Deep Learning

Modelling Sequential Data

Sam Bond-Taylor
Department of Computer Science

Lecture Overview

Durham University

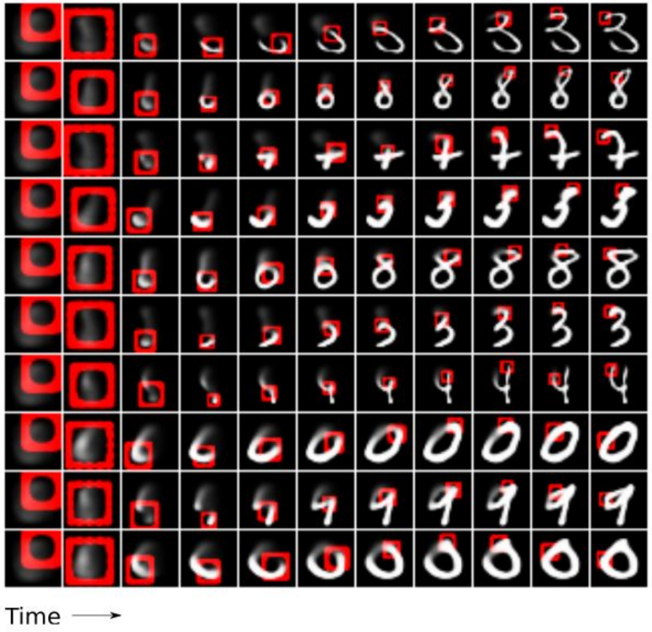
Recap

- FC, CNN, ResNet

Today's Lecture

- Recurrent Neural Networks
- Backpropagation Through Time
- Vanishing Gradients and LSTMs
- Neural Attention
- Transformers

RNN Transformer



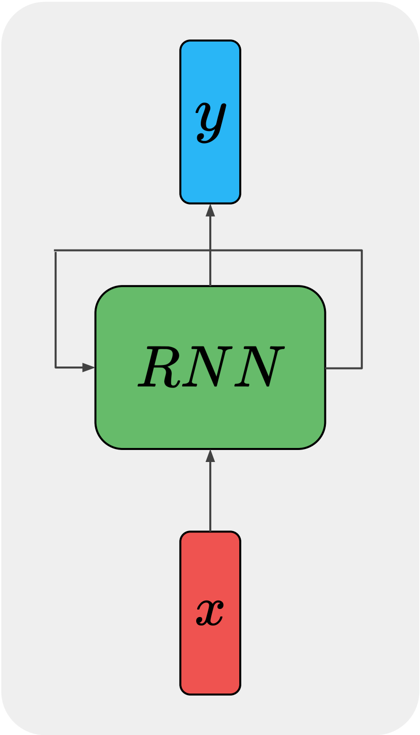
Sequence Modelling: Design Criteria

To model sequences we need to:

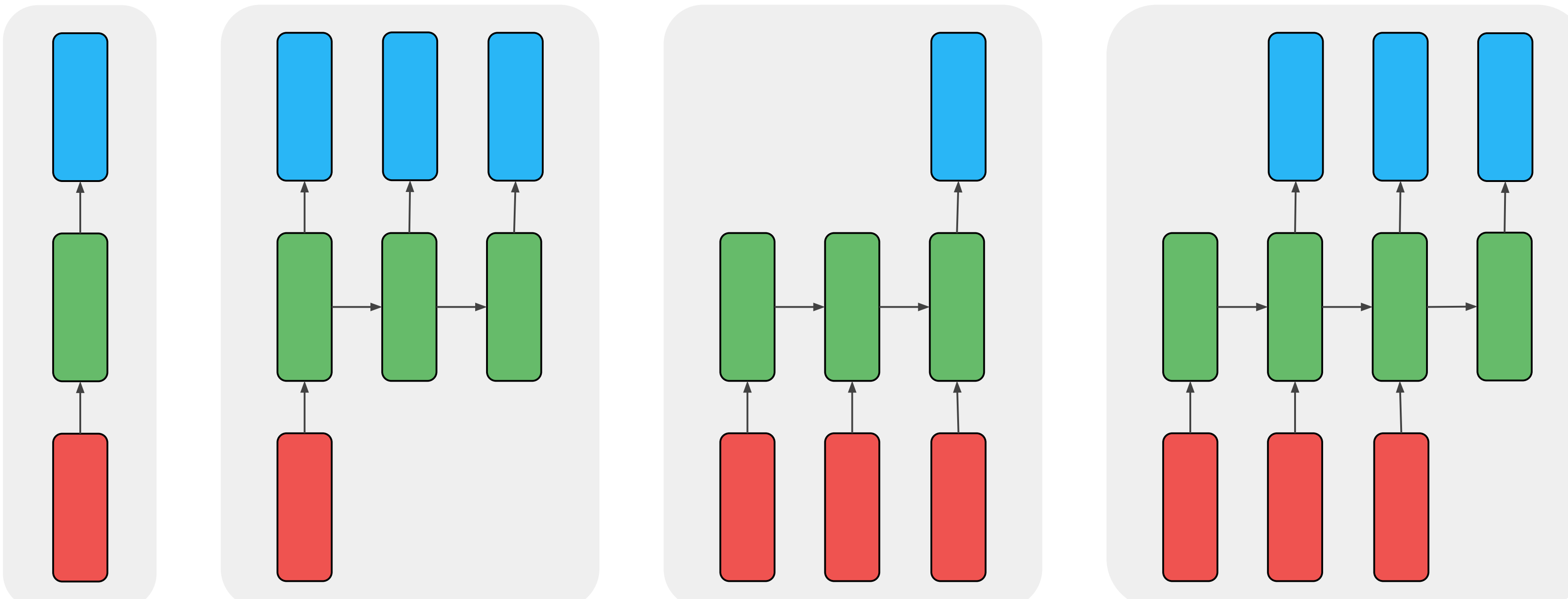
1. Handle **variable-length** sequences
2. Track **long-term** dependencies
3. Maintain information about **order**
4. **Share parameters** across the sequence

RNNs are Turing Complete!

$$h_{t+1} = f_\theta(h_t, x_t)$$

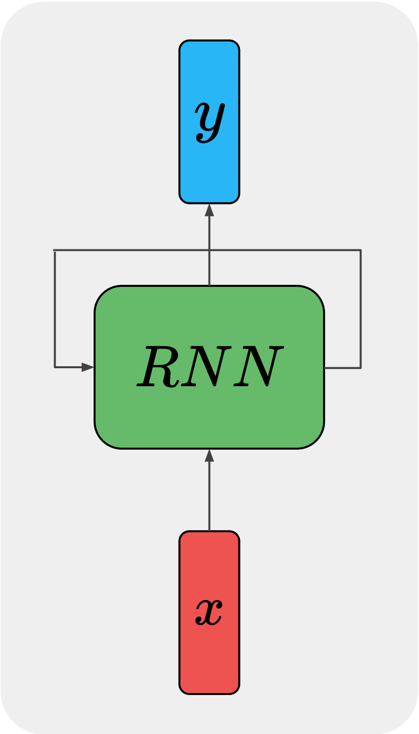


Unrolling RNNs



One-to-One One-to-Many Many-to-One Many-to-Many

Simple Implementation



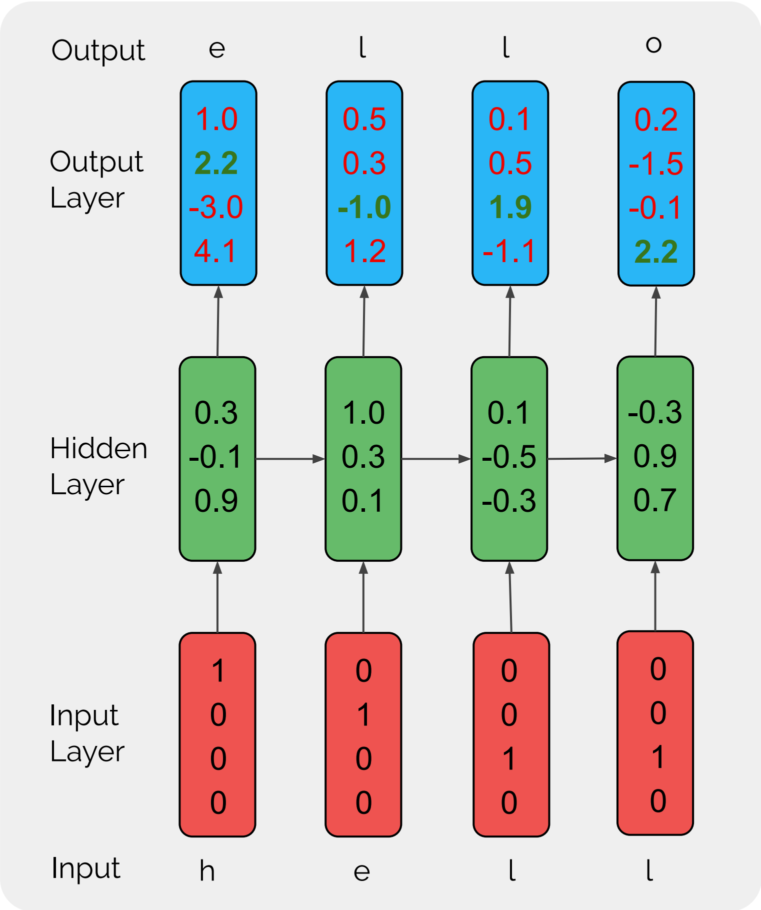
State consists of a single *"hidden"* vector **h**

$$h_{t+1} = f_\theta(h_t, x_t)$$

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$$

$$y_t = W_{hy} h_t$$

Generative RNNs



Autoregressive

- Output variable depends on its own previous values and on a stochastic term

Teacher Forcing

- During training, past y in input is from training data
- At generation time, past y in input is generated
- Mismatch can cause compounding error

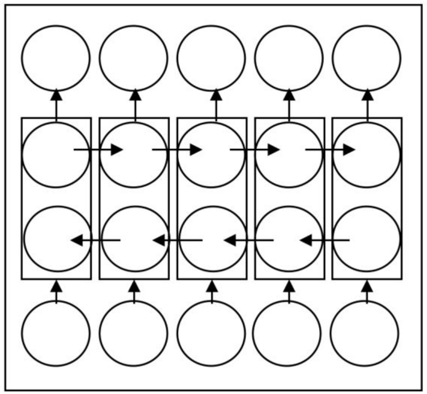
Scheduled Sampling

- Randomly pass output as input with probability ε
- Linearly increase ε through training

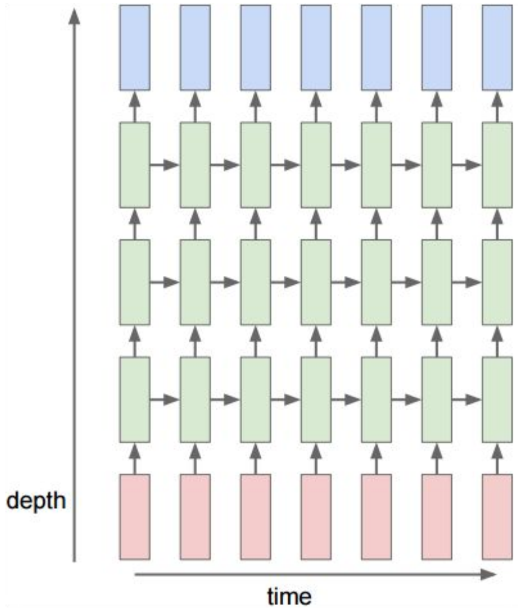
$$\mathcal{L}_t = -\log P(x_t | x_{t-1}, x_{t-2}, \ldots, x_1)$$

$$P(\mathbf{x}) = \prod_t P(x_t | x_{t-1}, x_{t-2}, \ldots, x_1)$$

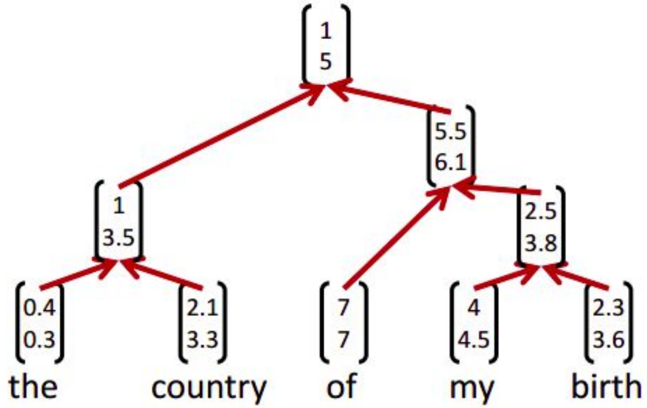
Other RNN Architectures



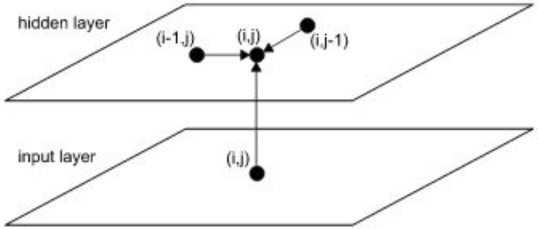
Bi-Directional



Stacked

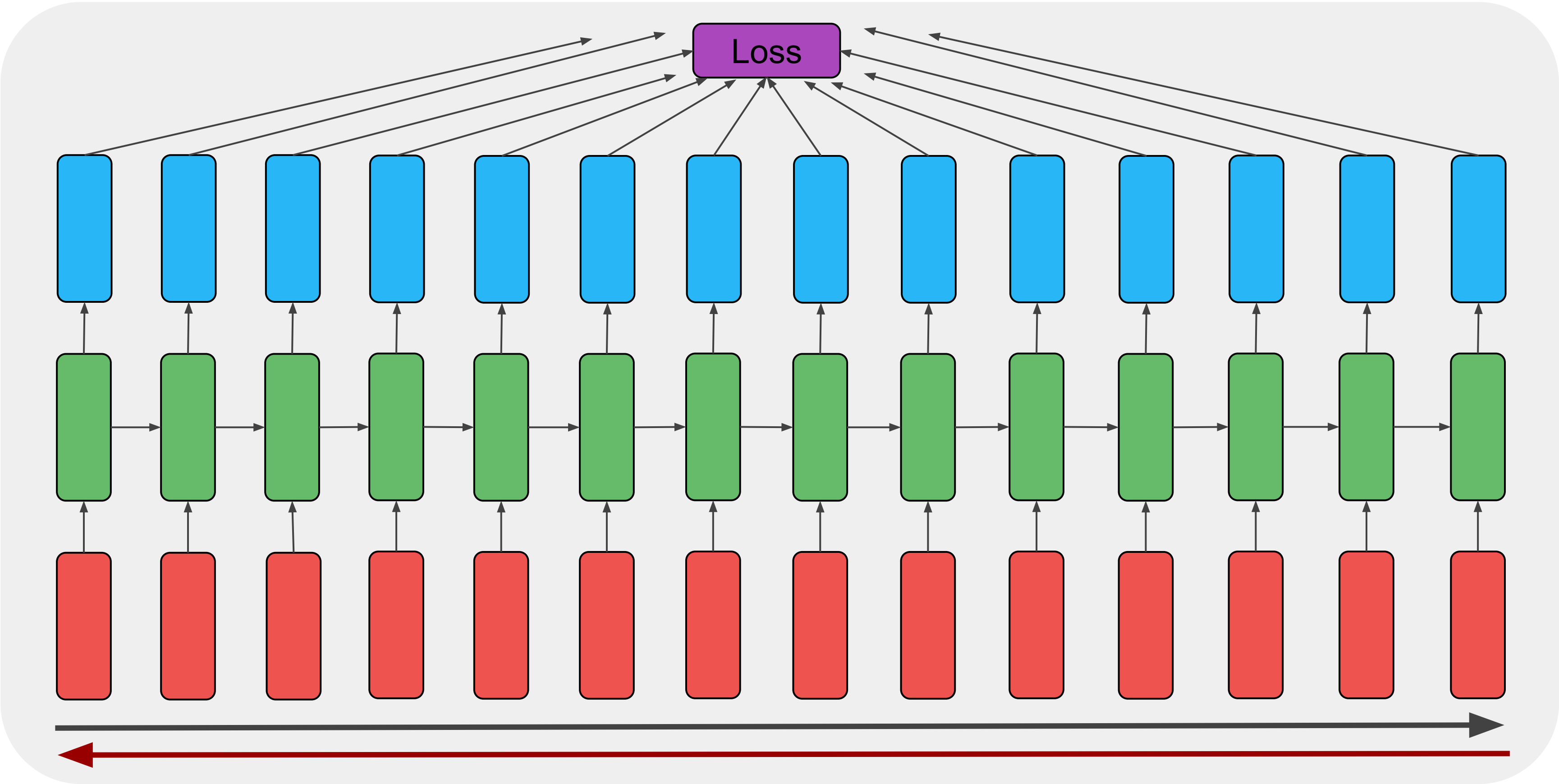


Recursive

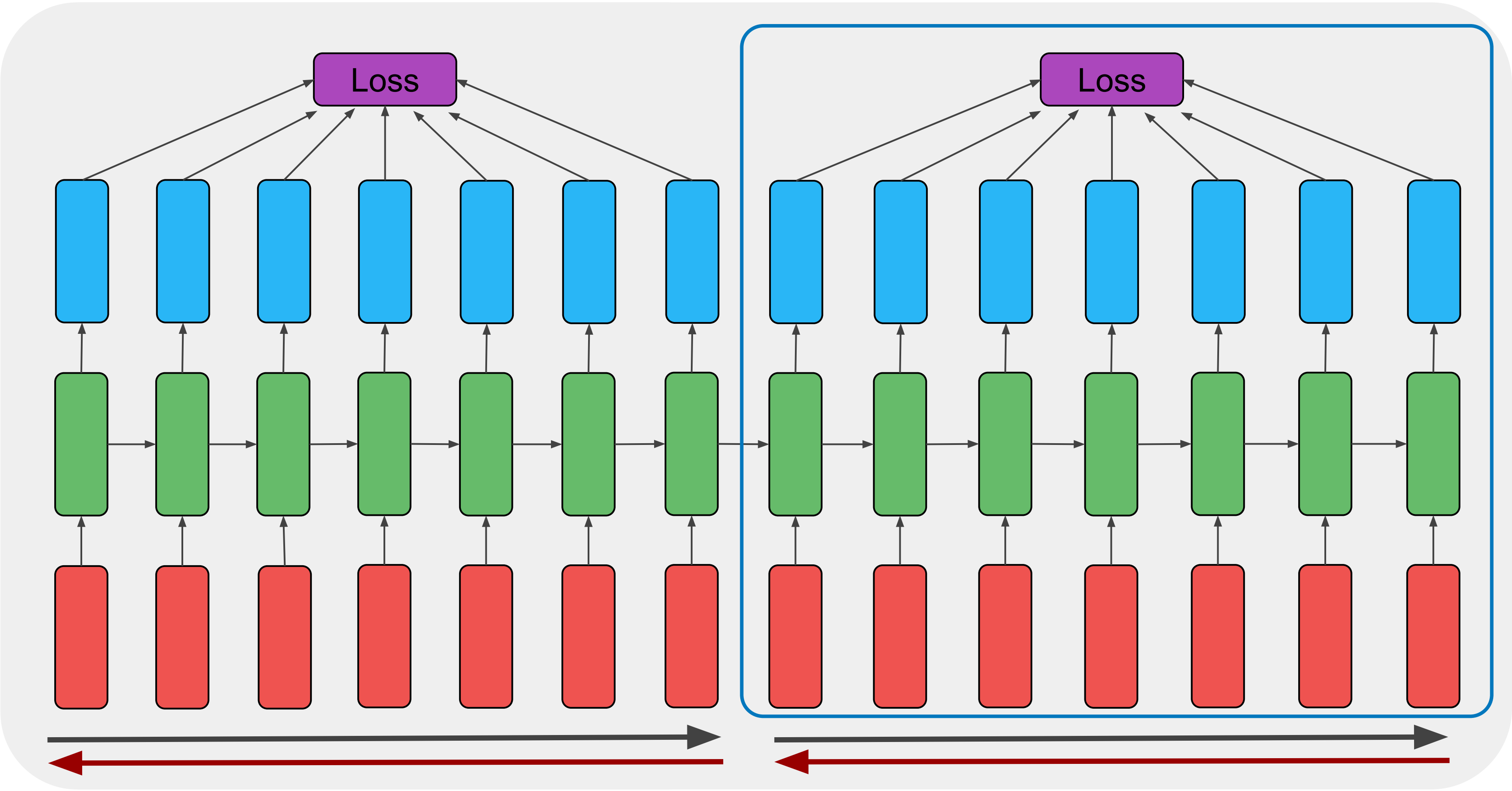


Multidimensional

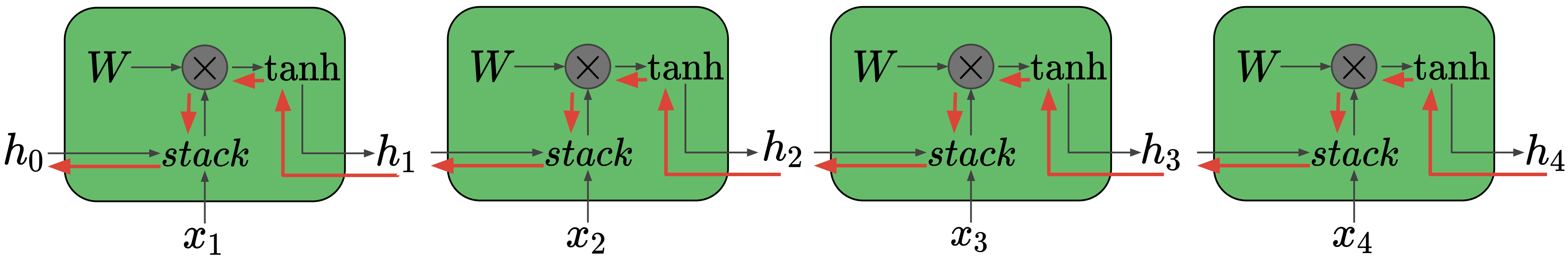
Backpropagation Through Time



Truncated Backprop Through Time



RNN Gradient Flow



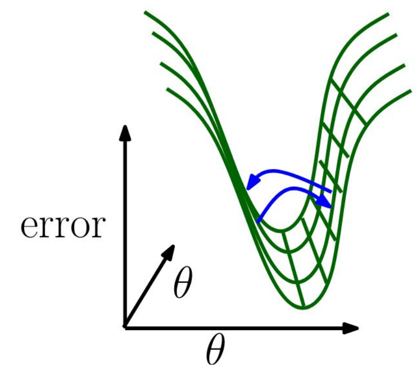
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$
$$= \tanh\left(W\begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$

$$\frac{\partial \mathcal{L}_T}{\partial W} = \sum_{t \le T} \frac{\partial \mathcal{L}_T}{\partial h_t}\frac{\partial h_t}{\partial W}$$
$$= \sum_{t \le T} \frac{\partial \mathcal{L}_T}{\partial h_T}\frac{\partial h_T}{\partial h_t}\frac{\partial h_t}{\partial W}$$

- Computing gradient of h_0 involves many factors of W (and repeated tanh).
- The product of T matrices whose spectral radius is < 1 is a matrix whose spectral radius converges to 0 at exponential rate in T

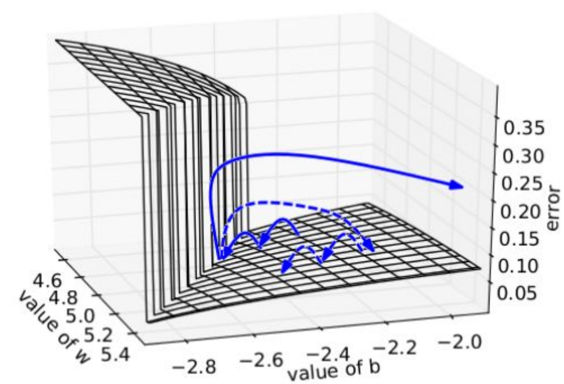
Exploding Gradients

- As parameters change, the asymptotic behaviour changes smoothly almost everywhere except for certain points where drastic changes occur
- The crossing of boundaries is sufficient for gradients to explode



Pascanu, Mikolov, Bengio, ICML 2013

$$
\begin{aligned}
&\hat{\mathbf{g}} \leftarrow \frac{\partial error}{\partial \theta} \\
&\textbf{if} \quad \|\hat{\mathbf{g}}\| \geq threshold \ \textbf{then} \\
&\quad \hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}} \\
&\textbf{end if}
\end{aligned}
$$



Vanishing Gradients: LSTMs

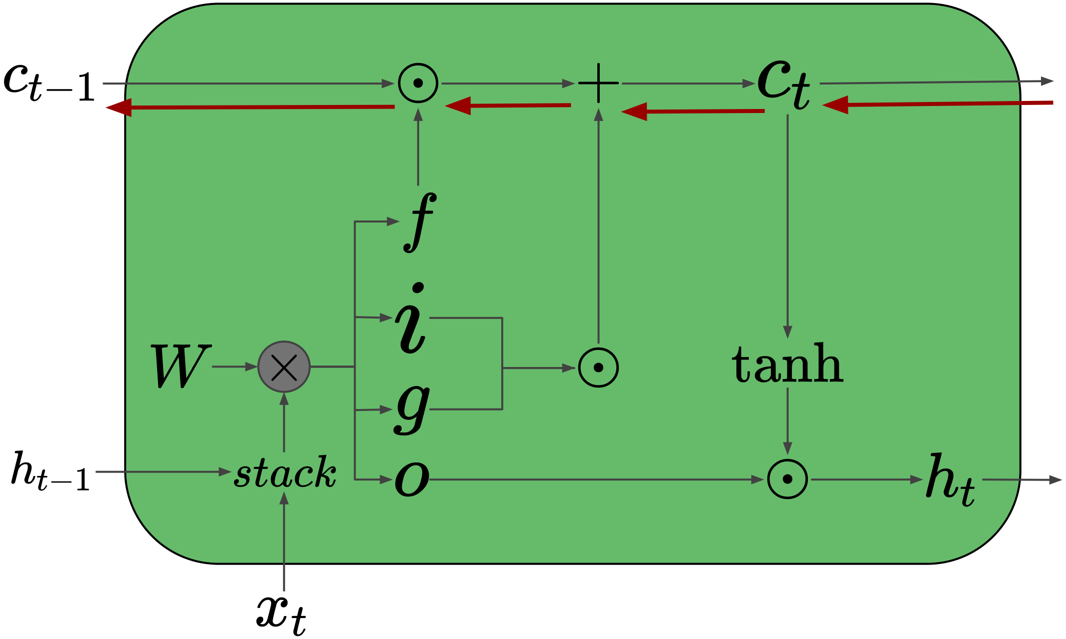
Backpropagation from c_t to c_{t-1} only elementwise multiplication by f , no matrix multiplication by W

f: Forget gate, whether to erase cell

i: Input gate, whether to write to cell

g: Gate gate, how much to write to cell

o: Output gate, how much to reveal cell



$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

Properties of RNNs

Main **Strengths**

- Allows for variable length sequences
- Efficient parameter usage
- <u>Theoretically</u> able to store arbitrarily old information

Main **Limitations**

- <u>Practically</u> unable to store very long term dependencies
- Limited by fixed size of hidden state
- Slow training and synthesis

Applications

Linux Source Code

```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)] | PFMR_CLOBATHINC_SECONDS << 12;
  return segtable;
}
```

Shakespeare

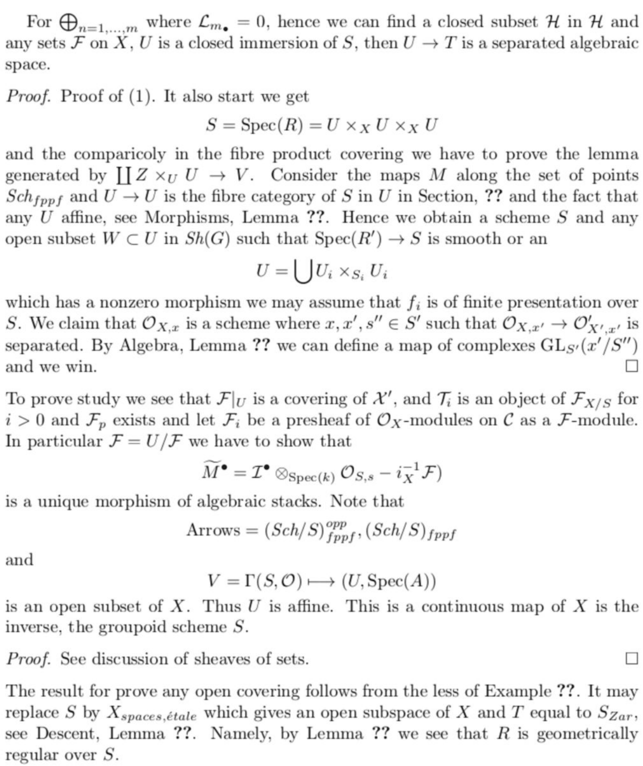
```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

Source: The Unreasonable Effectiveness of Recurrent Neural Networks, Andrej Karpathy

Applications

Durham University

Geometry (LaTeX)



For $\bigoplus_{n=1,\dots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ??. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

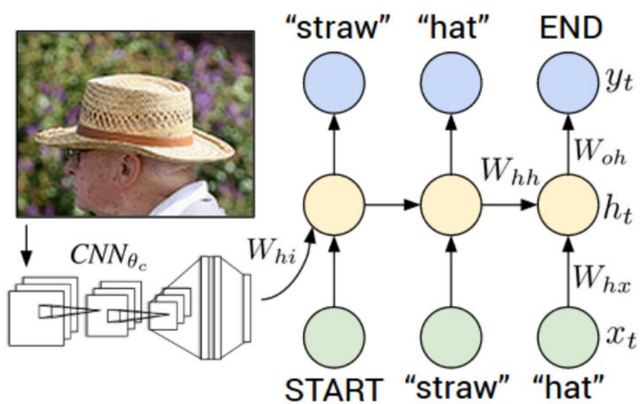
Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma ??. Namely, by Lemma ?? we see that R is geometrically regular over S.

Source: The Unreasonable Effectiveness of Recurrent Neural Networks, Andrej Karpathy

Applications

Image Captioning



A cat sitting on a suitcase on the floor

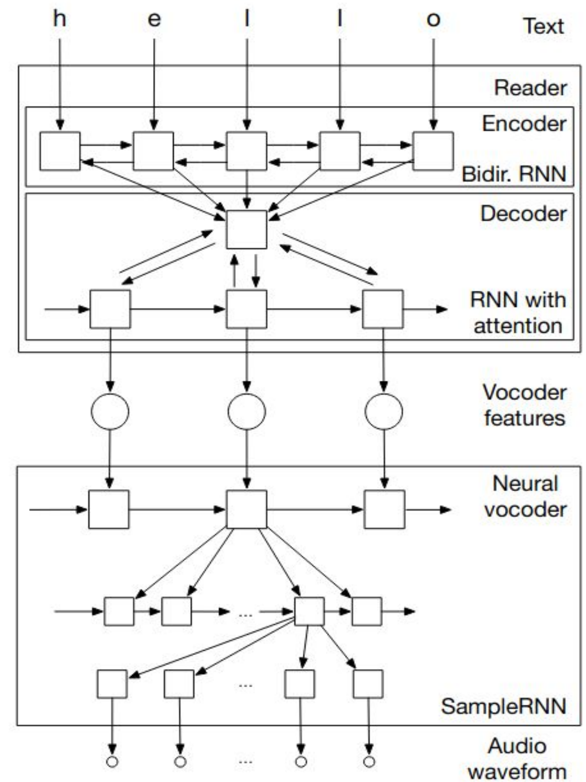
A cat is sitting on a tree branch

A dog is running in the grass with a frisbee

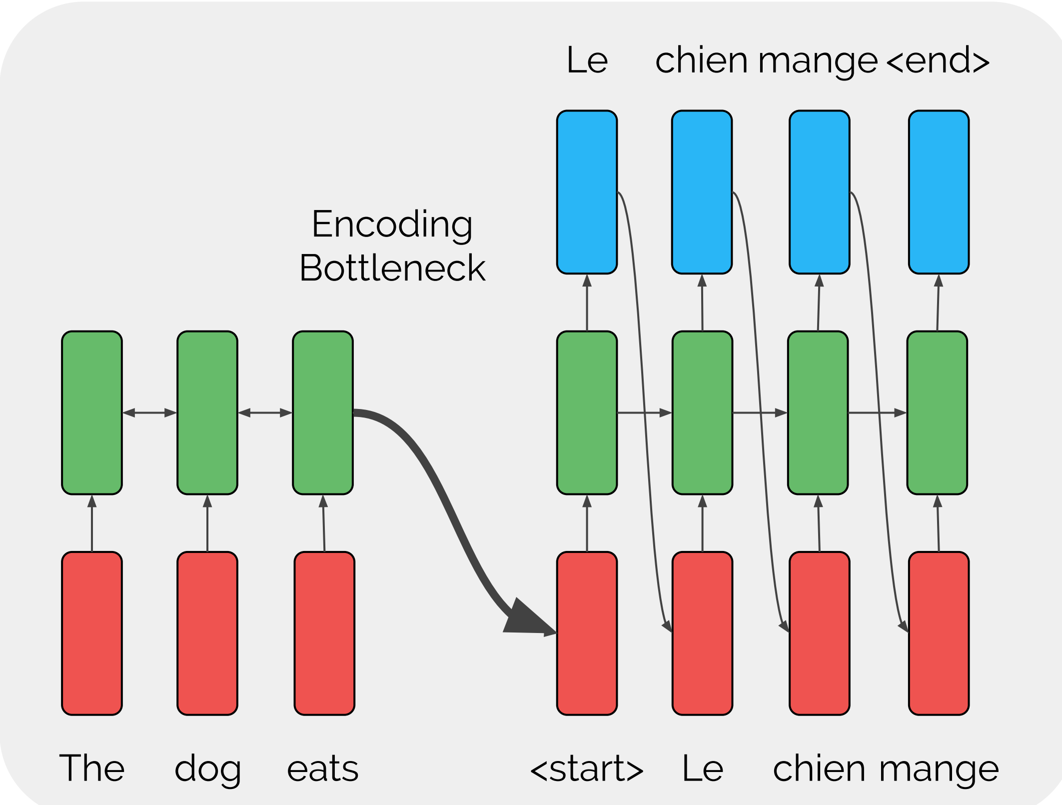
Source: Deep Visual-Semantic Alignments for Generating Image Descriptions

Source: Char2Wav: End-to-End Speech Synthesis

Speech Synthesis



Machine Translation



Architecture

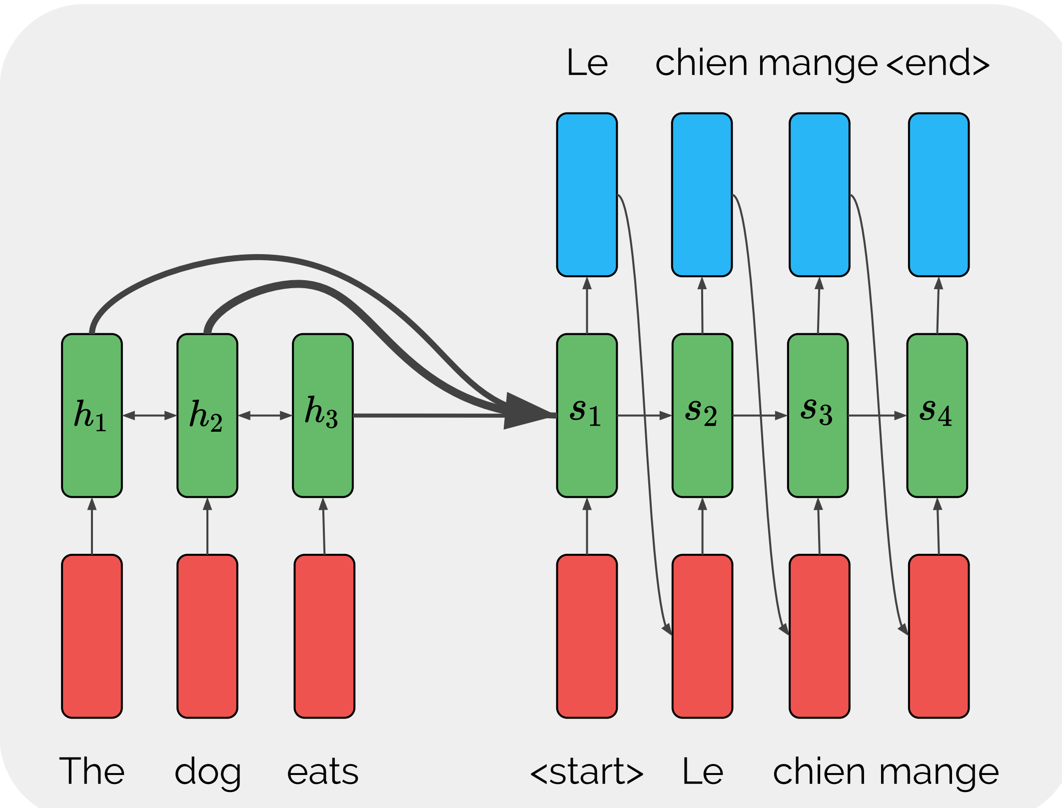
- Encoder: Bi-Directional RNN encodes input sentence
- Decoder: Autoregressive RNN synthesises translation
- Graph search decoder to find best translation

Problem

- Fixed length encoding vector is a bottleneck
- Want to access all hidden states of encoder

Neural Attention


Neural machine **translation** by **jointly** learning to **align** and **translate**
D Bahdanau, K Cho, Y Bengio - arXiv preprint arXiv:1409.0473, 2014 - arxiv.org
… The RNNencdec-50 correctly **translated** the source sentence until [a medical center … source sentences generated by the RNNencdec-50, RNNsearch-50 and Google **Translate** along with … traditional machine **translation** systems, all of the pieces of the **translation** system, including …
Cited by 10570 Related articles All 34 versions



Select what parts of encoding to look at during each step

Context vector

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

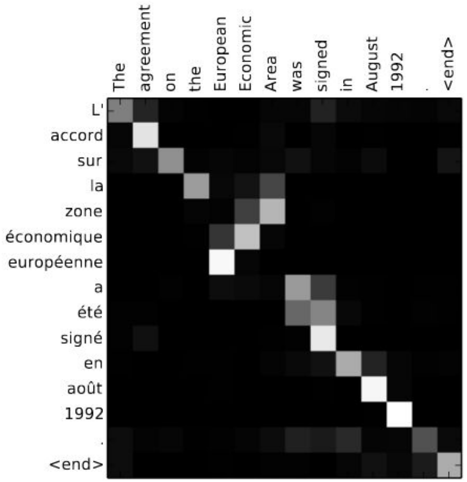
With weights

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

Using network a

$$e_{ij} = a(s_{i-1}, h_j)$$

Neural Attention: Examples



(a) **English→French (WMT-14)**

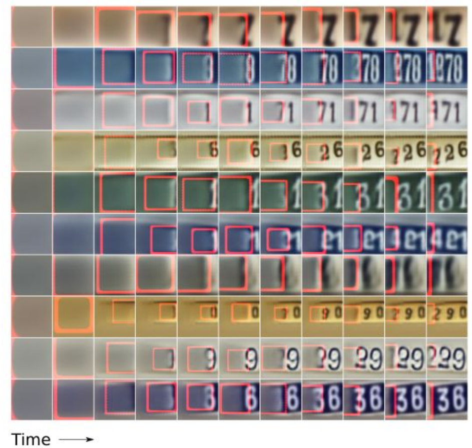
	NMT(A)	Google	P-SMT
NMT	32.68	30.6*	37.03•
+Cand	33.28	–	
+UNK	33.99	32.7°	
+Ens	36.71	**36.9°**	

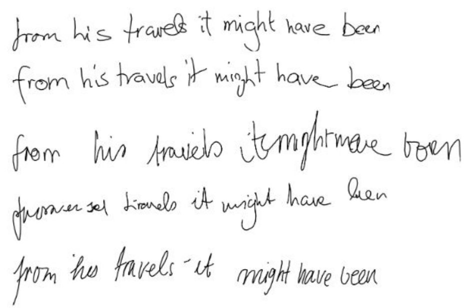
(b) **English→German (WMT-15)**

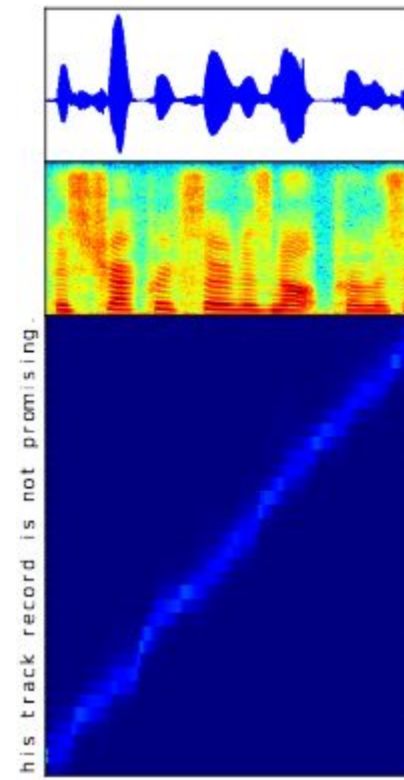
Model	Note
24.8	Neural MT
24.0	U.Edinburgh, Syntactic SMT
23.6	LIMSI/KIT
22.8	U.Edinburgh, Phrase SMT
22.7	KIT, Phrase SMT

(c) **English→Czech (WMT-15)**

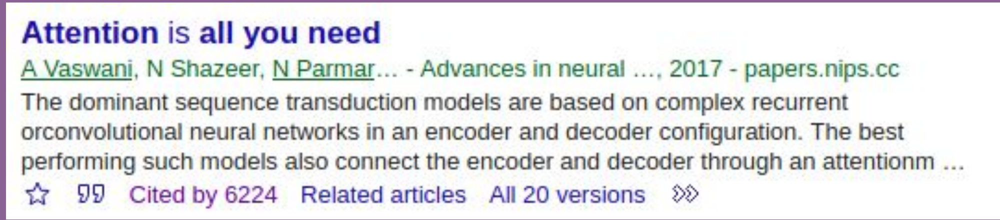
Model	Note
18.3	Neural MT
18.2	JHU, SMT+LM+OSM+Sparse
17.6	CU, Phrase SMT
17.4	U.Edinburgh, Phrase SMT
16.1	U.Edinburgh, Syntactic SMT







Transformers


Attention is all you need
A Vaswani, N Shazeer, N Parmar... - Advances in neural ..., 2017 - papers.nips.cc
The dominant sequence transduction models are based on complex recurrent orconvolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attentionm ...
Cited by 6224 Related articles All 20 versions

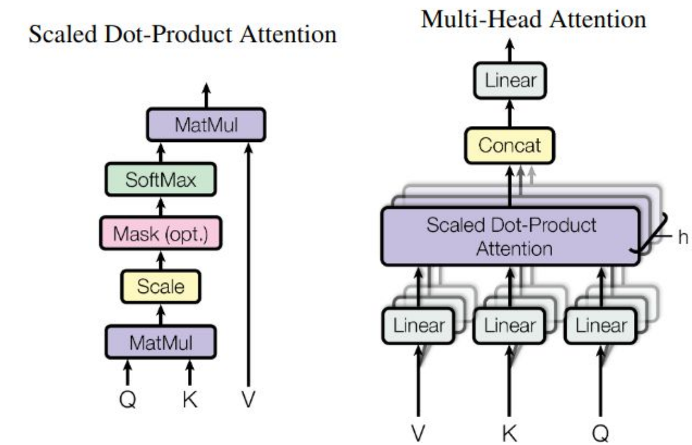
- RNN training is **sequential** and slow
- We can do everything with **attention**
- Dot product attention allows parallel training

Encode input sentence into

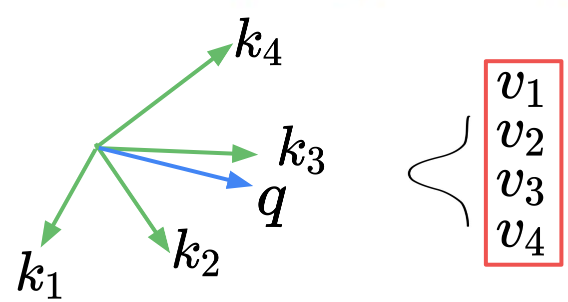
- **Values**: Information describing input, and
- **Keys**: A method to index Values

Discover attentively by making

- **Queries**: Requests for information

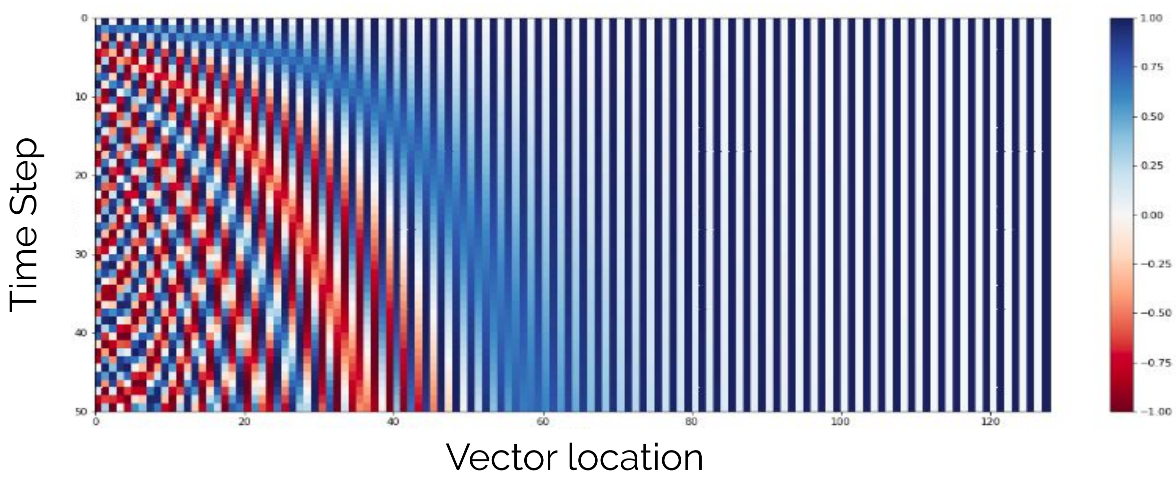


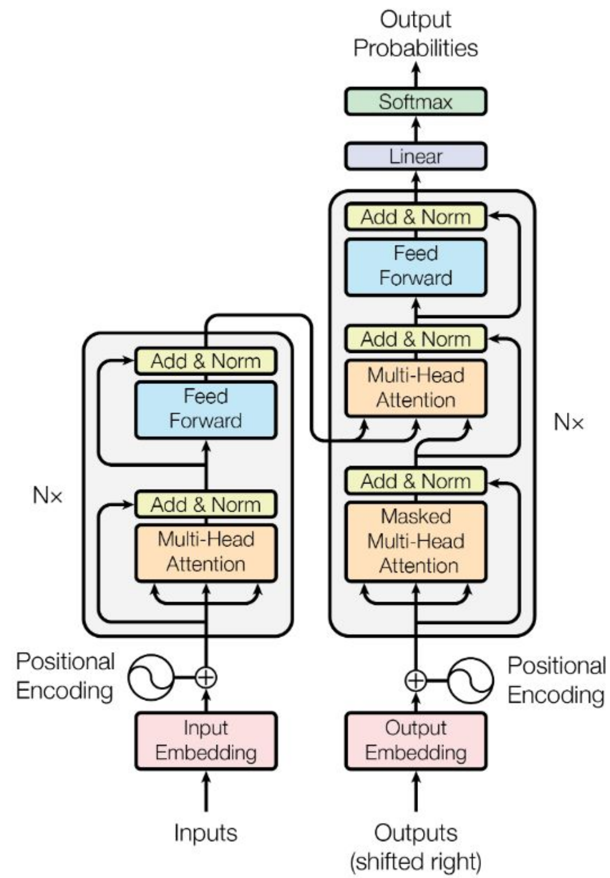
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Transformers

- Encoder-Decoder architecture
- Speedup allows for much larger powerful architectures
- Must add **sinusoidal encoding** as temporal information to allow attention by relative positions



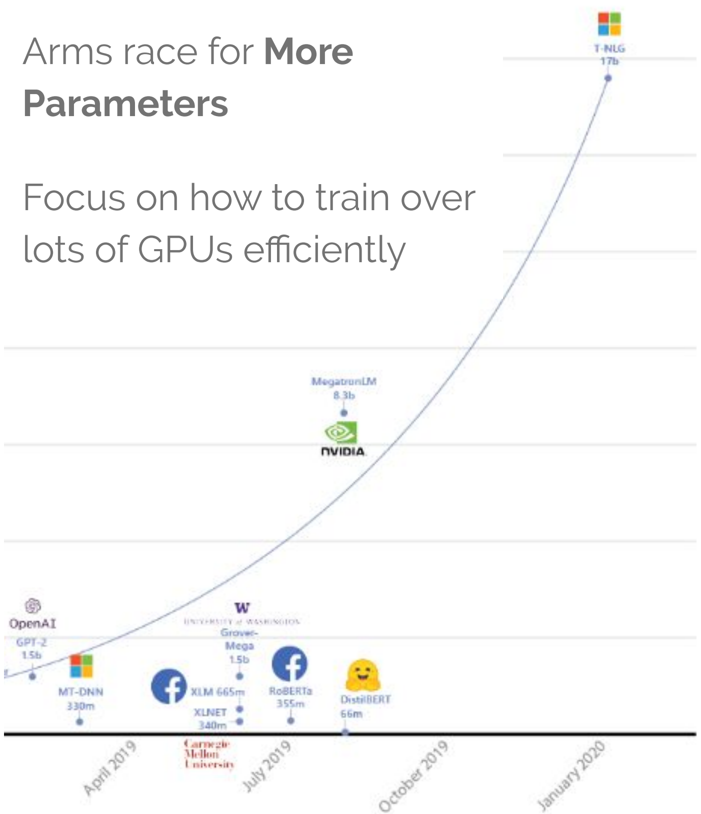


BIG Transformers

Durham University

Arms race for **More Parameters**

Focus on how to train over lots of GPUs efficiently



Context: The 36th International Conference on Machine Learning (ICML 2019) will be held in Long Beach, CA, USA from June 10th to June 15th, 2019. The conference will consist of one day of tutorials (June 10), followed by three days of main conference sessions (June 11-13), followed by two days of workshops (June 14-15).

Megatron-LM: With a broad scope, the conference addresses the challenges and opportunities in machine learning for practitioners and researchers. The conference program will feature sessions that cover different machine learning topics, from reinforcement learning to computational biology and from AI to deep learning. It is the world's premier forum for exploring and presenting advances in machine learning.

Question	Who was Jason Mraz engaged to?
Passage	Mraz was engaged to singer/songwriter and long-time close friend **Tristan Prettyman** on Christmas Eve 2010; they broke off the engagement six months later.
"Direct" Answer	Jason Mraz was engaged to Tristan Prettyman.

Link to original text	**T-NLG Summary**
"Microsoft will be carbon negative by 2030" by Brad Smith, Official Microsoft Blog"	Microsoft is committed to being carbon negative by 2030. We are launching an aggressive program to cut our carbon emissions by more than half by 2030, both for our direct emissions and for our entire supply and value chain. We are also launching an initiative to use Microsoft technology to help our suppliers and customers reduce their own carbon footprints and a new $1 billion climate innovation fund to accelerate the development of carbon reduction, capture, and removal technologies that will help us and the world become carbon negative. In addition to our aggressive carbon goals, we are launching a new Climate Innovation Fund to accelerate carbon reduction and removal opportunities. We are also launching a program to use our technology to improve the efficiency of our supply chain and reduce our own carbon footprint as well...

Take Away Points

Durham University

RNN   Transformer

- RNNs are great for sequential data
- But they struggle with backpropagation
  - Clip your gradients to prevent exploding gradients
  - Use an LSTM or GRU to prevent vanishing gradients
- Use Attention to help learn long term dependencies
- Transformers are currently the most popular
  - They perform similar or much better than RNNs and scale much more nicely

Further Reading

- https://karpathy.github.io/2015/05/21/rnn-effectiveness/
- https://weberna.github.io/blog/2017/11/15/LSTM-Vanishing-Gradients.html
- https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html