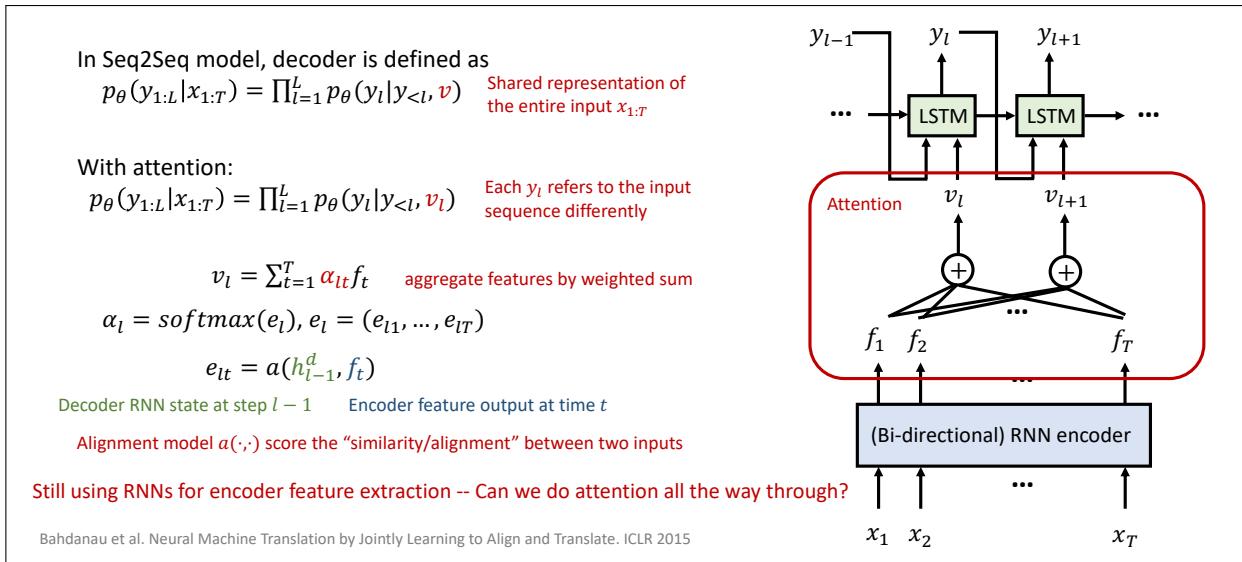
In sequence-to-sequence models, the goal is to learn the following model

$$p_{\theta}(\mathbf{y}_{1:L} | \mathbf{x}_{1:T}) = \prod_{l=1}^L p_{\theta}(\mathbf{y}_l | \mathbf{y}_{<l}, \mathbf{v}), \quad \mathbf{v} = enc(\mathbf{x}_{1:T}), \quad (1)$$

with $p_{\theta}(\mathbf{y}_l | \mathbf{y}_{<l}, \mathbf{v})$ and $enc(\mathbf{x}_{1:T})$ defined by LSTMs. Although LSTMs suffer less from the gradient vanishing/explosion problems, it can still be challenging for LSTMs to learn long-term dependencies. A practical trick to boost the performance is to reverse the order of the input, i.e. using $\mathbf{x}_{T:1}$ instead of $\mathbf{x}_{1:T}$ as the input to the encoder LSTM [Sutskever et al., 2014]. Still the difficulty of learning long-term dependencies remains unsolved even with this trick. Furthermore, it is possible that different \mathbf{y}_l words require different information extracted from $\mathbf{x}_{1:T}$, so a shared global representation \mathbf{v} for the input sequence might be sub-optimal.

People have found that as the sentence length has increased, the quality of the sentence decreases. This is a problem, since a lot of sequenced data is very long naturally.

4.2 NMT model



1. The first attempt is to improve the sequence-to-sequence model with attention. In Seq2Seq the global representation vector \mathbf{v} is used to represent the entire input sequence for all \mathbf{y} . This means it needs to capture as much information as possible, yet, it comes from the last state in the LSTM — this means that there may be problems of learning long-term dependencies.
2. Bahdanau proposed to use different feature vectors at different decoding steps to represent the input sequence. The idea: decoding words at different states, the context can be different. Therefore, don't use only the last state - we ask the encoder to produce a feature output f at each time step, and compute a weighted sum of those features to obtain a feature representation of the input sequence.
3. α is different at each decoding step. This depends on the alignment/similarity between the feature f_t and the representation of the current decoding step.

Bahdanau et al. [2015] proposed an attention-based approach (the authors called it as “alignment”) to address the above issues. Recall that at time t the encoder LSTM updates its internal recurrent states \mathbf{c}_t^e and \mathbf{h}_t^e using the current input \mathbf{x}_t :

$$\mathbf{h}_t^e, \mathbf{c}_t^e = LSTM_{\theta}^{enc}(\mathbf{x}_t, \mathbf{h}_{t-1}^e, \mathbf{c}_{t-1}^e). \quad (2)$$

Similarly for the decoder LSTM, we need to maintain its internal recurrent states \mathbf{c}_l^d and \mathbf{h}_l^d :

$$p_{\theta}(\mathbf{y}_l | \mathbf{y}_{<l}, \mathbf{v}) = p_{\theta}(\mathbf{y}_l | \mathbf{h}_l^d, \mathbf{c}_l^d), \quad \mathbf{h}_l^d, \mathbf{c}_l^d = LSTM_{\theta}^{dec}(\mathbf{y}_{l-1}, \mathbf{h}_{l-1}^d, \mathbf{c}_{l-1}^d). \quad (3)$$

In the original sequence-to-sequence model, the global representation \mathbf{v} is obtained by transforming the last hidden state \mathbf{h}_T^e of the encoder LSTM, and it is used to initialise \mathbf{h}_0^d and \mathbf{c}_0^d . Instead, Bahdanau et al. [2015] proposes using different representations of the input sequence $\mathbf{x}_{1:T}$ at different steps for predicting \mathbf{y}_l , i.e.

$$\begin{aligned} p_{\theta}(\mathbf{y}_{1:L} | \mathbf{x}_{1:T}) &= \prod_{l=1}^L p_{\theta}(\mathbf{y}_l | \mathbf{y}_{<l}, \mathbf{v}_l), \\ p_{\theta}(\mathbf{y}_l | \mathbf{y}_{<l}, \mathbf{v}_l) &= p_{\theta}(\mathbf{y}_l | \mathbf{h}_l^d, \mathbf{c}_l^d), \quad \mathbf{h}_l^d, \mathbf{c}_l^d = LSTM_{\theta}^{dec}([\mathbf{y}_{l-1}, \mathbf{v}_l], \mathbf{h}_{l-1}^d, \mathbf{c}_{l-1}^d). \end{aligned} \quad (4)$$

where the representation \mathbf{v}_l is obtained as follows:

$$\mathbf{f}_t = T_{\theta}(\mathbf{h}_t^e), \quad (\text{feature output of the encoder at time } t) \quad (5)$$

$$e_{lt} = a(\mathbf{h}_{l-1}^d, \mathbf{f}_t), \quad (\text{compute similarity/alignment score}) \quad (6)$$

$$\alpha_l = softmax(e_{lt}), \quad \mathbf{e}_l = (e_{l1}, \dots, e_{lT}), \quad (7)$$

$$\mathbf{v}_l = \sum_{t=1}^T \alpha_{lt} \mathbf{f}_t. \quad (\text{weighted aggregation of input features}) \quad (8)$$

The key idea of using \mathbf{v}_l as a weighted average of individual features \mathbf{f}_t is to allow the decoder LSTM to directly access the representation for each input \mathbf{x}_t , therefore the issue of lacking long-term dependencies between \mathbf{y}_l and \mathbf{x}_t is addressed. This is in contrast with the global representation vector \mathbf{v} in the original Seq2Seq model: since \mathbf{v} is computed using the last recurrent state of the encoder LSTM, it is questionable whether \mathbf{v} can capture long-term dependencies within $\mathbf{x}_{1:T}$. Another notable difference is that in Bahdanau et al. [2015] the features \mathbf{v}_l are used as the input to the decoder LSTM (together with \mathbf{y}_{l-1}), while in the original Seq2Seq model the global representation \mathbf{v} is used to initialise the decoder LSTM’s recurrent states.

The alignment score $e_{lt} = a(\mathbf{h}_{l-1}^d, \mathbf{f}_t)$ is computed between \mathbf{h}_{l-1}^d (which summarises $\mathbf{y}_{<l}$, required for the auto-regressive model to predict \mathbf{y}_l) and \mathbf{f}_t (a feature representation for \mathbf{x}_t but is also dependant on $\mathbf{x}_{<t}$). These scores are then passed through a softmax layer to obtain the *attention weight* α_{lt} , with larger attention weight value α_{lt} the final representation \mathbf{v}_l will focus more on the input feature \mathbf{f}_t for \mathbf{x}_t .

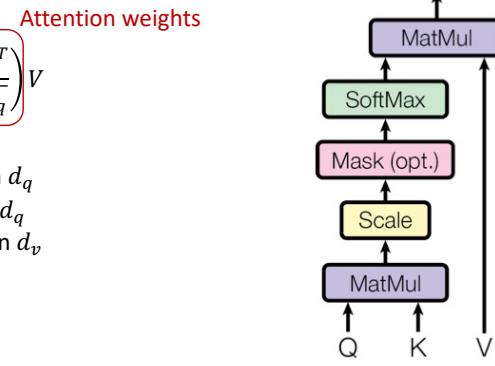
4.3 Attention in Transformers — Single-head attention

- Single head attention

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

$Q \in R^{N \times d_q}$: N query inputs, each of dimension d_q
 $K \in R^{M \times d_q}$: M key vectors, each of dimension d_q
 $V \in R^{M \times d_v}$: M value vectors, each of dimension d_v
 $a(\cdot)$: activation function applied row-wise

Self attention: $K = Q$



The *scaled dot product attention* method is introduced by [Vaswani et al. \[2017\]](#). Intuitively it can be understood from an information retrieval point of view. Queries are submitted to the system which are represented by the *query vectors* $\mathbf{q}_i \in \mathbb{R}^{d_q}$, and the system will first check the matching/alignment/“similarity” between the query and the keys (which are represented by *key vectors* $\mathbf{k}_j \in \mathbb{R}^{d_q}$), then return the retrieved values to the user. Each key vector \mathbf{k}_j is associated with a *value vector* $\mathbf{v}_j \in \mathbb{R}^{d_v}$, so that if \mathbf{q}_i and \mathbf{k}_j are aligned, the value vector \mathbf{v}_j will be returned in some form.

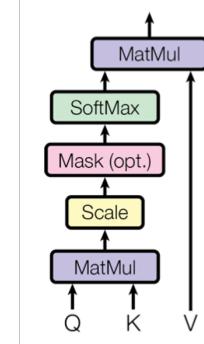
In detail, the mathematical form for single-head attention is the following:

$$\text{Attention}(Q, K, V; a) = a \left(\frac{QK^T}{\sqrt{d_q}} \right) V \quad (9)$$

$$Q = (\mathbf{q}_1, \dots, \mathbf{q}_N)^\top \in \mathbb{R}^{N \times d_q} \quad (10)$$

$$K = (\mathbf{k}_1, \dots, \mathbf{k}_M)^\top \in \mathbb{R}^{M \times d_q} \quad (11)$$

$$V = (\mathbf{v}_1, \dots, \mathbf{v}_M)^\top \in \mathbb{R}^{M \times d_v} \quad (12)$$



There are two important ingredients in the scaled dot product attention process. First, an *attention matrix* $A = a \left(\frac{QK^T}{\sqrt{d_q}} \right)$ is computed to indicate the alignment of each query vector \mathbf{q}_i to the key vectors \mathbf{k}_j . Then given the attention matrix A , the attention output for each query vector \mathbf{q}_i is the weighted sum of the value vectors $\sum_{j=1}^M A_{ij} \mathbf{v}_j$, which is similar to the final output of the attention method by Bahdanau et al. Here $a(\cdot)$ is an activation function applied row-wise, and in *soft attention*, $a(\cdot)$ is the softmax function, i.e. $A_{ij} = \text{softmax}(\langle \mathbf{q}_i, \mathbf{k}_1 \rangle, \dots, \langle \mathbf{q}_i, \mathbf{k}_M \rangle) / \sqrt{d_q}$. This is in contrast with *hard attention* where $a(\cdot)$ returns a one-hot vector for each row with the $j^* = \arg \max_j \langle \mathbf{q}_i, \mathbf{k}_j \rangle$ element equals to 1. Hard attention can be interpreted exactly as an information retrieval system since it returns the corresponding value for the best key match to the query.

When the dimensionality of the query/key vector d_q is large, the dot product $\langle \mathbf{q}_i, \mathbf{k}_j \rangle$ can be large as well. This is likely to increase the gap between the dot product values so that the softmax output could be dominated by a single entry (i.e. close to hard attention). Normalisation of the logits $\langle \mathbf{q}_i, \mathbf{k}_j \rangle$ could be useful to address this issue. For the specific choice of $\sqrt{d_q}$, this comes from the assumption that the elements in \mathbf{q}_i and \mathbf{k}_j are independently distributed with variance 1. If so then the variance of $\langle \mathbf{q}_i, \mathbf{k}_j \rangle$ is d_q , so by normalising the dot product with $\sqrt{d_q}$, the logit will have its variance equal to 1.

In some cases, *masking* is applied to the attention procedure. For example, a typical masking strategy will define a mask matrix $M \in \{0, 1\}^{N \times M}$, so that if $M_{ij} = 0$, then the corresponding attention weight $A_{ij} = 0$, so that \mathbf{v}_j does not contribute to the final output for query \mathbf{q}_i .

Self-attention is also in wide usage which sets $K = Q$.

4.3.1 Hard-attention

- Single head attention

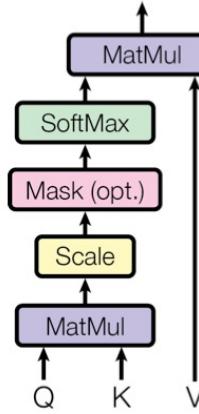
$$\text{Attention}(Q, K, V; a) = a \left(\frac{QK^T}{\sqrt{d_q}} \right) V$$

(QK^T)_{ij} = ⟨q_i, k_{ji} and key k_j

Retrieve values according to the weighting

Hard attention: $a(x) = \text{onehot}(\text{argmax } x_i)$ for $x = (x_1, \dots, x_d)$

- Each key vector k_i is associated with a value vector v_i ($K = (k_1, \dots, k_M)^T, V = (v_1, \dots, v_M)^T$)
- The query matrix packs the query vectors $Q = (q_1, \dots, q_N)^T$
- $a(\cdot)$ applied row-wise: the n^{th} row of the output is $\text{onehot}(\text{argmax } \langle q_n, k_i \rangle)$
- Therefore, the n^{th} row of the attention output will be v_{i_n} for $i_n = \text{argmax } \langle q_n, k_i \rangle$



4.3.2 Soft-attention

- Single head attention

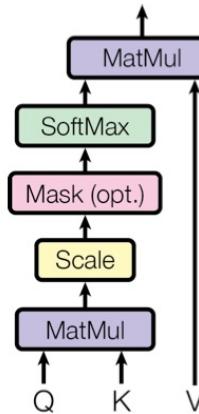
$$\text{Attention}(Q, K, V; a) = a \left(\frac{QK^T}{\sqrt{d_q}} \right) V$$

(QK^T)_{ij} = ⟨q_i, k_{ji} and key k_j

Retrieve values according to the weighting

Soft attention: $a(x) = \text{softmax}(x)$ for $x = (x_1, \dots, x_d)$

- Each key vector k_i is associated with a value vector v_i ($K = (k_1, \dots, k_M)^T, V = (v_1, \dots, v_M)^T$)
- The query matrix packs the query vectors $Q = (q_1, \dots, q_N)^T$
- $a(\cdot)$ applied row-wise: the n^{th} row of the output is $\text{softmax}(q_n K^T)$
- Therefore, the n^{th} row of the attention output will be a weighted sum of value vectors v_j with weighting proportional to $\exp(\langle q_n, k_j \rangle)$



4.3.3 Masked-attention

- Single head attention

$$\text{MaskedAttention}(Q, K, V; a, M) = a \left(\text{mask} \left(\frac{QK^T}{\sqrt{d_q}}, M \right) \right) V$$

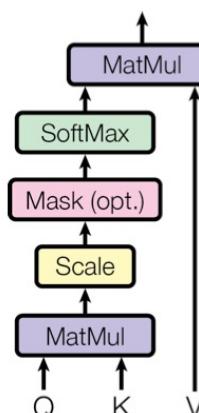
(QK^T)_{ij} = ⟨q_i, k_{ji} and key k_j

Retrieve values according to the weighting and masking

Masked attention: mask out some of the attention values,

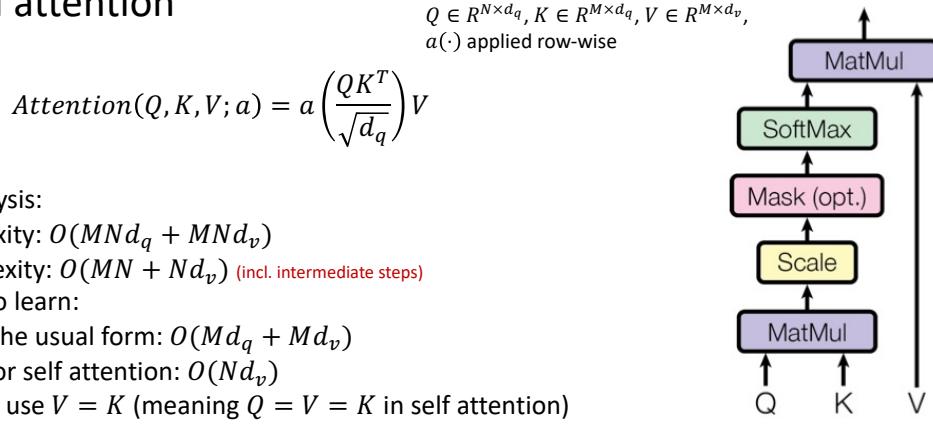
Example when $a(\cdot)$ is softmax:

- M_{nm} takes values 0 (mask out) or 1 (keep in)
- With $M_{nm} = 0$, set $(QK^T)_{nm} = -\infty$ so that $\exp \left(\frac{QK^T}{\sqrt{d_q}} \right)_{nm} = 0$
- So value v_m will NOT contribute to the attention output for query q_n
- Useful for sequence prediction with a given ordering: in test time, “future” is not available for the “current” to attend



4.3.4 Complexity analysis

- Single head attention



The time complexity for scaled dot-product attention is $\mathcal{O}(MNd_q + MNd_v)$. These include the dot product QK^\top which has $\mathcal{O}(MNd_q)$ run-time cost, the computation of A matrix given the dot product which has $\mathcal{O}(MN)$ cost, and the dot product AV which has $\mathcal{O}(MNd_v)$ cost.

The space complexity for the scaled dot-product is $\mathcal{O}(MN + Nd_v)$ since back-propagation requires storing some intermediate results. These include the dot product QK^\top which has $\mathcal{O}(MN)$ memory cost and the final output AV which has $\mathcal{O}(Nd_v)$ cost.

For self-attention, since $Q = K$, then the time complexity and space complexity figures are $\mathcal{O}(N^2d_q + N^2d_v)$ and $\mathcal{O}(N^2 + Nd_v)$, respectively.

4.4 Attention in Transformers — Multi-head attention

- Multi-head attention

$$\text{Multihead}(Q, K, V, a) = \text{concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

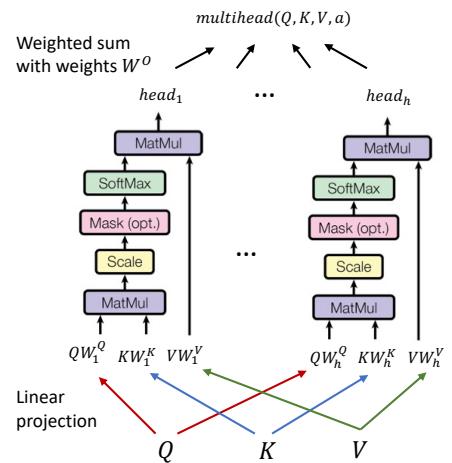
$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V, a)$$

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

Different head represent different alignments, e.g.

when each query represents a word and self-attention is used:

- Head 1: find keys that are semantically similar to q
- Head 2: find keys that makes (q, k) as a subject-verb pair
- ...



- Multi-head attention

$$\text{Multihead}(Q, K, V, a) = \text{concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

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

Complexity analysis (assume $Q \in R^{N \times d_q}$ projected to $R^{N \times \tilde{d}_q}$ and so on):

- Time complexity:

$$O(hMN(\tilde{d}_q + \tilde{d}_v) + h(\tilde{d}_q d_q(M + N) + \tilde{d}_v d_v M) + Nh\tilde{d}_v d_{out})$$

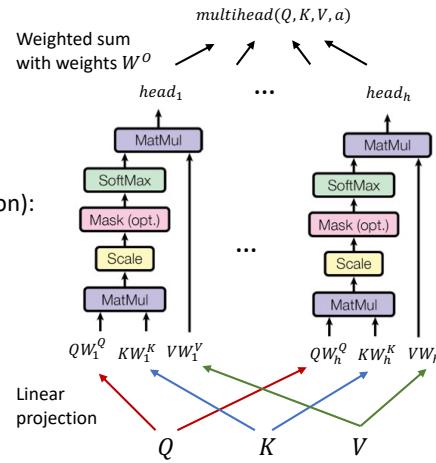
- Space complexity:

$$O(hN(M + \tilde{d}_v) + h((N + M)\tilde{d}_q + M\tilde{d}_v) + Nd_{out})$$

- Parameters to learn (apart from K and V):

- Projection parameters W_i^Q, W_i^K, W_i^V

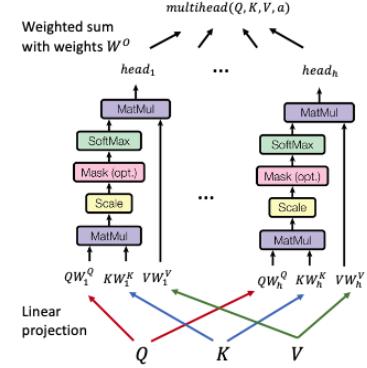
- Output weight matrix W^O



Multi-head attention repeats the single-head attention process with different “views”. Intuitively, consider information retrieval with the input query representing “cat”. Then depending on the definition of similarity, it can match to key “tiger” or key “pet” and so on.

Multi-head attention allows the definition of multiple alignment processes, by projecting the inputs into different sub-spaces then performing dot product attention in such sub-spaces. The outputs of each attention head are concatenated and projected to produce the final output.

$$\begin{aligned} \text{Multihead}(Q, K, V; a) &= \text{concat}(\text{head}_1, \dots, \text{head}_h)W^O, \\ \text{head}_i(Q, K, V; a) &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V; a). \end{aligned} \quad (13)$$



It is clear that the time and space complexity figures of multi-head attention are h times of those for a single head plus the extra costs for linear projections. Assume the projection matrices have sizes $W_i^Q \in \mathbb{R}^{d_q \times \tilde{d}_q}$, $W_i^K \in \mathbb{R}^{d_q \times \tilde{d}_q}$, $W_i^V \in \mathbb{R}^{d_v \times \tilde{d}_v}$ and $W^O \in \mathbb{R}^{h\tilde{d}_v \times d_{out}}$. This means

$$\text{time complexity: } \mathcal{O}\left(\underbrace{h(MN\tilde{d}_q + MN\tilde{d}_v)}_{\text{attention heads}} + \underbrace{h(\tilde{d}_q d_q(M + N) + \tilde{d}_v d_v M)}_{\text{input projections}} + \underbrace{Nh\tilde{d}_v d_{out}}_{\text{combined outputs}}\right),$$

$$\text{space complexity: } \mathcal{O}\left(\underbrace{h(MN + Nd_v)}_{\text{attention heads}} + \underbrace{h(\tilde{d}_q(M + N) + \tilde{d}_v M)}_{\text{input projections}} + \underbrace{Nd_{out}}_{\text{combined outputs}}\right).$$

To keep the costs close to performing single-head attention in the original space, often the \tilde{d}_q , \tilde{d}_v dimensions are set to be $\tilde{d}_q = \lfloor \frac{d_q}{h} \rfloor$ and $\tilde{d}_v = \lfloor \frac{d_v}{h} \rfloor$, respectively.

4.5 Transformer Architecture

Transformer (proposed in [Vaswani et al. \[2017\]](#)) is an encoder-decoder type of architecture, which is visualised in Figure 1. We discuss some of the key ingredients as follows.

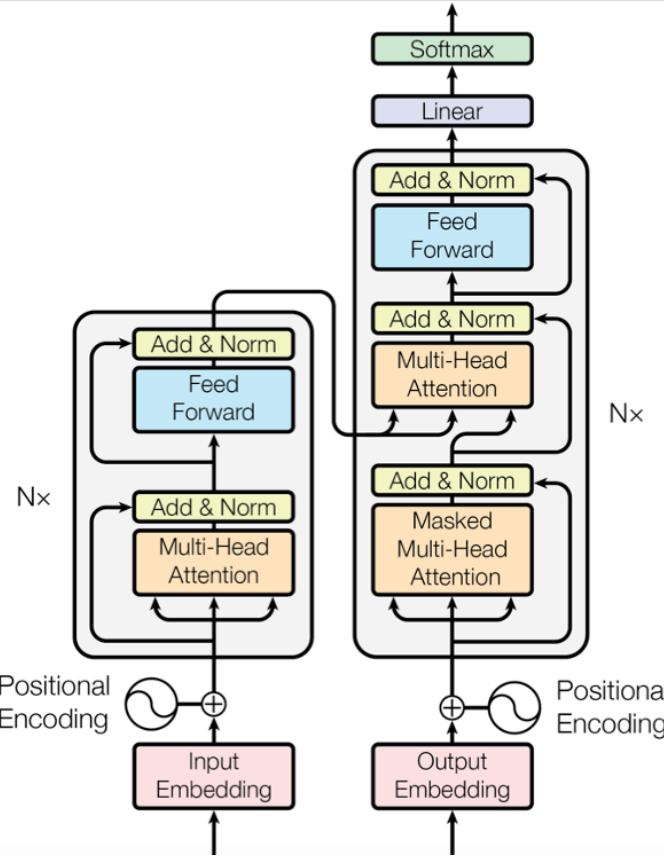
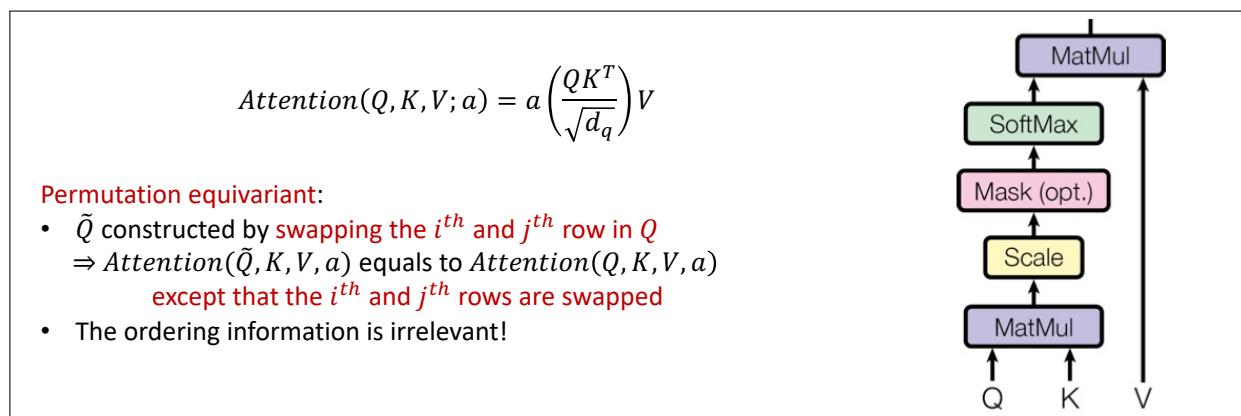


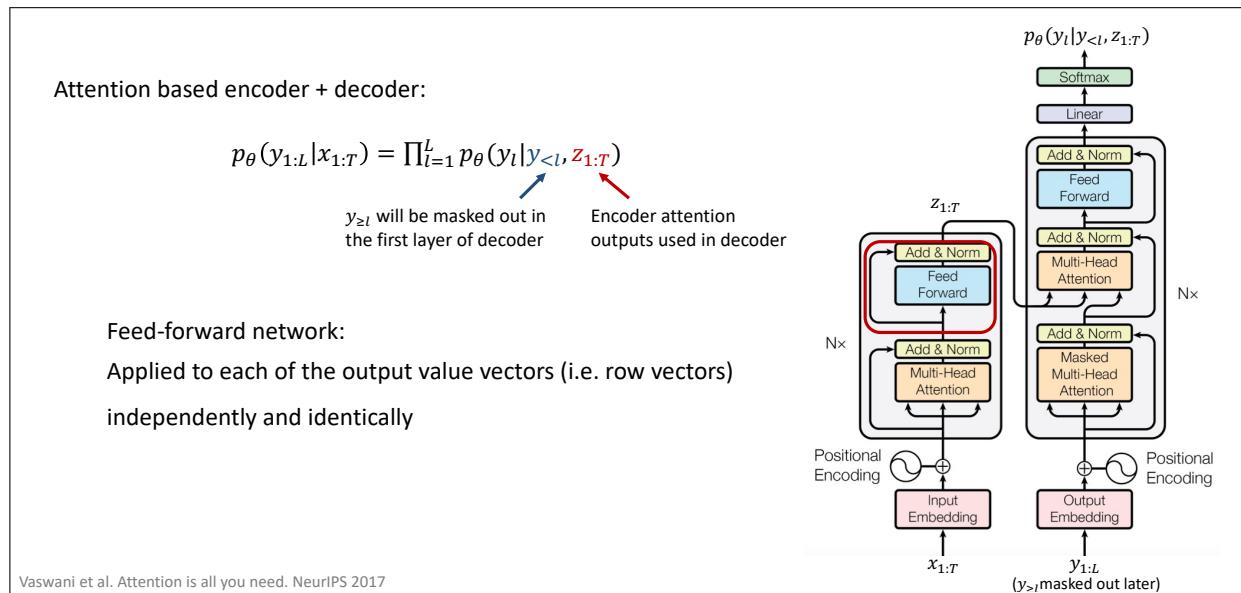
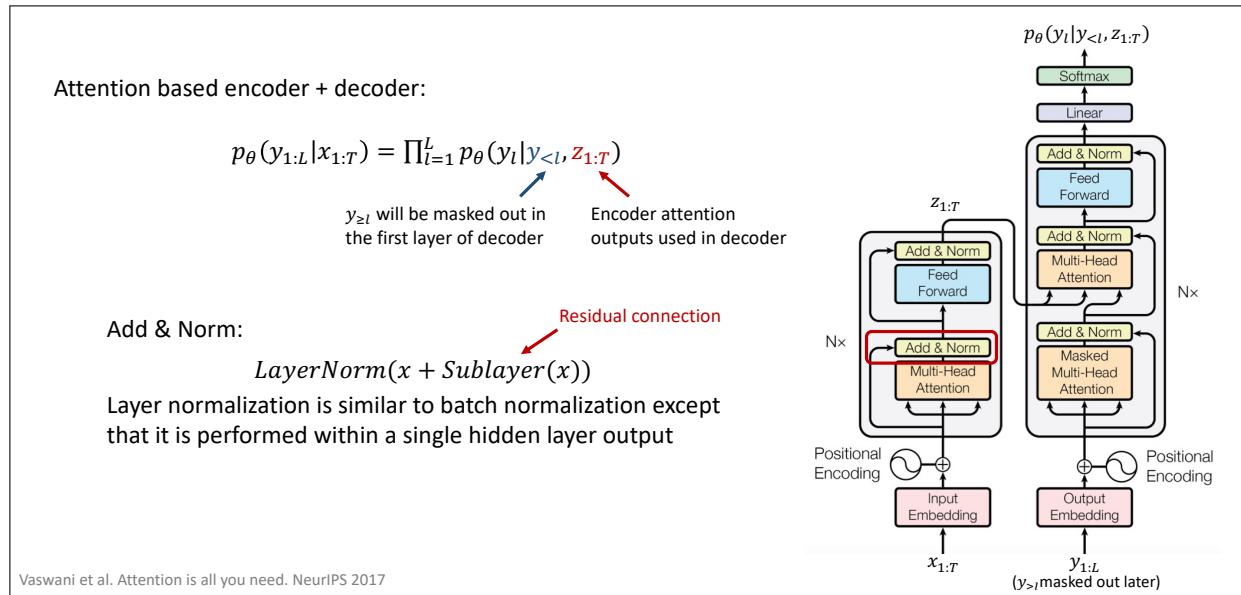
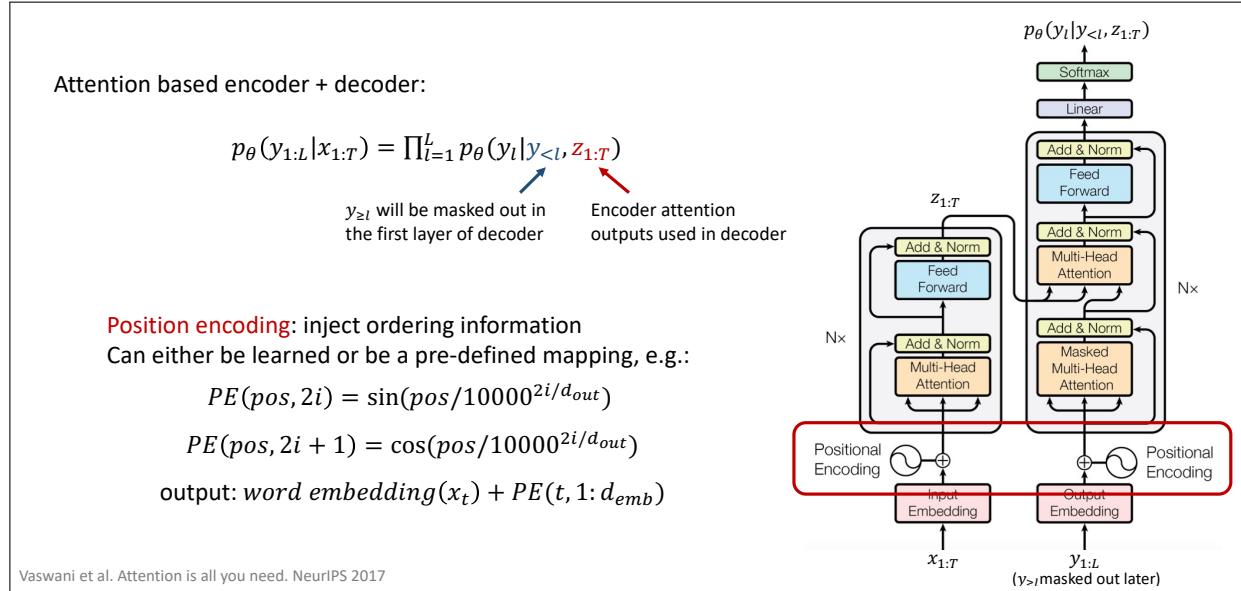
Figure 1: The Transformer architecture in [Vaswani et al. \[2017\]](#).

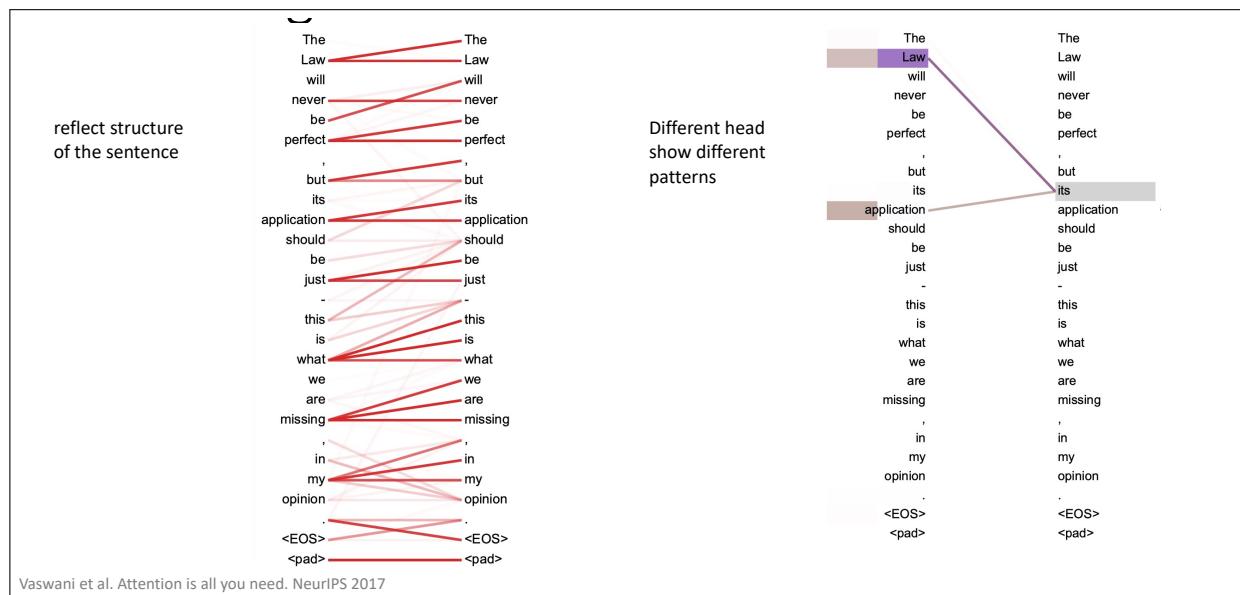
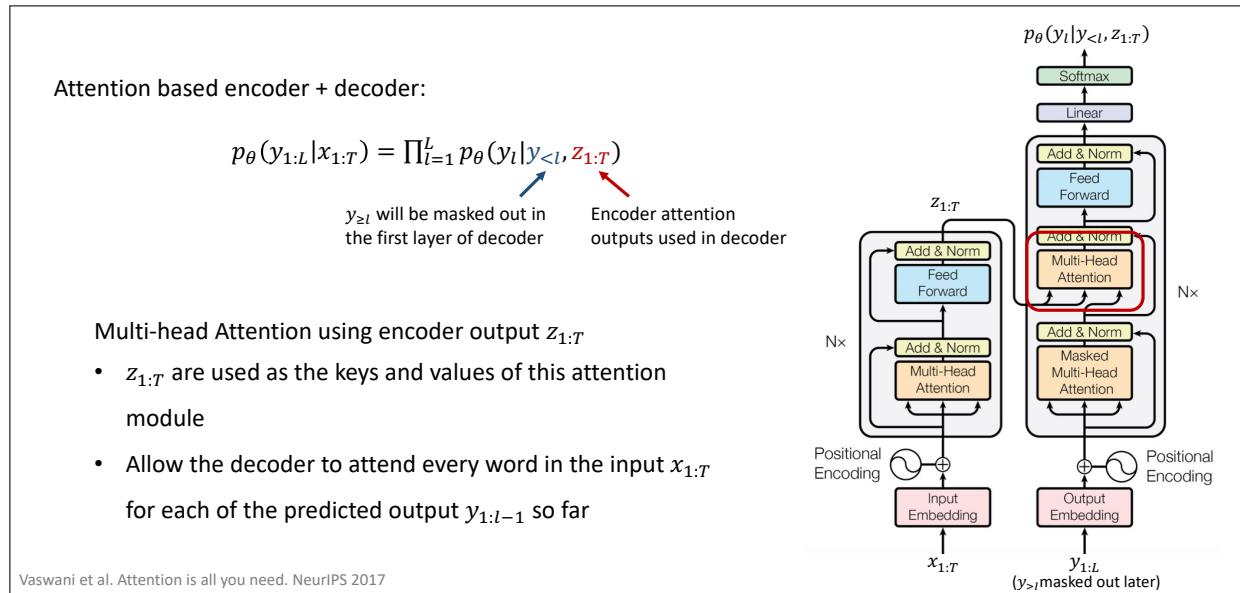
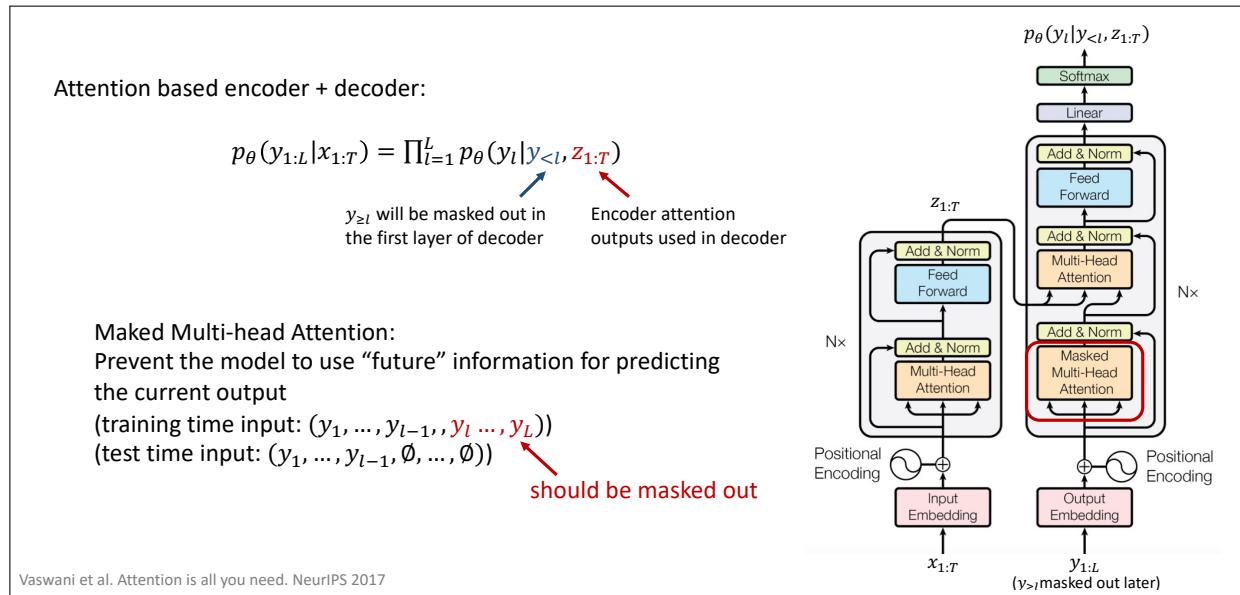
We first want to use the Attention based encoder to extract features for input sequences $x_{1:T}$. Then we will follow the autoregressive decoding process again to produce the output sequence, but different from the Sequence to sequnce model; now the decoder is also an attention based network,rather than LSTM.

$$p_\theta(y_{1:L}|x_{1:T}) = \prod_{l=1}^L p_\theta(y_l|y_{<l}, z_{1:T}) \quad (4.5.1)$$

The first step in the encoding process is to gether input embeddings for the input sequence. In an NLP setting, this is the word embeddings. BEfore we feed in the input representationso, we append the input embedding with a position encoding. **TODO**







5 Advances and Applications

5.1 Attention Applications

Attention used in RNN generators back in 2013:

- Idea: align two different sequences with different length
 - $x_{1:T}$: x_t is the pen position at time t
 - $y_{1:L}$: y_l is the l^{th} character in the text

RNN parameterised mixture of Gaussian attention

$$\text{Attention}(x_t, y_{1:L}) = \sum_{l=1}^L \phi(t, l) \text{emb}(y_l)$$

$$\phi(t, l) = \sum_{k=1}^K \alpha_t^k \exp(-\beta_t^k(l - \kappa_t^k)^2)$$

$$\alpha_t^k, \beta_t^k, \kappa_t^k = \text{MLP}(h_t), h_{1:T} = \text{RNN}(x_{1:T})$$

Graves. Generating Sequences with Recurrent Neural Networks. arXiv:1308.0850

pen position and pen position wasn't aligned — author developed an RNN

DRAW model, extending Graves (2013):

- A VAE generative model for images combining RNNs and Gaussian attention
- Treat the intermediate “drawing state” as latent variable:
 $p(x|z) = p(x|z_{1:T})$
- Decoder uses “read” and “write” operations guided by attention
- Gaussian attention to select focused patch at time t
- The Gaussian attention filter parameters are obtained from the RNN in the decoder

Figure 1



Figure 2: before we use CNNs to extract features from images, and then use RNNs to generate captions. Now, since the CNN captures spatial location we can treat these features as an input sequence indexed by the spatial location. Then apply Seq2Seq.

5.2 Transformer in NLP

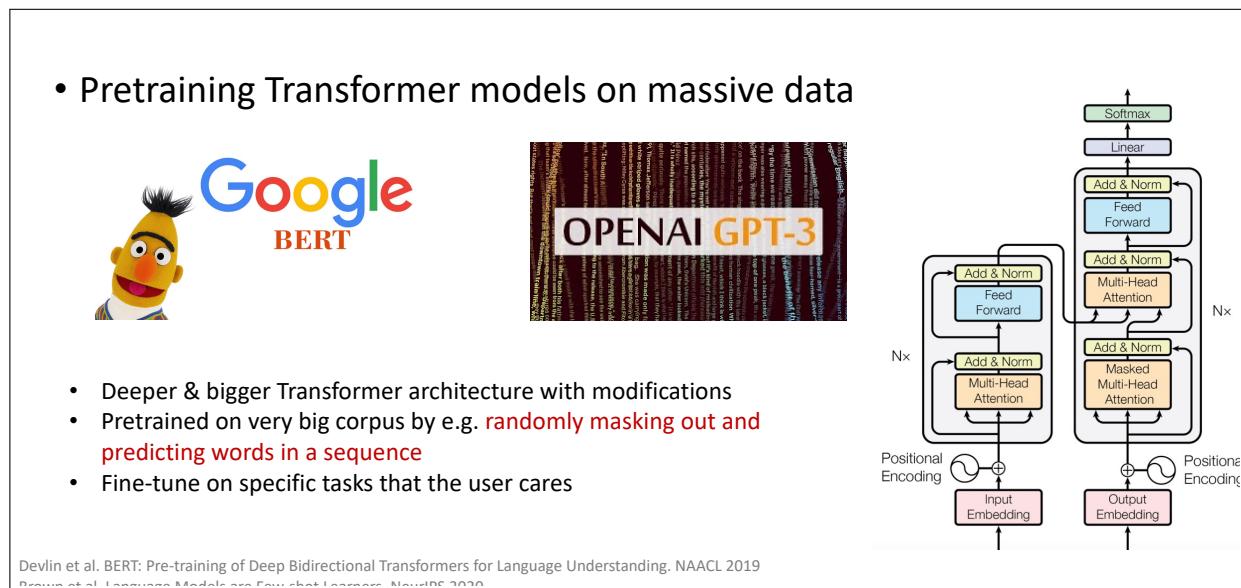


Figure 3: Attention based. We mask and pre-train.

The screenshot shows a web-based interface for generating text from unstructured data. On the left, there is a text input field labeled "Parse unstructured data" containing the following code:

```
Please make a table summarizing the fruits from Goocrux
| Fruit | Color | Flavor |
| Neoskizzles | Purple | Sweet |
| Loheckles | Grayish blue | Tart |
|"""

response = openai.Completion.create(model="davinci",
prompt=prompt, stop="\n\n", temperature=0,
max_tokens=300)

print(response)
```

Below the code is a button labeled "See cached response". To the right, the generated response is displayed in a dark box:

Fruit Color Flavor
Neoskizzles Purple Sweet
Loheckles Grayish blue Tart
Pounits Bright green Savory
Loopnovas Neon pink Cotton candy
Glowlis Pale orange Sour

At the bottom left of the interface, it says "beta.openai.com".

Figure 4

5.3 Multi-head Attention in other applications

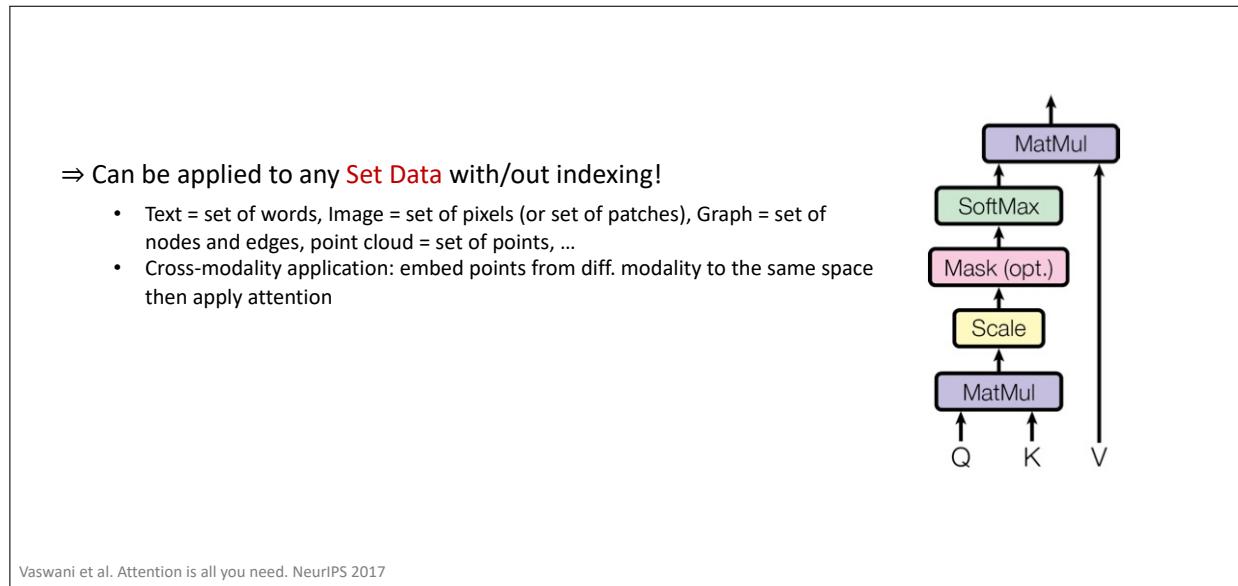


Figure 5: Trnasformers can be applied to any domain where data comes in sets

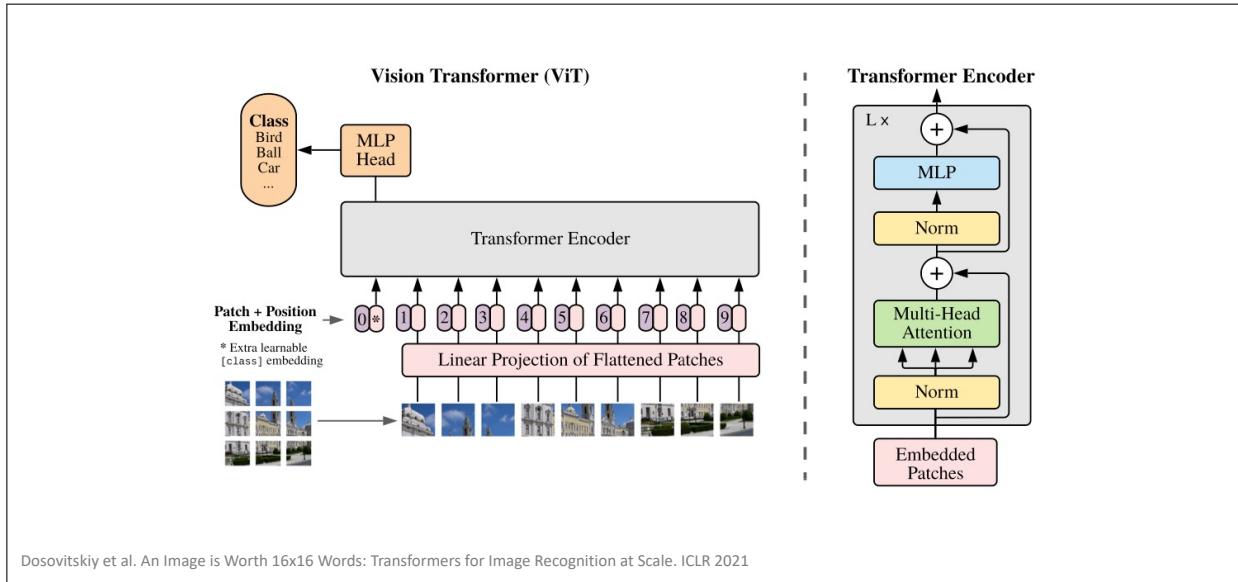
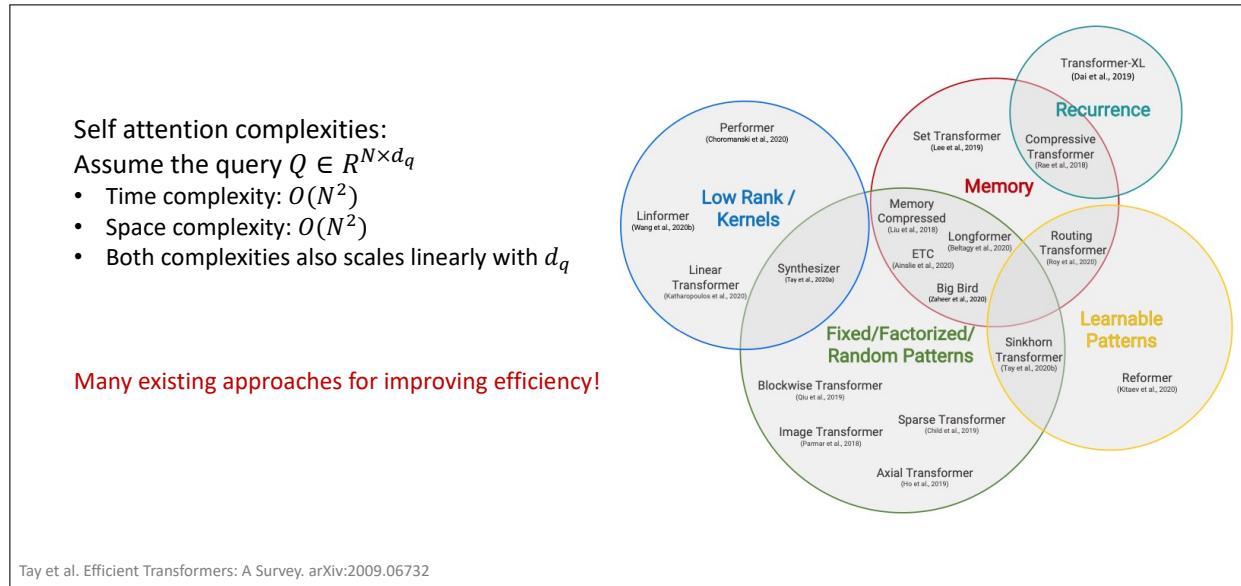


Figure 6: Here, we can view an image as spatial locations. This has achieved state of the art results in image classification. Convolutions are still more efficient pixelwise than just using attention on images.



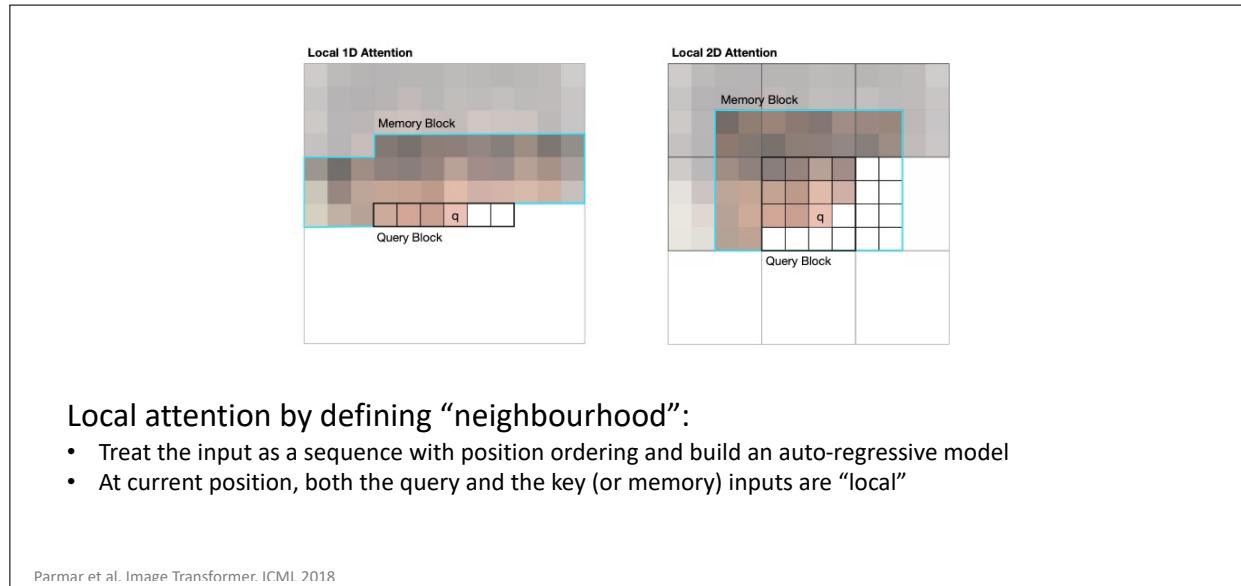
Figure 7

5.4 Efficient Transformers



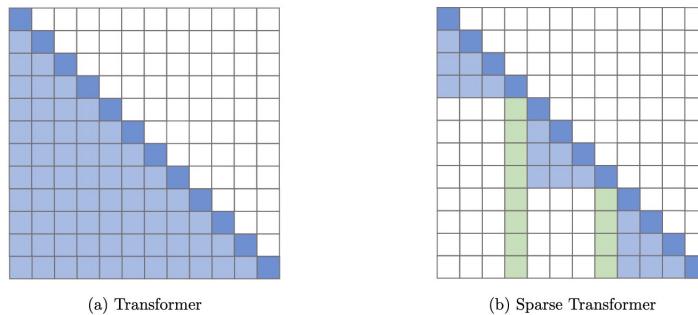
Tay et al. Efficient Transformers: A Survey. arXiv:2009.06732

Figure 8



Parmar et al. Image Transformer. ICML 2018

Figure 9: attention is only used in the neighbourhood of the focal pixel

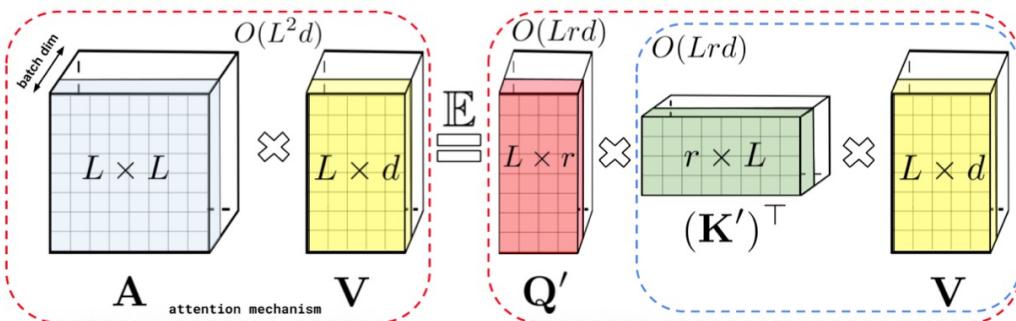


Sparse attention by defining attention weight matrix patterns:

- Full attention matrix before non-linearity: $(QK^T)_{ij} = \langle q_i, k_j \rangle$
 - Sparse attention matrix: fix rules such that only for certain pairs of (i, j) the attention entry is computed
 - Different attention heads can have different sparsity patterns

Child et al. Generating Long Sequences with Sparse Transformers. arXiv:1904.10509

Figure 10: sparseness



Low-rank approximations:

- If the activation $a(\cdot)$ in $\text{Attention}(Q, K, V, a)$ is linear: $(QK^T)V = Q(K^TV)$
 - In general the attention matrix has entries $A_{ij} = K(q_i, k_j)$ with $K(\cdot, \cdot)$ a kernel function
 - Random feature approximation can be done to (approximately) compute A fastly

Choromanski et al. Rethinking attention with Performers. ICLR 2021

Figure 11: Low rank approximation, if A is linear then we can remove quadratic scaling and make it linear.

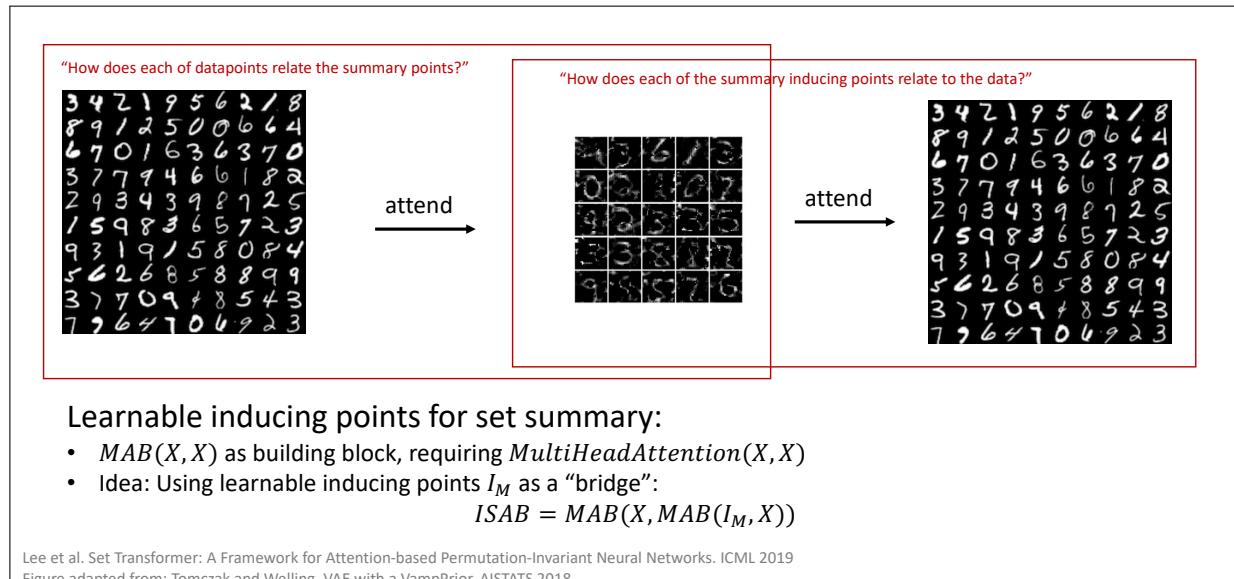


Figure 12

5.5 Memorisation in Transformers

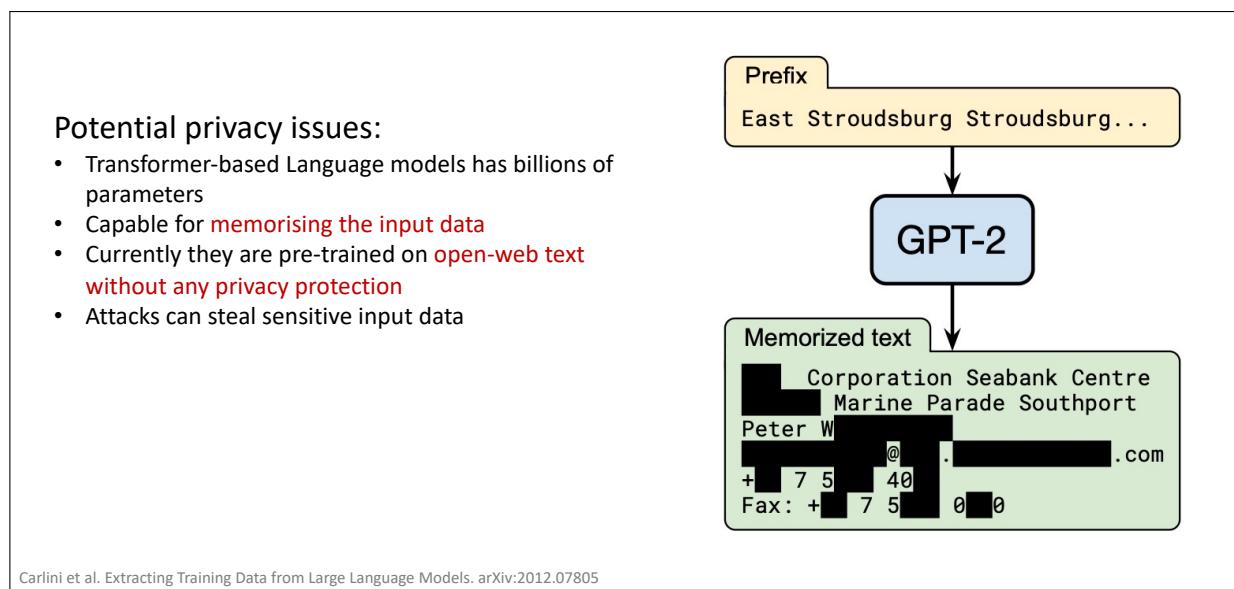


Figure 13

A RNN Notes

A.1 Tricks to fix the gradient vanishing/explosion issue

There are a handful of empirical tricks to fix the gradient vanishing/explosion issues discussed above.

- Gradient clipping:

This trick is often used to prevent the gradient from explosion. With a fixed hyper-parameter γ , a gradient \mathbf{g} is clipped when $\|\mathbf{g}\| > \eta$:

$$\mathbf{g} \leftarrow \frac{\gamma}{\|\mathbf{g}\|} \mathbf{g}.$$

This trick ensures the gradients used in optimisation has their maximum norm bounded by a pre-defined hyper-parameter. It introduces biases in the gradient-based optimisation procedure, but in certain cases it can be beneficial. Figure 2 visualises such an example, where with gradient clipping, the updates can stay in the valley of the loss function.

- Good initialisation of the recurrent weight matrix W_h :

The IRNN approach [Le et al., 2015] uses ReLU activation's for ϕ_h and initialise $W_h = \mathbf{I}$, $\mathbf{b}_h = \mathbf{0}$. This makes $\phi'_h(t) = \delta(t > 0)$ and $\frac{d\mathbf{h}_{l+1}}{d\mathbf{h}_l} = \delta(W_x \mathbf{x}_{l+1} > 0)$ at initialisation. While there is no guarantee of eliminating the gradient vanishing/explosion problem during the whole course of training, empirically RNNs with this trick have achieve competitive performance to LSTMs in a variety of tasks.

- Alternatively, one can construct the recurrent weight matrix W_h to be orthogonal or unitary matrix. See e.g. Saxe et al. [2014]; Arjovsky et al. [2016] for examples.

B Transformer Notes

B.1 Position encoding

From the equations of scaled dot-product attention, we see that attention is equivariant to row permutations in the query matrix Q . To see this, notice that permuting rows in a matrix is equivalent to left multiplying a permutation matrix P to Q . Since the non-linearity $a(\cdot)$ is applied row-wise, this leads to the permutation equivariance result:

$$\text{Attention}(PQ, K, V; a) = a \left(\frac{PQK^\top}{\sqrt{d_q}} \right) V = Pa \left(\frac{QK^\top}{\sqrt{d_q}} \right) V = P \text{Attention}(Q, K, V; a). \quad (14)$$

Therefore, for an input sequence of queries $\mathbf{q}_1, \dots, \mathbf{q}_N$ in which the ordering information (i.e. the subscript n) is useful for the task (e.g. time series regression), these ordering information needs to

be added in some form to the attention inputs. This issue is addressed by position encoding, which constructs the query input \tilde{Q} to the attention module as:

$$\tilde{Q} = (\tilde{q}_1, \dots, \tilde{q}_N)^\top, \quad \tilde{q}_n = f(\mathbf{q}_n, PE(n)), \quad (15)$$

where the f transformation is often set to be simple operations such as summation and concatenation, etc. $PE(n)$ is called the *position encoding* of input index n , which can either be learned (e.g. having a set of learnable parameters $\{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ or using an NN-parameterised function), or computed using a pre-defined function. Learned embedding of the form $\{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ is well suited if the maximum value of the index is known. Otherwise, in the case where unseen index can occur in test time, a pre-defined function might be preferred.

A popular choice of such pre-defined functions is the *sinusoid embedding* [Vaswani et al., 2017]:

$$\begin{aligned} PE(n) &= (PE(n, 0), \dots, PE(n, 2I)), \\ PE(n, 2i) &= \sin(n/10000^{2i/d_q}), \\ PE(n, 2i + 1) &= \cos(n/10000^{2i/d_q}). \end{aligned} \quad (16)$$

We see that sinusoid embeddings use multiple periodic functions to embed the input index n , where the frequency of each sine/cosine wave is determined by i . The authors of Vaswani et al. [2017] did not give a formal justification of this approach and instead stated that “we hypothesized it would allow the model to easily learn to attend by relative positions”. My personal hypothesis is that, in the case of using concatenation to construct the new query (i.e. $\tilde{q}_n = f(\mathbf{q}_n, PE(n)) = [\mathbf{q}_n, PE(n)]$) which is then multiplied by learnable weight matrices, the sinusoid embedding allows the network to learn a very flexible position encoding function as a combination of many sine/cosine waves of different frequencies (think about the Fourier series of a continuous function).

B.2 Layer normalization

Layer normalisation [Ba et al., 2016] is a normalisation technique used to stabilise neural network training. Assuming a feed-forward layer with the pre-activation computed as $\mathbf{a} = (a_1, \dots, a_H) = W\mathbf{x} + \mathbf{b}$, layer normalisation can be applied to the pre-activation vector before feeding to the non-

linearity:

$$\mathbf{a} \leftarrow \frac{\mathbf{a} - \mu}{\sigma}, \quad \mu = \frac{1}{H} \sum_{h=1}^H a_h, \quad \sigma = \sqrt{\frac{1}{H} \sum_{h=1}^H (a_h - \mu)^2}. \quad (17)$$

In transformers, layer normalisation is applied together with a residual link: $Add\&Norm(x) = LayerNorm(x + Sublayer(x))$, where $Sublayer(\cdot)$ can either be a multi-head attention block or a point-wise feed-forward network.

B.3 Point-wise feed-forward network

The point-wise feed-forward network is used as the “feed forward” layer in Transformer’s architecture (see Figure 1). The multi-head attention (after “Add&Norm”) returns a matrix of size $N \times d_{out}$, representing the attention results for N queries. These output can be processed “point-wise” (i.e. row-wise) with a feed-forward neural network, which effectively treat the rows in the attention outputs as “datapoint” inputs for the next layer.