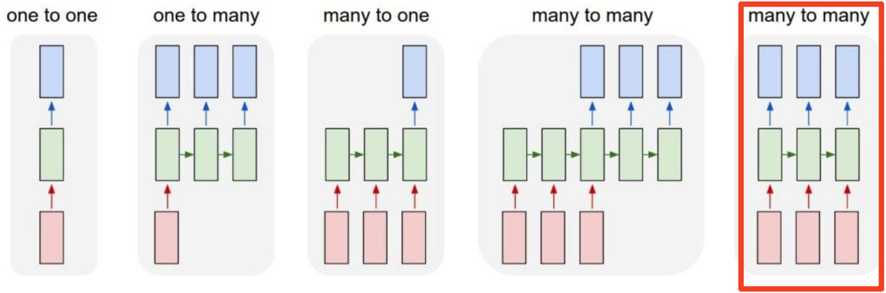
Lecture 11: Sequence-to-sequence architectures. Neural attention and memory.

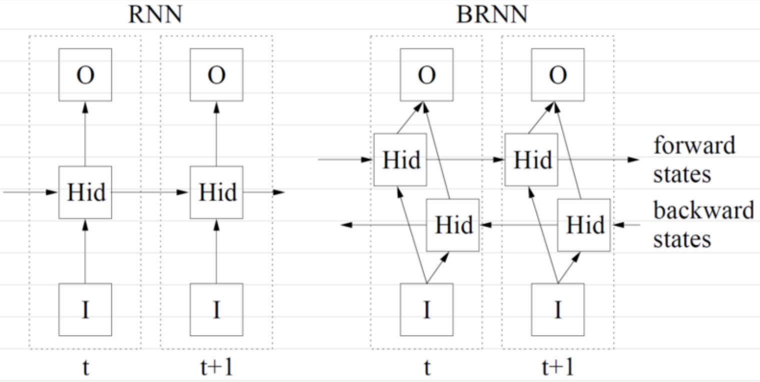
Learning settings

slide credit: A. Karpathy



One-to-one: image to class label
One-to-many: text generation/image captioning
Many-to-one: sentiment analysis
Many-to-many 1: machine translation
Many-to-many 2: online classification (e.g. POS tagging)

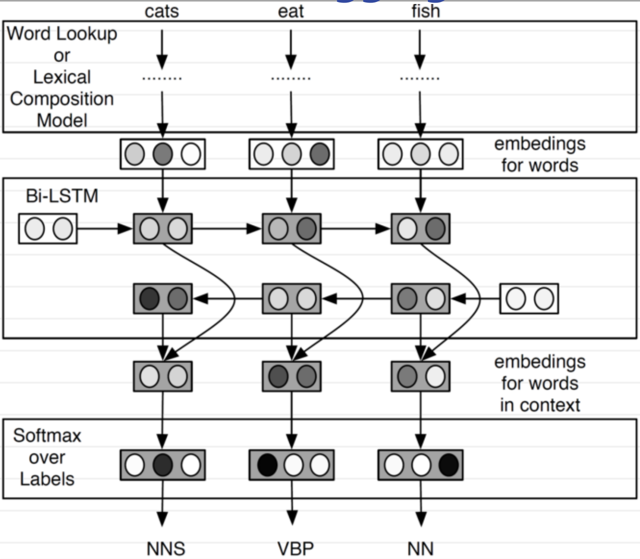
Bi-directional RNN




```
for t = 1 to T do
  Do forward pass for the forward hidden layer, storing activations at
  each timestep
for t = T to 1 do
  Do forward pass for the backward hidden layer, storing activations at
  each timestep
for t = 1 to T do
  Do forward pass for the output layer, using the stored activations from
  both hidden layers
```

[A Graves, PhD thesis]

Bi-LSTM POS tagging



$$\mathbf{l}_i = \tanh(\mathbf{L}^f \mathbf{s}_i^f + \mathbf{L}^b \mathbf{s}_i^b + \mathbf{b}_l)$$

[Ling et al. EMNLP15]

Bi-LSTM POS tagging

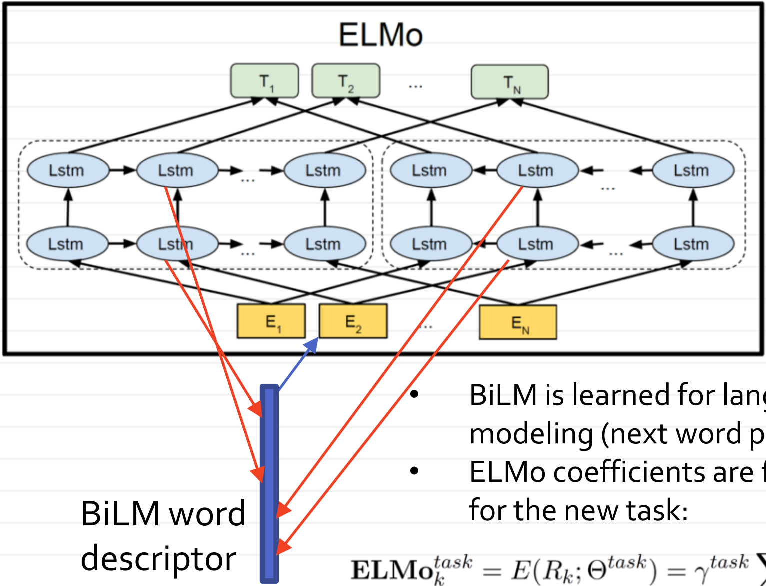
	acc	parameters	words/sec
Word Lookup	96.97	2000k	6K
Convolutional (S&Z)	96.80	42.5k	4K
Forward RNN	95.66	17.5k	4K
Backward RNN	95.52	17.5k	4K
Bi-RNN	95.93	40k	3K
Forward LSTM	97.12	80k	3K
Backward LSTM	97.08	80k	3K
Bi-LSTM $d_{CS} = 50$	97.22	70k	3K
Bi-LSTM	**97.36**	150k	2K

[Ling et al. EMNLP15]

Uni-directional vs bi-directional

- Bi-directional is not applicable when "future" is unavailable
- When future is available bi-directional is almost always better
- E.g. NLP (batch mode), bioinformatics

BiLM/ELMO



BiLM word descriptor

- BiLM is learned for language modeling (next word prediction)
- ELMo coefficients are fine-tuned for the new task:

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}$$

[Peters et al. ACL18]

BiLM/ELMO

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

SNLI task [Bowman et al. EMNLP15]:

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

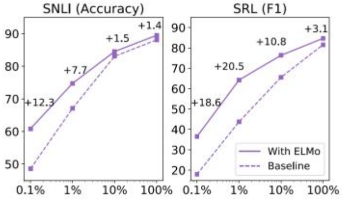
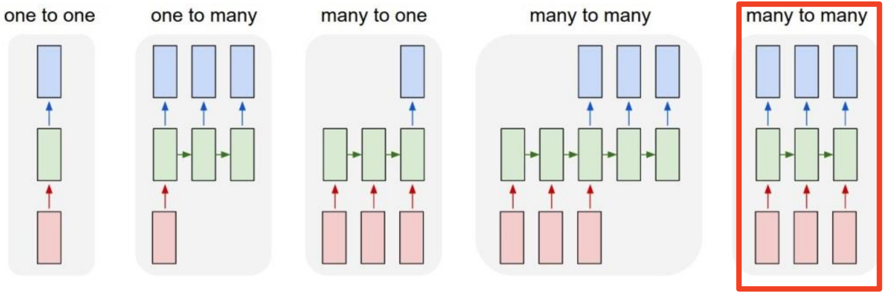


Figure 1: Comparison of baseline vs. ELMo performance for SNLI and SRL as the training set size is varied from 0.1% to 100%.

[Peters et al. ACL18]

Learning settings

slide credit: A. Karpathy

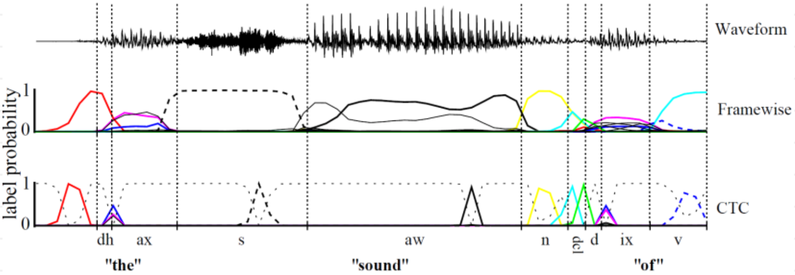


One-to-one: image to class label
One-to-many: text generation/image captioning
Many-to-one: sentiment analysis
Many-to-many 1: machine translation
Many-to-many 2: online classification (e.g. POS tagging)

Online seq2seq with monotonic alignment

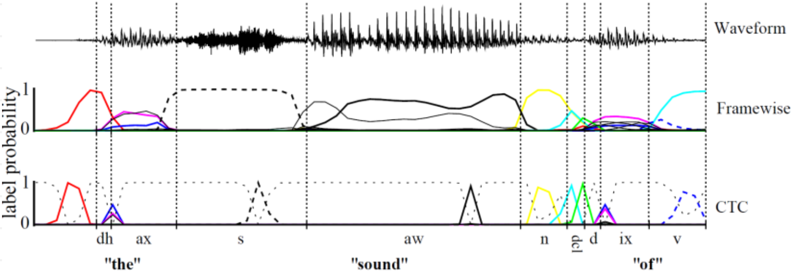
Many problems are sequence 2 sequence with monotonic alignment:

- Not one-to-one as sequence prediction or POS tagging
- More constrained than general seq2seq



[Graves et al. 2006]

Online seq2seq with monotonic alignment



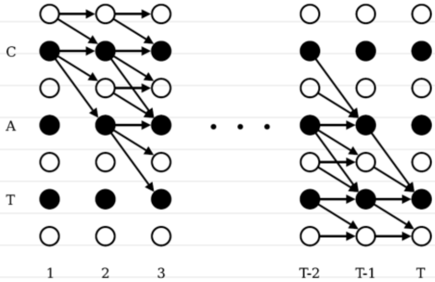
Decoding: ‘aaa__bb_c____ddaa’ → abcda

What should be the loss that encourage correct parsing?

Answer: connectionist temporal classification (CTC) loss

[Graves et al. 2006]

CTC-loss

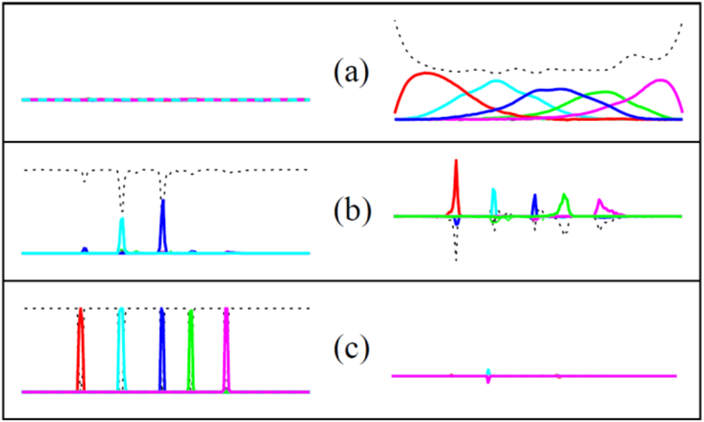


- Augment the output state with *blank*
- Predict probabilities of each symbol (inc. blank) at each time moment
- Compute the probability of each lattice vertex under correct paths using forward-backward
- Push log-probabilities up (*ML training*) proportionally to the current probability

[Graves et al. 2006]

Evolution of the CTC signal

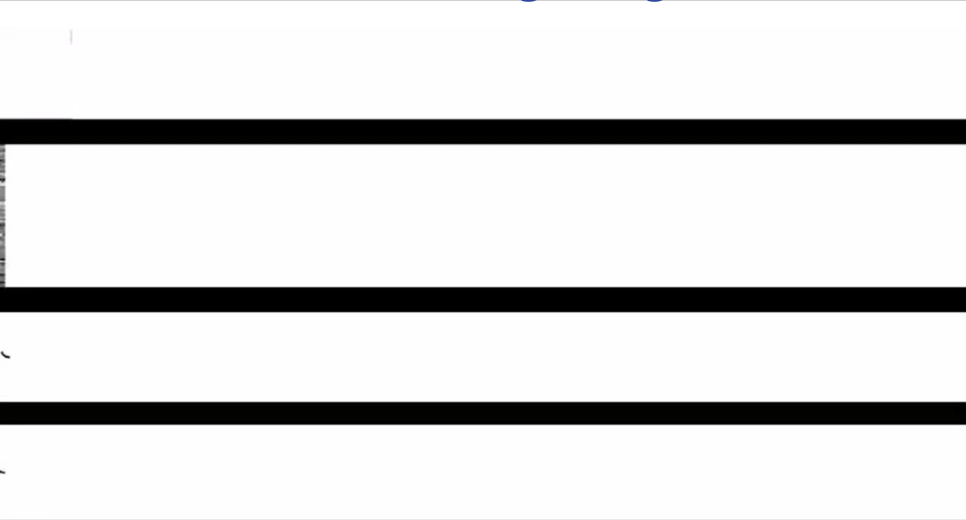
GT sequence: 



Prediction Gradient w.r.t. prediction

[Graves et al. 2006]

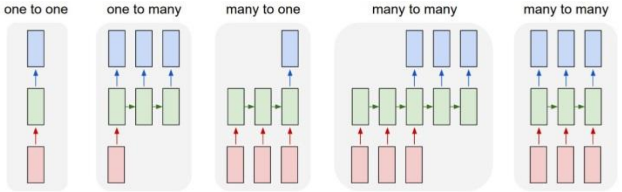
LSTM demo: handwriting recognition

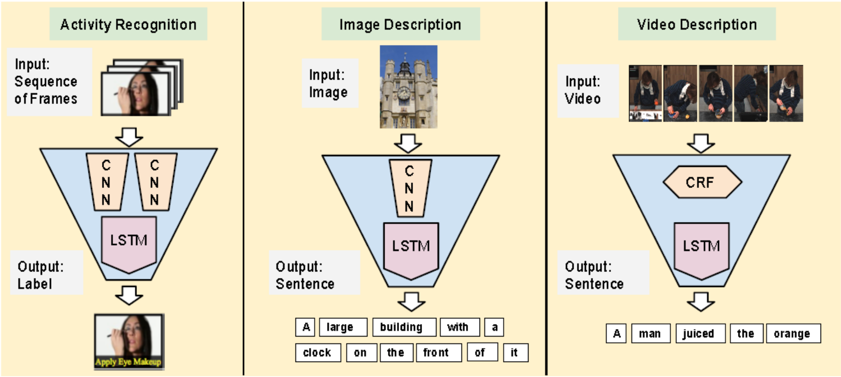


LSTM RNN Demo by Nikhil Buduma:

https://www.youtube.com/watch?v=mLxsbWAYIpw

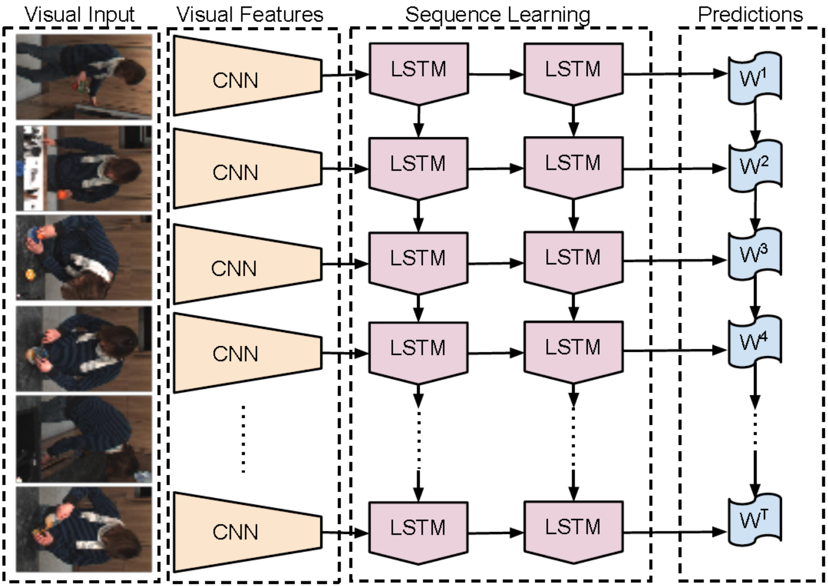
Image/video captioning





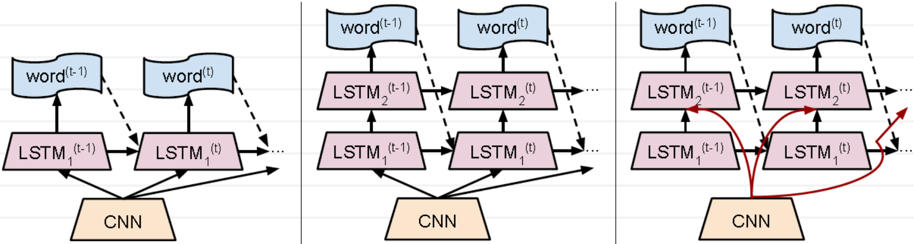
[Donahue et al. 2015]

Image/video captioning



[Donahue et al. 2015]*

Image/video captioning



Single Layer **LRCN$_{1u}$** | Two Layers, Unfactored **LRCN$_{2u}$** | Two Layers, Factored **LRCN$_{2f}$**

- Train on 108,000 images with descriptions
- Test on 1000 images (5 descr per image)
- For each image score 5000 descriptions
- See if top-k has a correct description:

	R@1	R@5	R@10	Med$r$
LRCN$_{1u}$	14.1	31.3	39.7	24
LRCN$_{2u}$	3.8	12.0	17.9	80
LRCN$_{2f}$	**17.5**	**40.3**	**50.8**	**9**
LRCN$_{4f}$	15.8	37.1	49.5	10

[Donahue et al. 2015]

Image/video captioning

Best results:

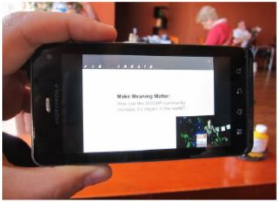
A female tennis player in action on the court.

A group of young men playing a game of soccer

A man riding a wave on top of a surf-board.

A baseball game in progress with the batter up to plate.

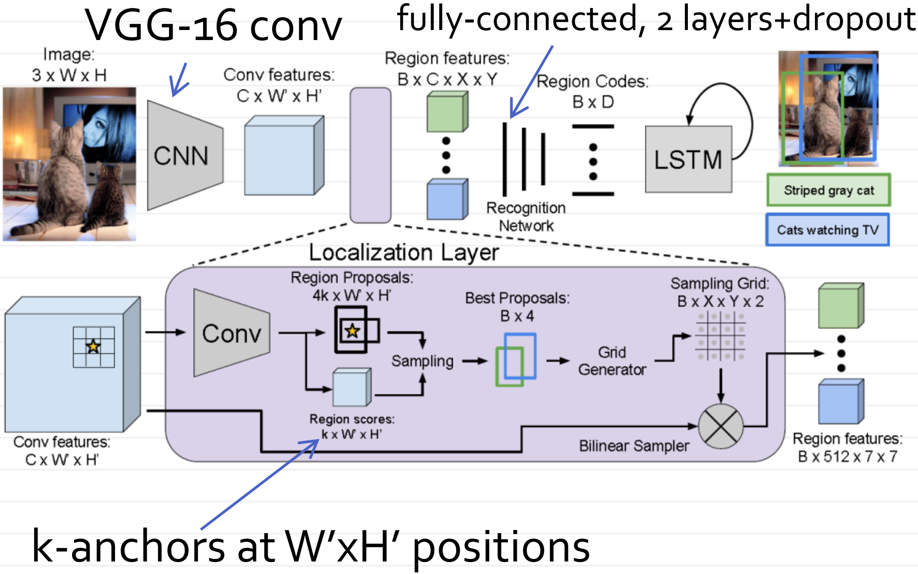
A brown bear standing on top of a lush green field.



A person holding a cell phone in their hand.

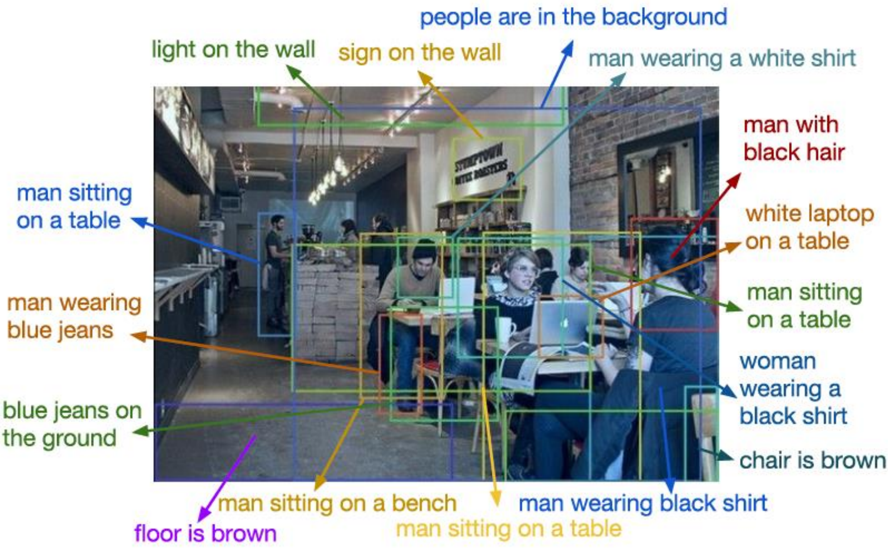
[Donahue et al. 2015]

End-to-end dense image captioning



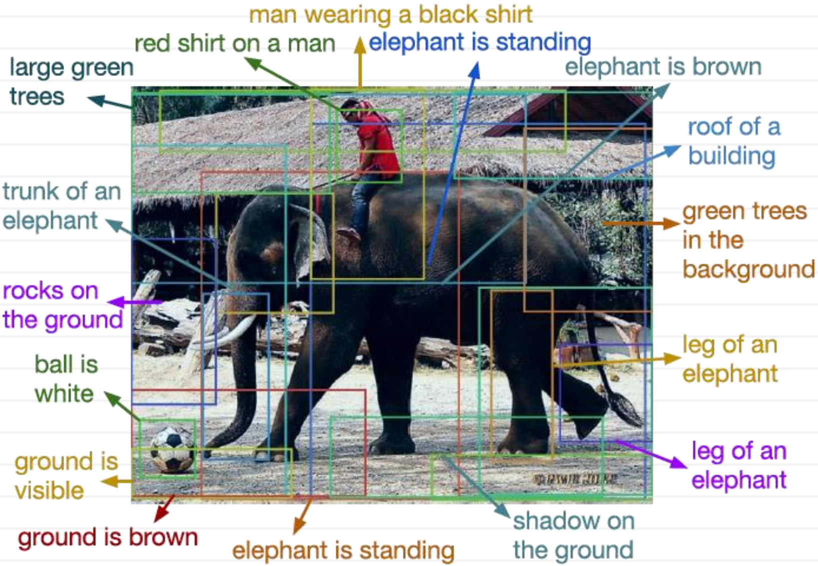
[Johnson et al, CVPR16]

End-to-end dense image captioning



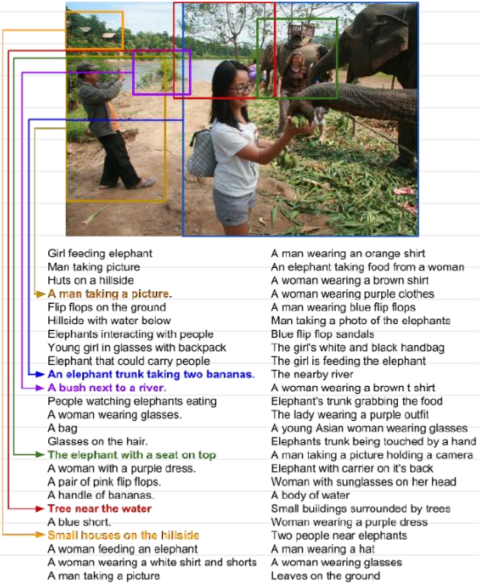
[Johnson et al, CVPR16]

End-to-end dense image captioning



[Johnson et al, CVPR16]

Training set: “visual genome”



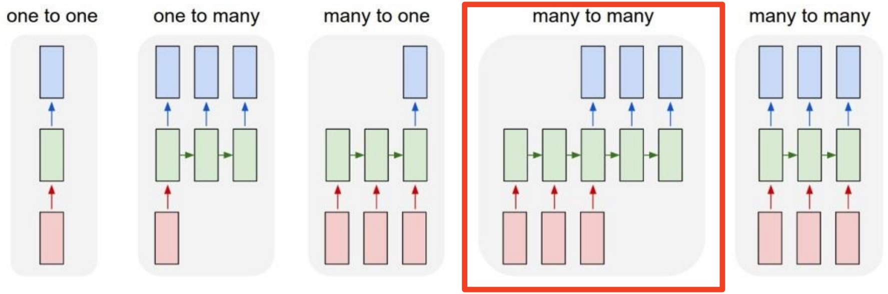
- “New Image-net”

108,249 Images
4.2 Million Region Descriptions
1.7 Million Visual Question Answers
2.1 Million Object Instances
1.8 Million Attributes
1.8 Million Relationships
Everything Mapped to Wordnet Synsets

[Krishna et al. 2016]

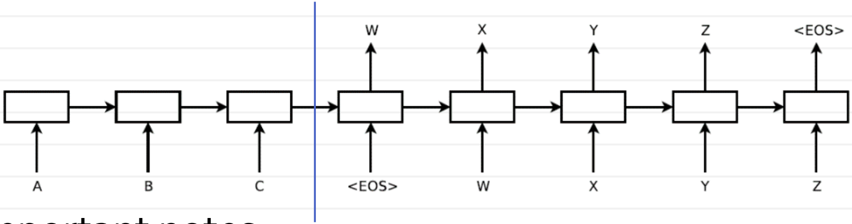
Learning settings

slide credit: A. Karpathy



aka "seq2seq"

Sequence-to-sequence machine translation



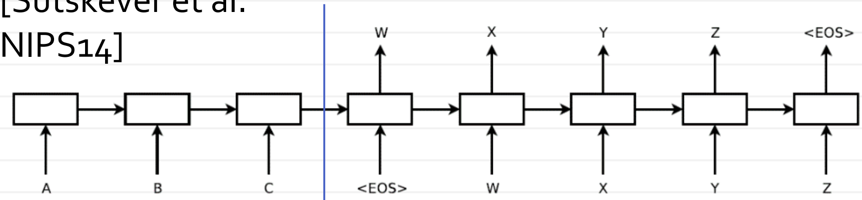
Important notes:

1. Fixed lexicon (160,000 English, 80,000 French) + 'UNK' word
2. Deep (four layers, 1000 cells in each)
3. Reversing input sequence helps a lot
4. Using two different LSTMs
5. Decoding proceeds by *beam search*

[Sutskever et al. NIPS14]

Sequence-to-sequence machine translation

[Sutskever et al. NIPS14]



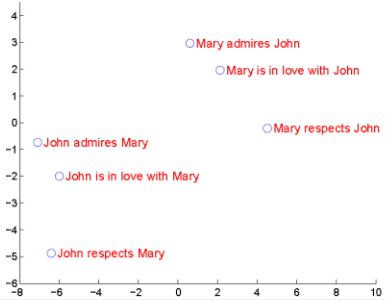
Decoding proceeds by *beam search:*

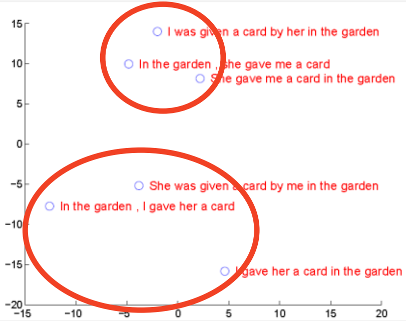
1. At the first step generate top-K words
2. At each step, expand each of the K in top-L ways (gives KL results)
3. Pick the best K out of KL results

NB: needs some mechanism to compare sequences of different lengths

Sequence-to-sequence machine translation

Learned embeddings:





PCA 1000-> 2

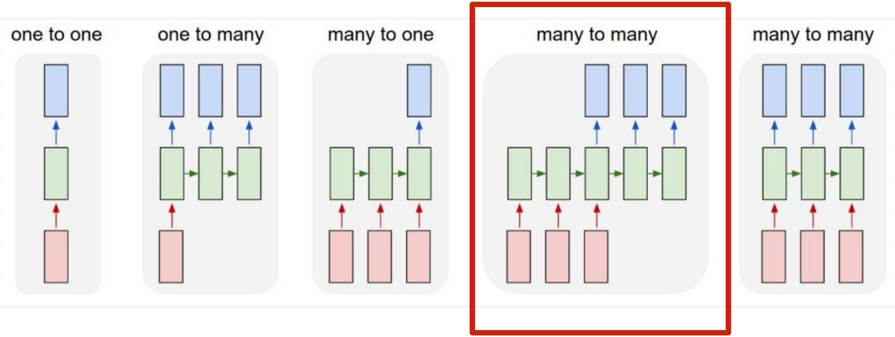
[Sutskever et al. NIPS14]

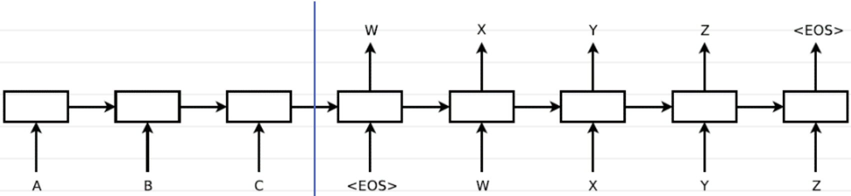
Sequence-to-sequence machine translation

Type	Sentence
Our model	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
Truth	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
Our model	" Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK .
Truth	" Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker .
Our model	Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " .
Truth	Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " .

[Sutskever et al. NIPS14]

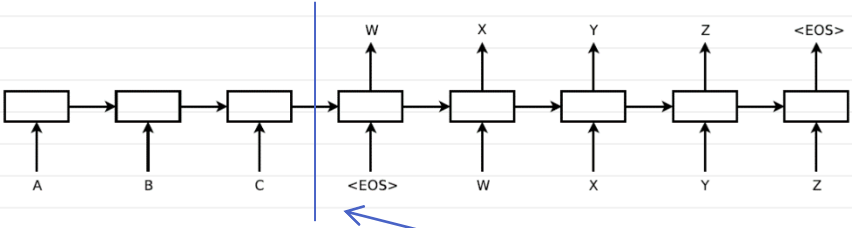
Sequence-to-sequence machine translation





[Sutskever et al. NIPS14]

Sequence-to-sequence machine translation

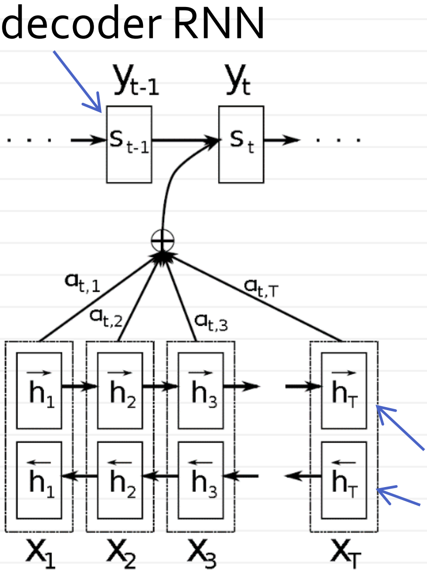


Problem:
all the meaning has to be carried from here

- Large memory needed
- Information has to survive for a very long time

[Sutskever et al. NIPS14]

Translation with attention



$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

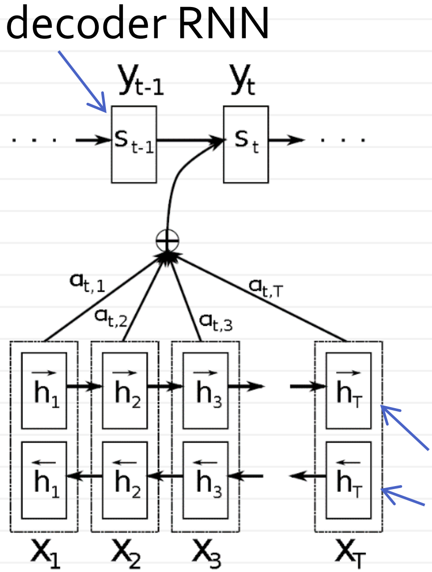
$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

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

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

[Bahdanau et al. 2015]

Translation with attention

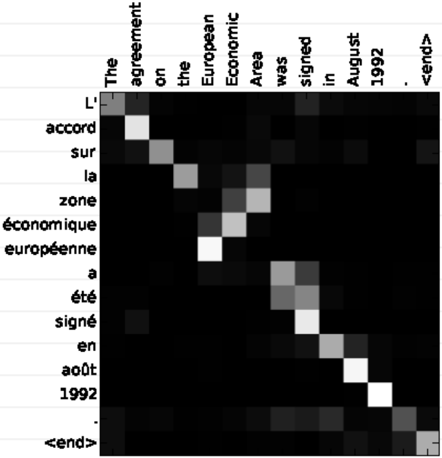


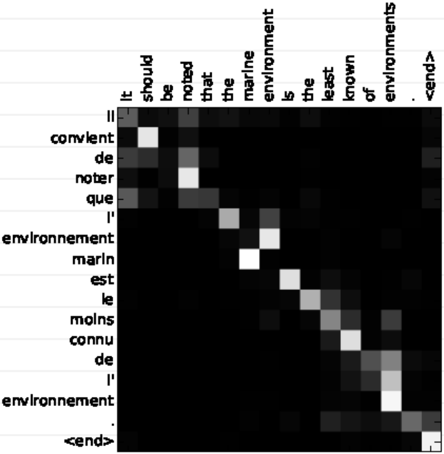
$$e_{ij} = a(s_{i-1}, h_j)$$

- *Attention model:* feed-forward neural network
- All components are trained end-to-end

[Bahdanau et al. 2015]

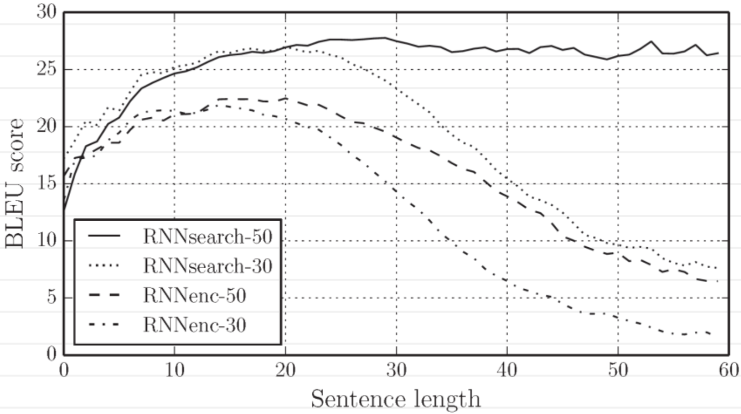
Translation with attention





[Bahdanau et al. 2015]

Translation with attention



- BLEU-score ≈ precision over n-grams
- Trained either with <30 word phrases or with <50 word phrases

[Bahdanau et al. 2015]

Translation with attention

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

LSTM system:

Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

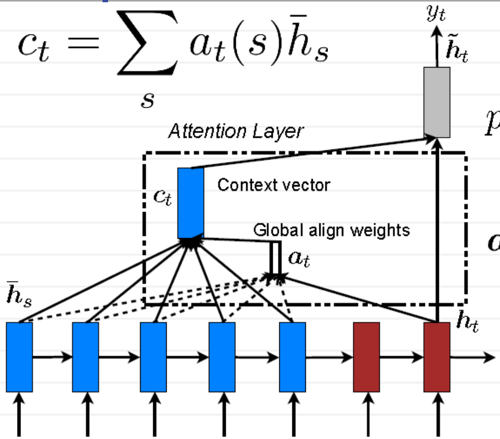
Attention-based system:

Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.

[Bahdanau et al. 2015]

Simpler translation with attention

$$c_t = \sum_s a_t(s)\bar{h}_s$$



$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W_c}[\boldsymbol{c}_t;\boldsymbol{h}_t])$$

$$p(y_t|y_{<t},x) = \mathrm{softmax}(\boldsymbol{W_s}\tilde{\boldsymbol{h}}_t)$$

$$\boldsymbol{a}_t(s) = \mathrm{align}(\boldsymbol{h}_t,\bar{\boldsymbol{h}}_s) = \frac{\exp\left(\mathrm{score}(\boldsymbol{h}_t,\bar{\boldsymbol{h}}_s)\right)}{\sum_{s'}\exp\left(\mathrm{score}(\boldsymbol{h}_t,\bar{\boldsymbol{h}}_{s'})\right)}$$

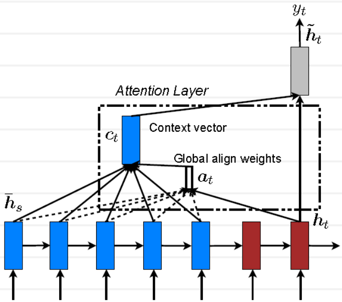
Quadratic complexity!

$$\mathrm{score}(\boldsymbol{h}_t,\bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top\bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top\boldsymbol{W_a}\bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top\tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t;\bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

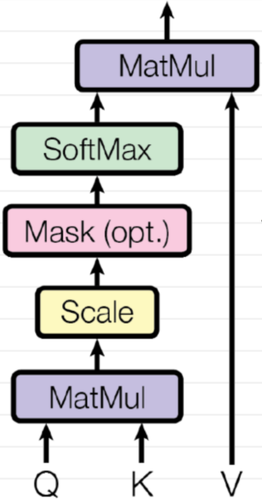
[Luong et al. 2015]

Recap

- Attention solved the limited memory problem
- Complexity is quadratic (in the length of sequence)



"Attention is all you need": single head



General purpose "single-head" attention:

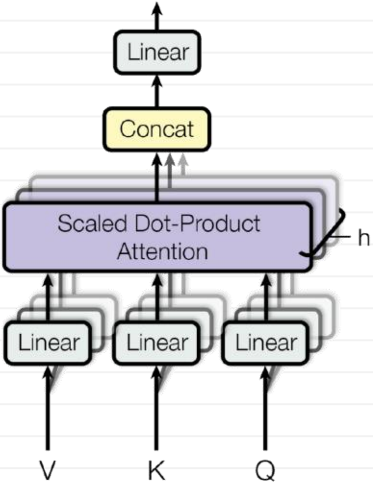
$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Recombining previous layer (V) in a long-range way using few parameters

[Vaswani et al. NIPS17]

"Attention is all you need": multiple heads

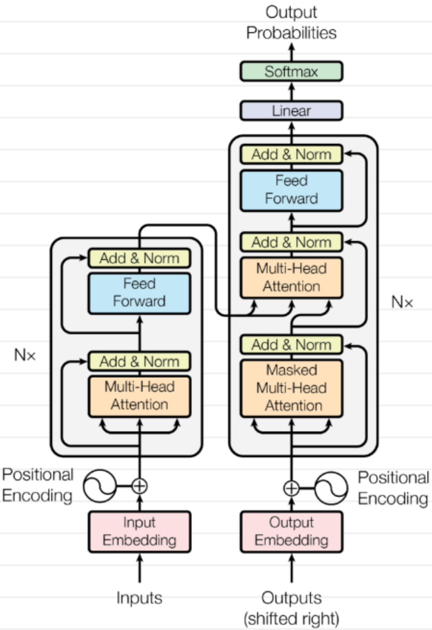
[Vaswani et al. NIPS17]



General purpose "multi-head" attention:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where}\ \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Transformer architecture



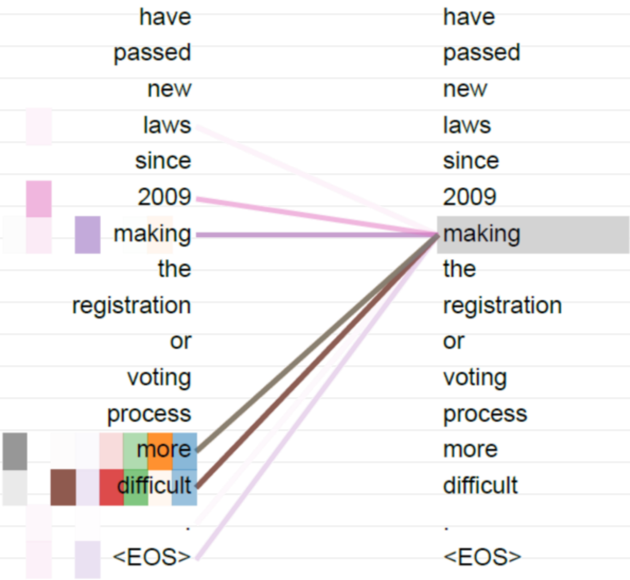
- Applying multi-head attention several times first for input, then for output
- Each unit is residual
- Emitting output one word at a time
- Positional encoding adds position-dependent features:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

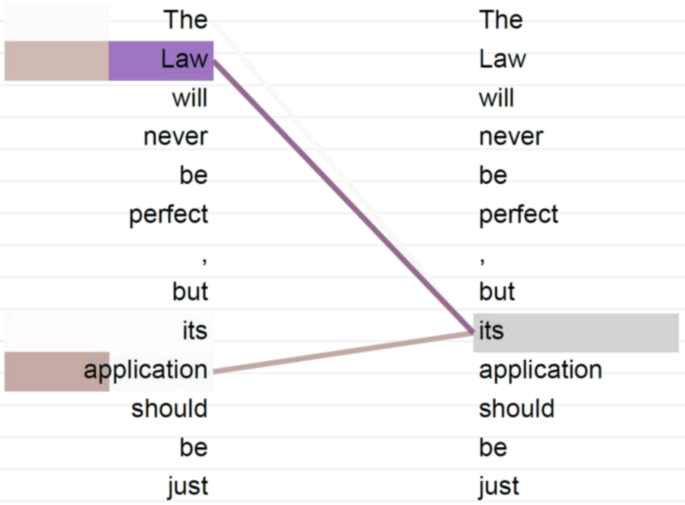
[Vaswani et al. NIPS17]

“Attention is all you need”



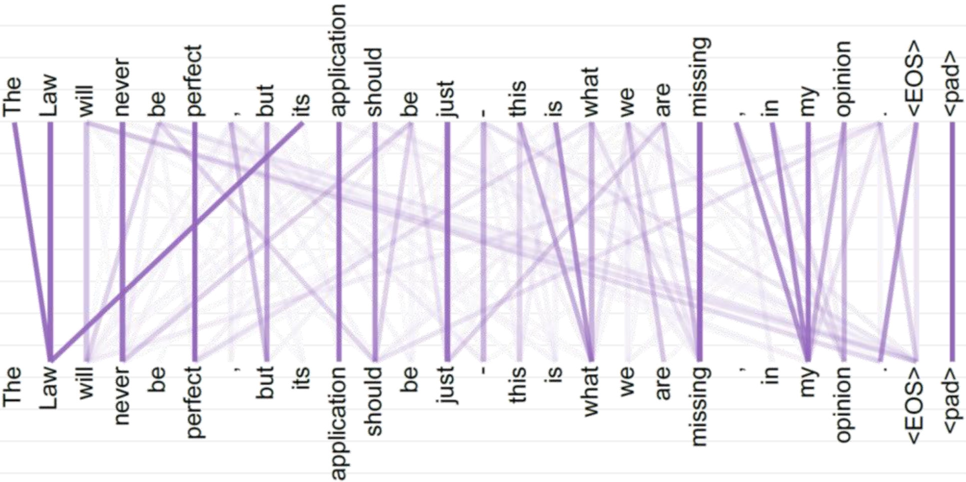
[Vaswani et al. NIPS17]

“Attention is all you need”



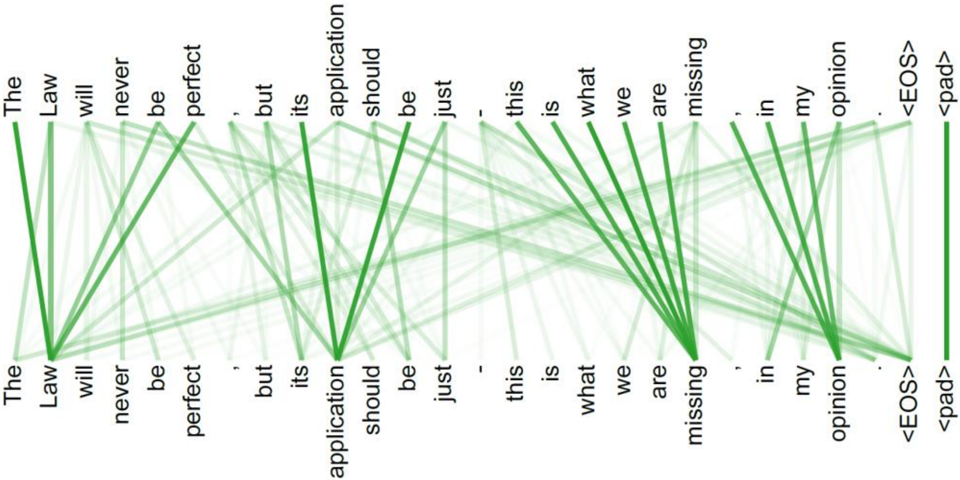
[Vaswani et al. NIPS17]

“Attention is all you need”



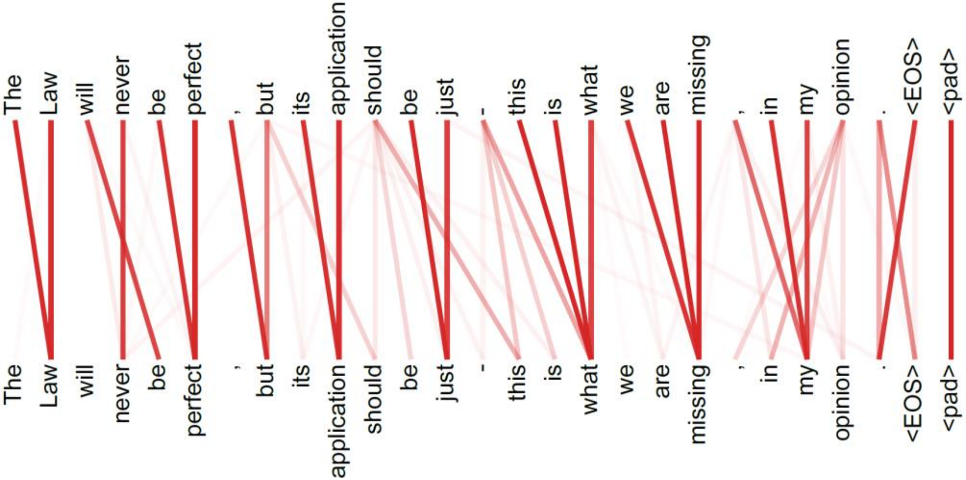
[Vaswani et al. NIPS17]

“Attention is all you need”



[Vaswani et al. NIPS17]

"Attention is all you need"



[Vaswani et al. NIPS17]

"Attention is all you need"

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

seq. of dilated conv

- Key property: attention layer complexity scales linearly in d (in attention layers) – same as depthwise separable convolution

[Vaswani et al. NIPS17]

"Attention is all you need"

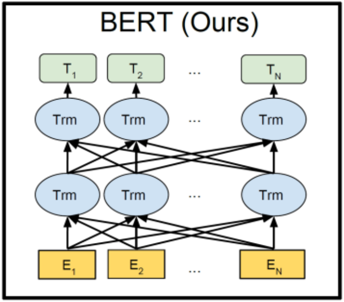
Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations
Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

[Vaswani et al. NIPS17]

BERT: pretraining by language modeling

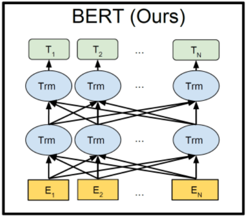
[Devlin et al. 2018]



- Pretraining six-level attention on language prediction tasks
- During training mask some words and replace others (~denoising autoenc idea)
- For sentence-level tasks, use the final layer representation of the first word
- **Fine-tune everything on the target task**

BERT: pretraining by language modeling

[Devlin et al. 2018]



System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); $\text{BERT}_{\text{BASE}}$ = (L=12, H=768, A=12); $\text{BERT}_{\text{LARGE}}$ = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from https://gluebenchmark.com/leaderboard and https://blog.openai.com/language-unsupervised/.

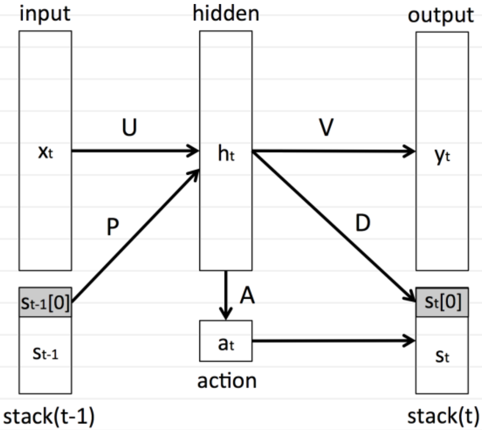
Stack augmented RNNs

- Inherent limitation of RNNs: memory capacity
- Increasing memory by n gives the increase of parameters by n^2
- **Conclusion:** we need to decouple memory and operations (thnik RAM and CPU!)

[Joulin and Mikolov NIPS15]

Stack augmented RNNs

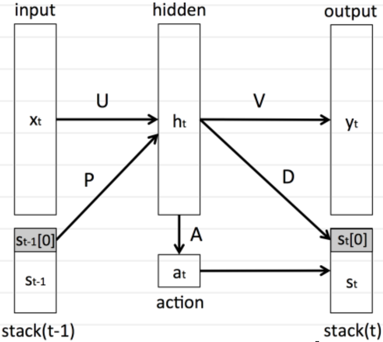
Conclusion: we need to decouple memory and operations (think RAM and CPU!)



[Joulin and Mikolov NIPS15]

Stack augmented RNNs

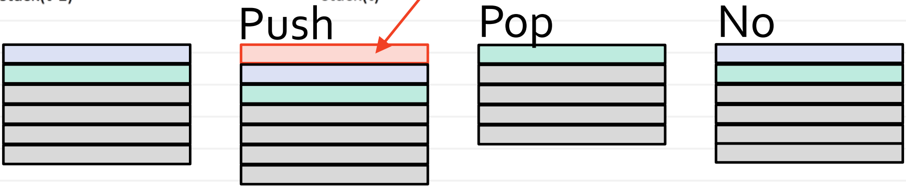
$$h_t = \sigma(Ux_t + Wh_{t-1} + Ps^k_{t-1})$$



$$y_t = \mathrm{SoftMax}(Vh_t)$$

$$a_t = \mathrm{SoftMax}(Ah_t)$$

$$\sigma(Dh_t)$$



[Joulin and Mikolov NIPS15]

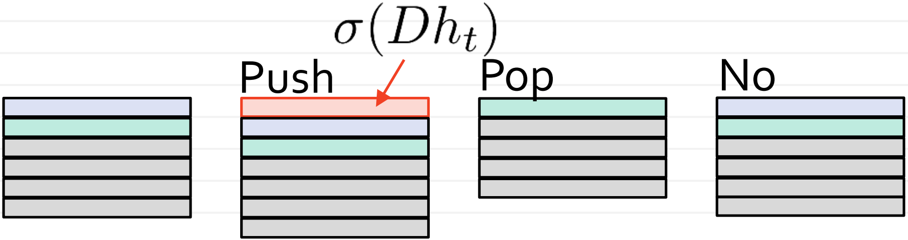
Stack augmented RNNs

$$a_t = \mathrm{SoftMax}(Ah_t)$$

Actions: Push, Pop, No

$$s_t^0 = a_t^{\mathrm{Push}}\sigma(Dh_t) + a_t^{\mathrm{Pop}} s_{t-1}^1 + a_t^{\mathrm{No}} s_{t-1}^0$$

$$s_t^i = a_t^{\mathrm{Push}} s_{t-1}^{i-1} + a_t^{\mathrm{Pop}} s_{t-1}^{i+1} + a_t^{\mathrm{No}} s_{t-1}^{i}$$



[Joulin and Mikolov NIPS15]

Binary addition with stack-RNN

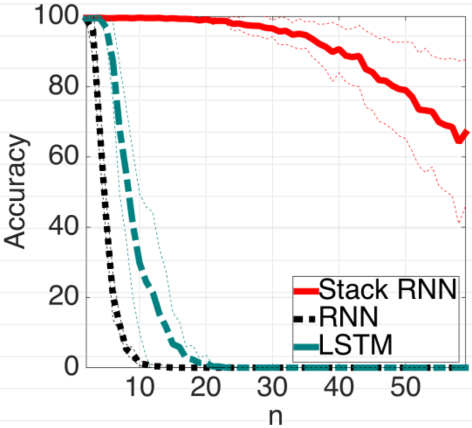
Goal: train a network that can add binary numbers.

Inputs:	.	1	0	0	0	1	1	+	1	1	1	0	=	1	0	0	0	1	1	.	
Predictions:	0	0	.	0	1	0	1	0	1	1	1	1	1	0	0	0	1	1	.	0	
Stack 1:	0							1					1							0	Counter
Stack 2:	1	-1											1			0				1	End of number 2
Stack 3:	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	Number 2
Stack 4:								1	0	0	0	0	0	0	0	1					Length of number 2
Stack 5:	0	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	Carry
Stack 6:			1	0	0	0	1	1					0	1	0	0	0	1	-1		Number 1
Stack 7:																					Junk
Stack 8:																					Junk
Stack 9:																					Junk
Stack 10:																					Junk

PUSH POP

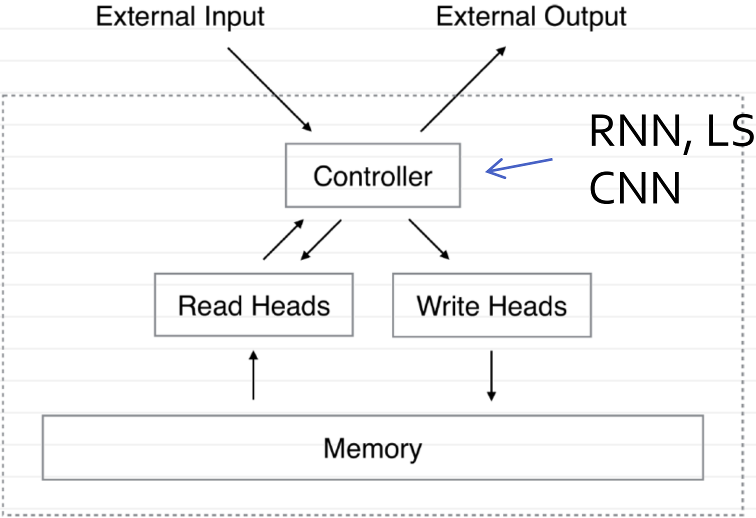
NB: the answer is reversed, i.e. 101+11 = 0001

Binary addition with stack-RNN



- Training with total lengths upto 20
- 100 hidden units and 10 1-dim stacks

Neural Turing Machine



RNN, LSTM, CNN

[Graves et al. 2014]

Outlook

- RNNs allow to solve many problems with sequences (as inputs or outputs)
- CTC-loss is useful for monotonically aligned input-output tasks
- The *attention* idea is working and is used across different domains (e.g. computer vision)
- Learning a computer to "program" is ambitious and promising
- Currently works only for simplistic algorithms
- Differentiability requires real-valued (soft) values
- Learning systems that make discrete choices is harder (but possible)

Bibliography

A. Graves. Supervised Sequence Labelling with Recurrent Neural Networks. Textbook, Studies in Computational Intelligence, Springer, 2012

Sepp Hochreiter, Jürgen Schmidhuber: Long Short-Term Memory. Neural Computation 9(8): 1735-1780 (1997)

Ilya Sutskever, Oriol Vinyals, Quoc V. Le:
Sequence to Sequence Learning with Neural Networks. NIPS 2014: 3104-3112

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, Kate Saenko:
Long-term recurrent convolutional networks for visual recognition and description. CVPR 2015: 2625-2634

Justin Johnson, Andrej Karpathy, Li Fei-Fei, DenseCap: Fully Convolutional Localization Networks for Dense Captioning. CVPR 2016

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, Fei-Fei Li: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. CoRR abs/1602.07332 (2016)

D. Bahdanau, K. Cho, and Y. Bengio: Neural machine translation by jointly learning to align and translate. In ICLR 2015.

Bibliography

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luís Marujo, Tiago Luís: Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. EMNLP 2015: 1520-1530

Minh-Thang Luong, Hieu Pham, Christopher D. Manning:
Effective Approaches to Attention-based Neural Machine Translation. CoRR abs/1508.04025 (2015)

Alex Graves, Santiago Fernández, Faustino J. Gomez, Jürgen Schmidhuber:
Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. ICML 2006: 369-376

Armand Joulin, Tomas Mikolov:
Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets. NIPS 2015: 190-198

Alex Graves, Greg Wayne, Ivo Danihelka:
Neural Turing Machines. CoRR abs/1410.5401 (2014)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin:
Attention is All you Need. NIPS 2017: 6000-6010

Bibliography

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer:
Deep Contextualized Word Representations. NAACL-HLT 2018: 2227-2237

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova:
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Arxiv 18