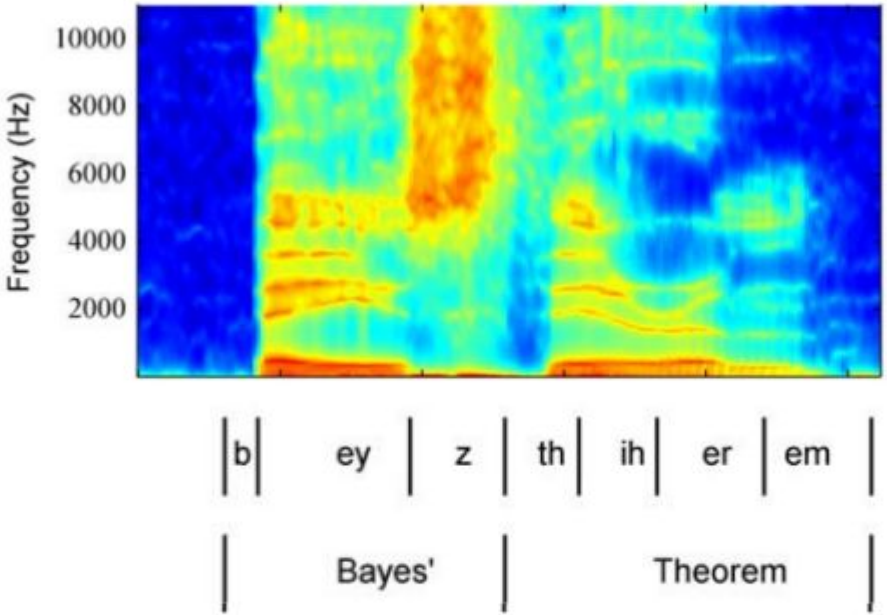
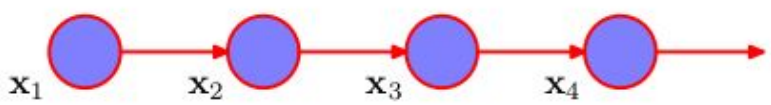
Sequence modeling

Ville Hautamäki

Modeling a time-dependent signal

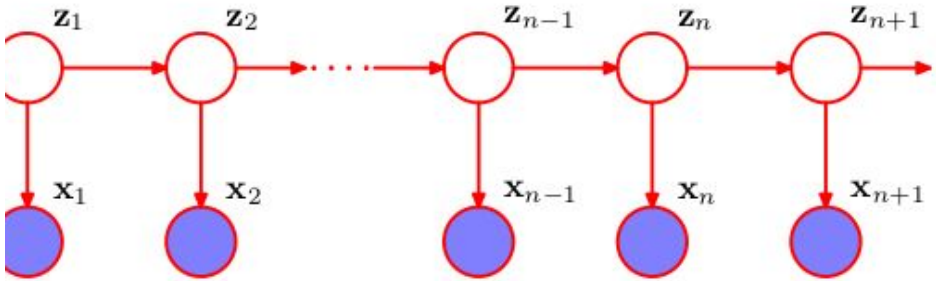


First order Markov

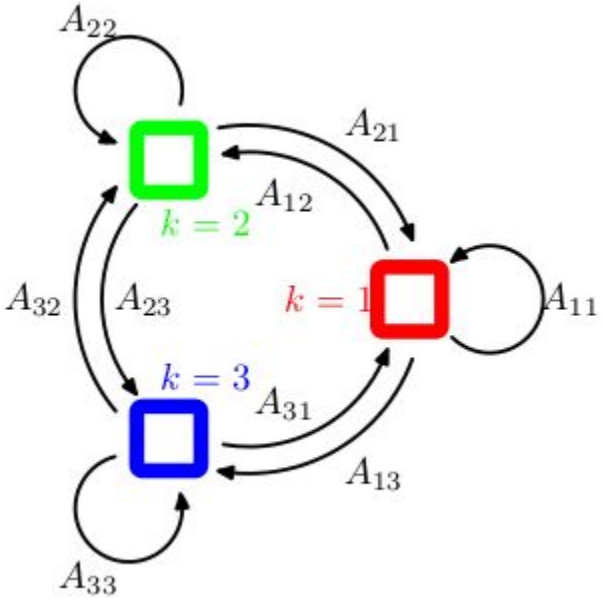


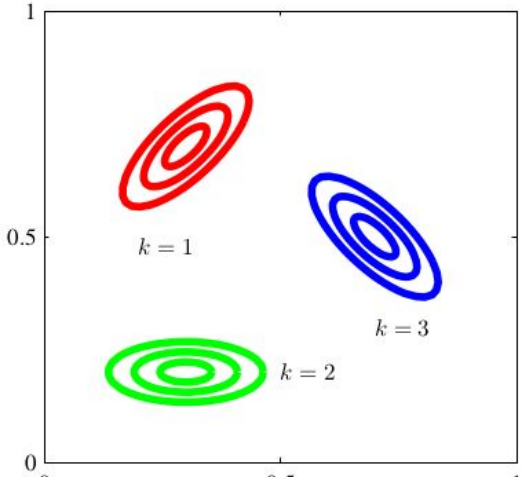
$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n | \mathbf{x}_{n-1})$$

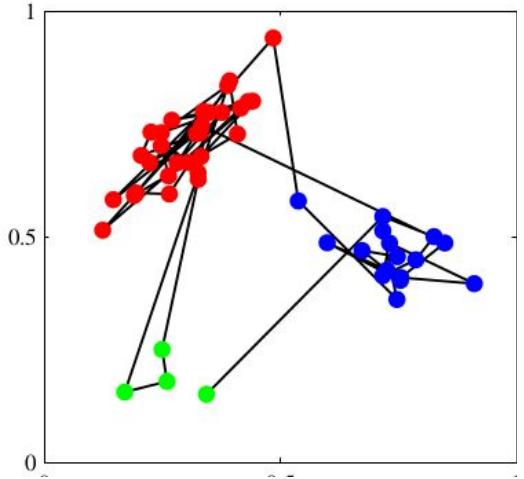
Hidden Markov model



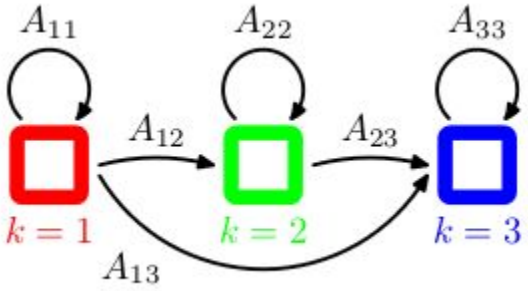
Hidden Markov model (HMM)

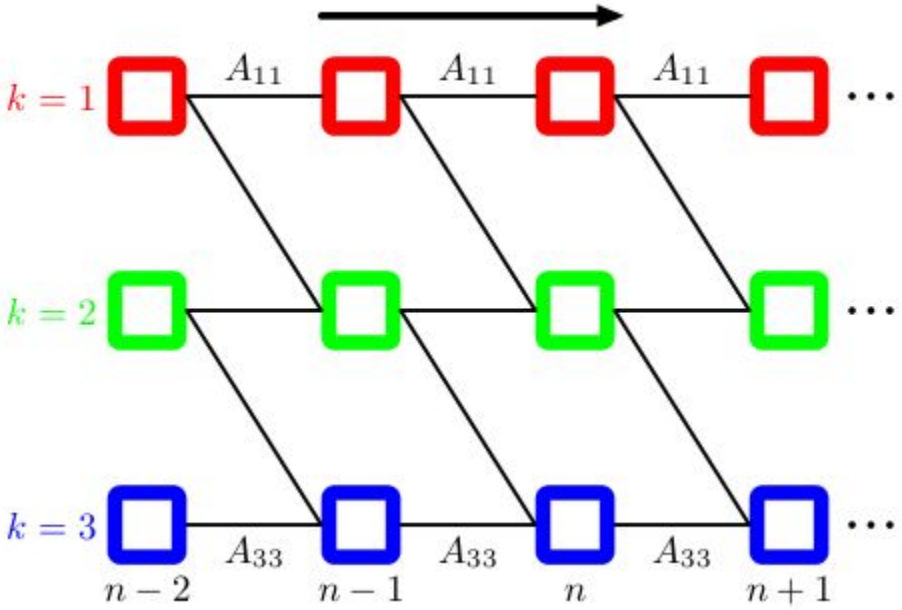






Left-to-right HMM

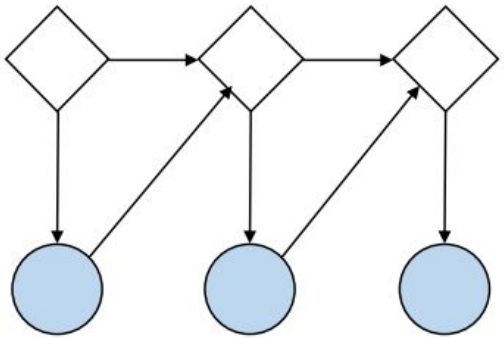




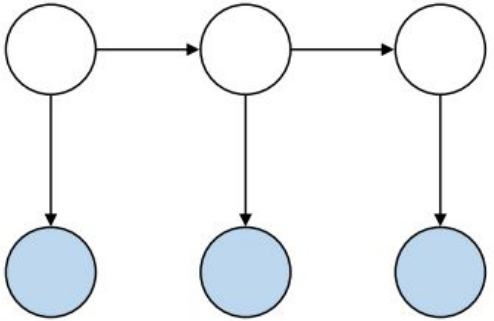
One way to use HMM for recognition

Can train an HMM to classify a sequence:

1. train a separate HMM per class
2. evaluate prob. of unlabelled sequence under each HMM
3. classify: HMM with highest likelihood

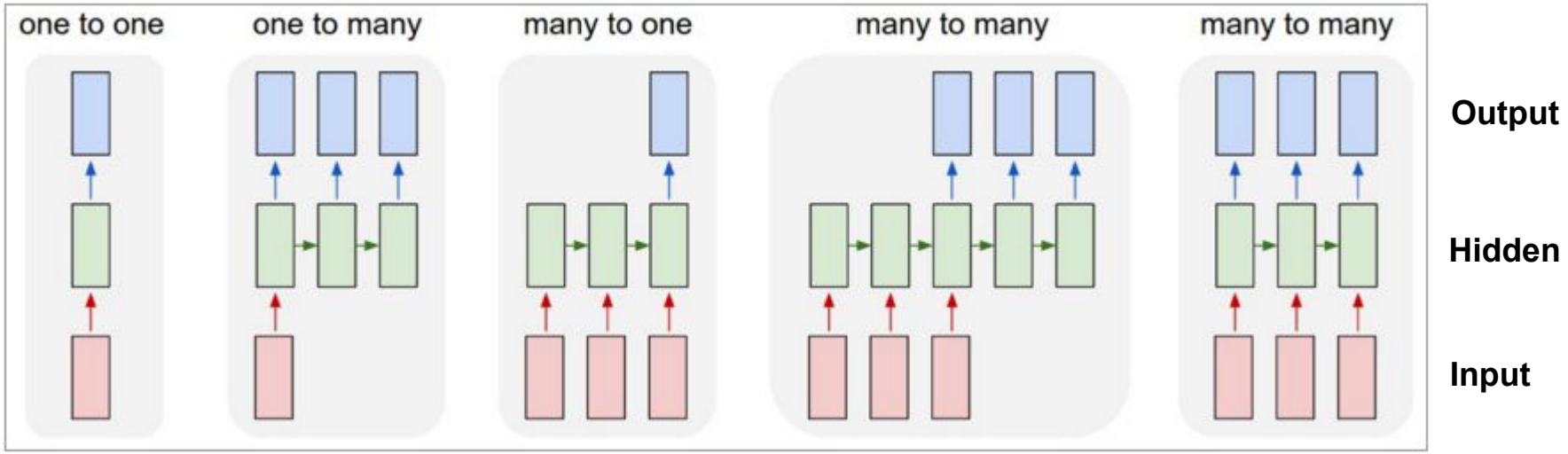


Neural model

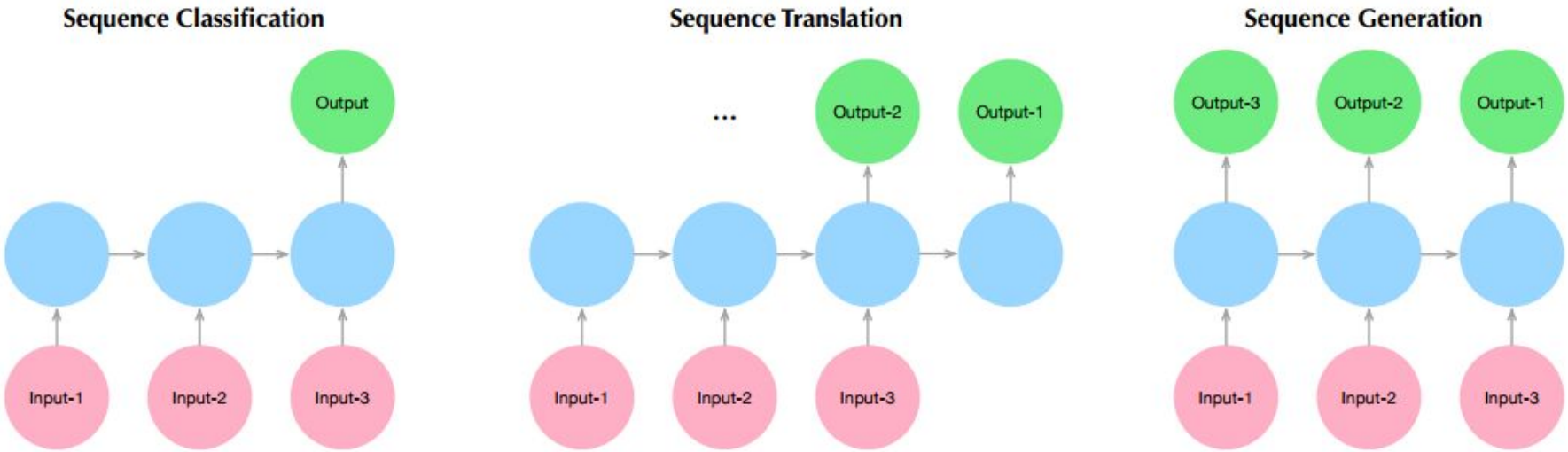


Classical probabilistic

Recurrent neural network (RNN) is flexible



Three basic tasks that RNN can do

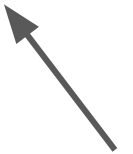


Core RNN equation

$$\mathbf{h}_t = F(\mathbf{x}_t, \mathbf{h}_{t-1}, \theta)$$

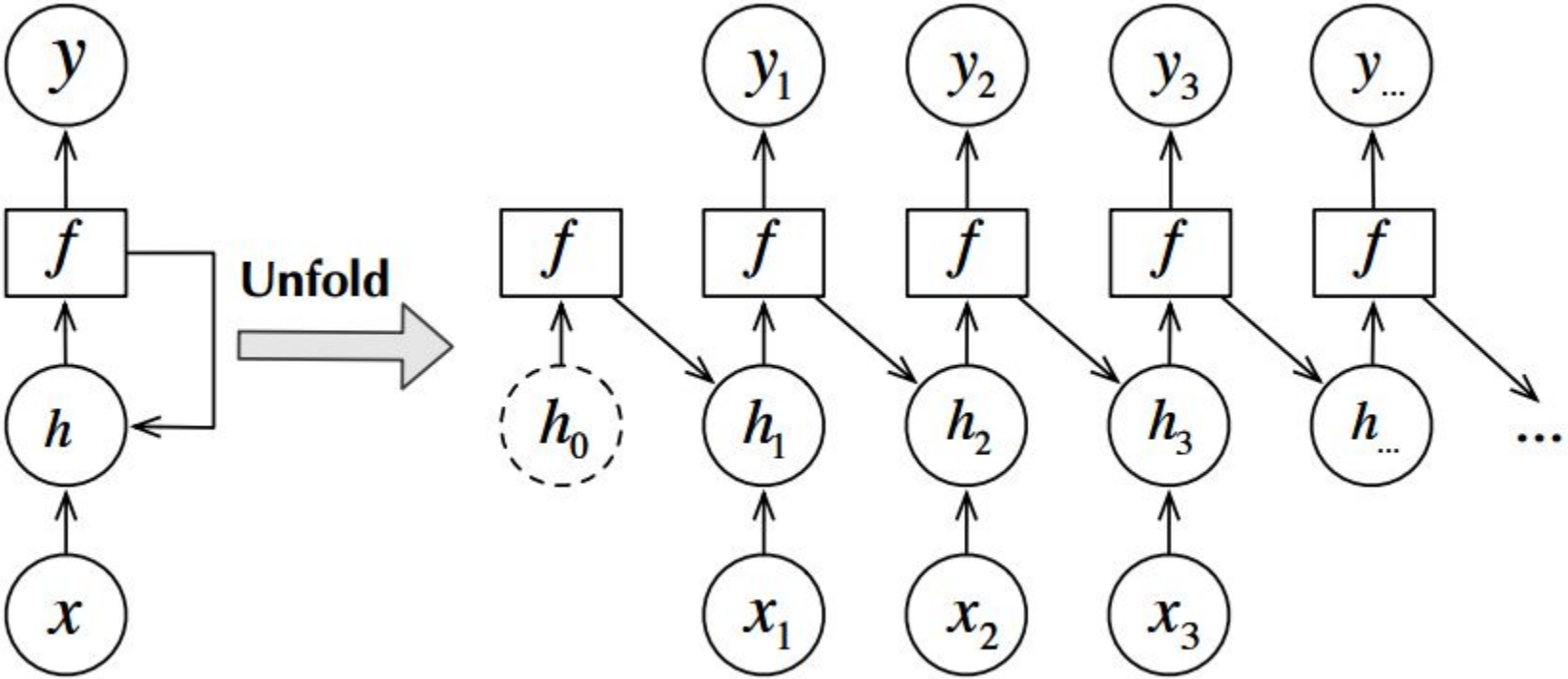
To get t:th hidden state from current observation and hidden state from the previous time step.

$$\mathbf{h}_t = f_a(\theta_{\mathbf{x}} \cdot \mathbf{x}_t + \theta_{\mathbf{h}} \cdot \mathbf{h}_{t-1} + b)$$

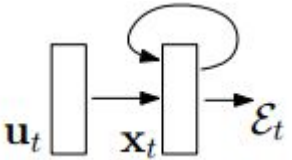


Activation function

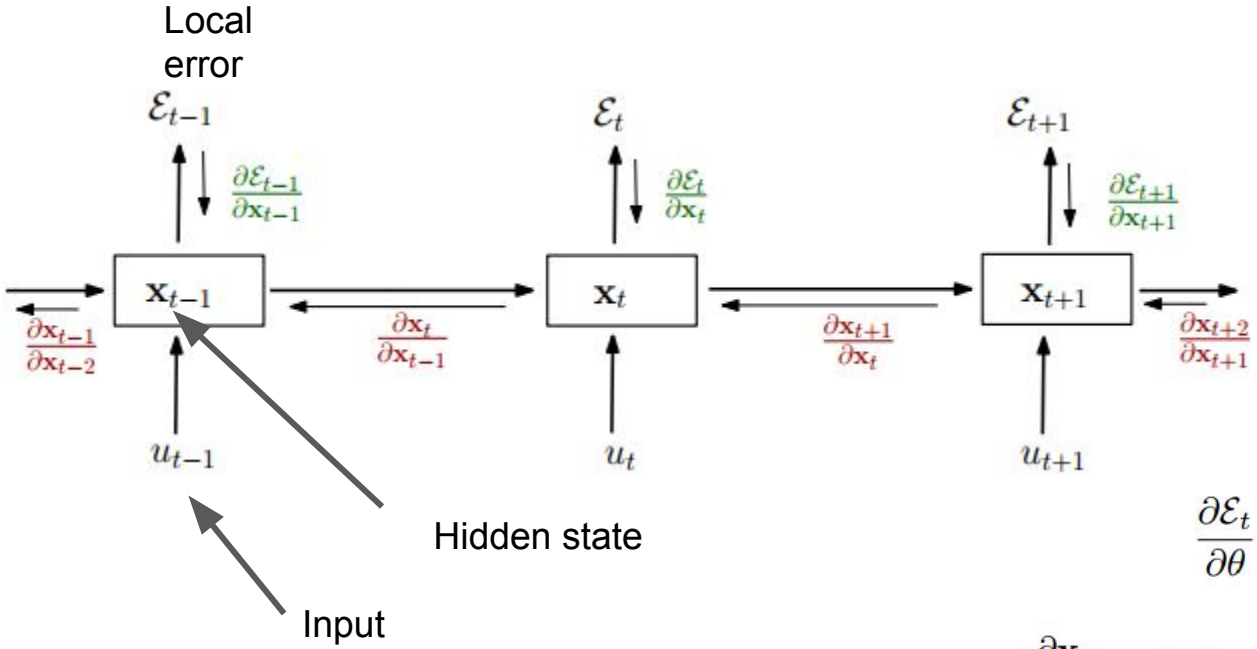
Unrolling the time-steps



Training RNN is difficult



$$\mathcal{E} = \sum_{1 \le t \le T} \mathcal{E}_t$$



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \tag{3}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \tag{4}$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \tag{5}$$

Long-term correlations are not modeled!

$$\left\| \frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \left(\prod_{i=k}^{t-1} \frac{\partial \mathbf{x}_{i+1}}{\partial \mathbf{x}_i} \right) \right\| \leq \eta^{t-k} \left\| \frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \right\|$$

- Gradient either explodes or vanishes
- For explosion, practical trick is gradient clipping:

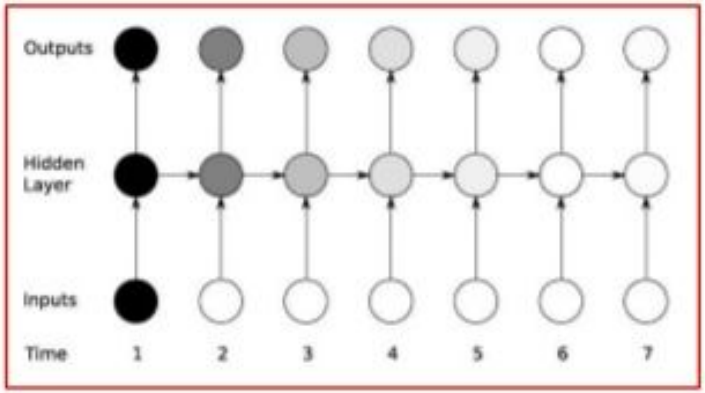
Algorithm 1 Pseudo-code for norm clipping

$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|} \hat{\mathbf{g}}$
end if

LSTM is the solution to gradient vanishing

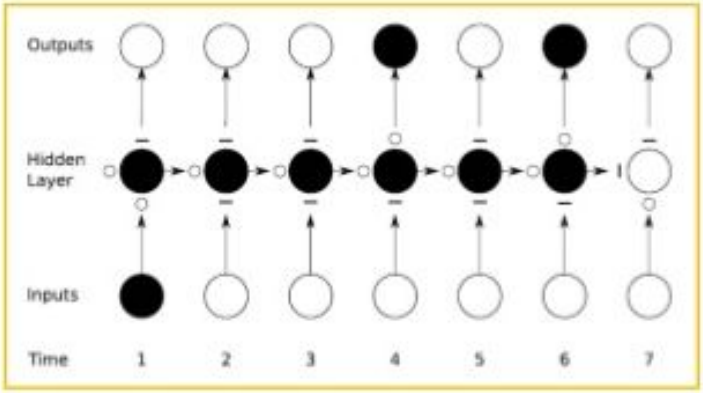
Conventional RNN with sigmoid

- The sensitivity of the input values decays over time
- The network forgets the previous input

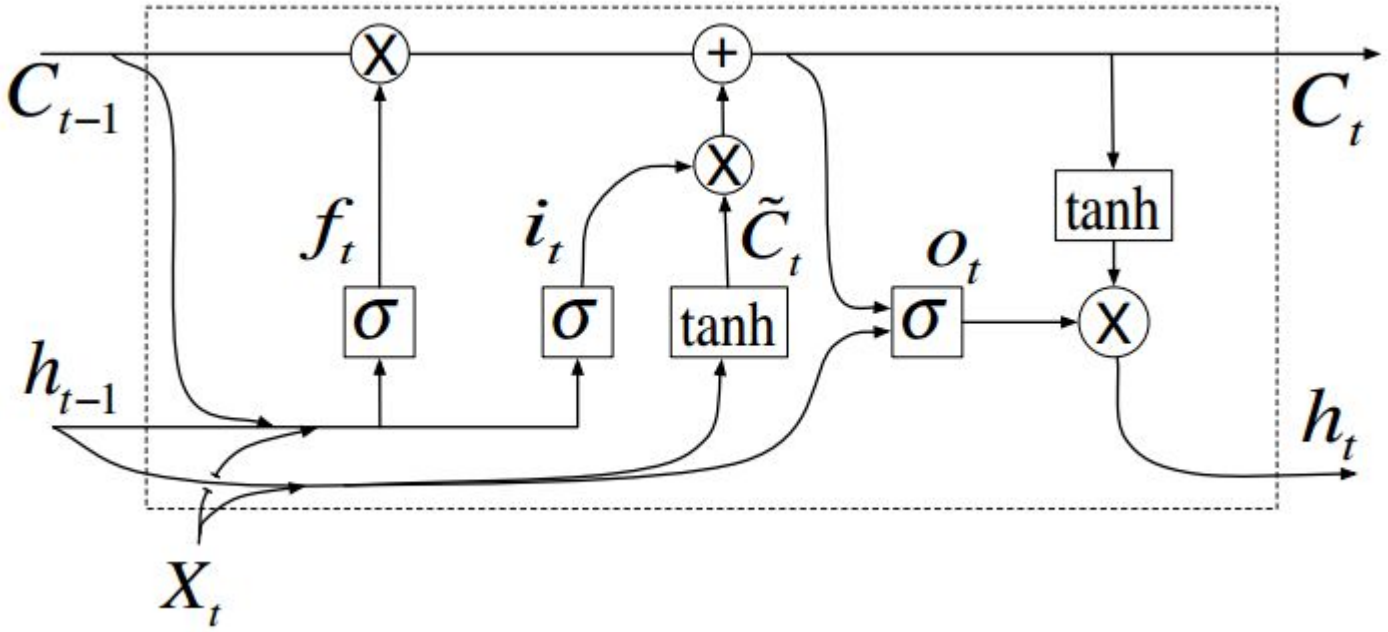


Long-Short Term Memory (LSTM) [2]

- The cell remember the input as long as it wants
- The output can be used anytime it wants



The LSTM model



LSTM equations

Main idea! Notice summation.

$$\mathbf{i}_t = \sigma_i(\mathbf{x}_t\mathbf{W}_{xi} + \mathbf{h}_{t-1}\mathbf{W}_{hi} + \mathbf{w}_{ci} \odot \mathbf{c}_{t-1} + \mathbf{b}_i),$$

$$\mathbf{f}_t = \sigma_f(\mathbf{x}_t\mathbf{W}_{xf} + \mathbf{h}_{t-1}\mathbf{W}_{hf} + \mathbf{w}_{cf} \odot \mathbf{c}_{t-1} + \mathbf{b}_f),$$

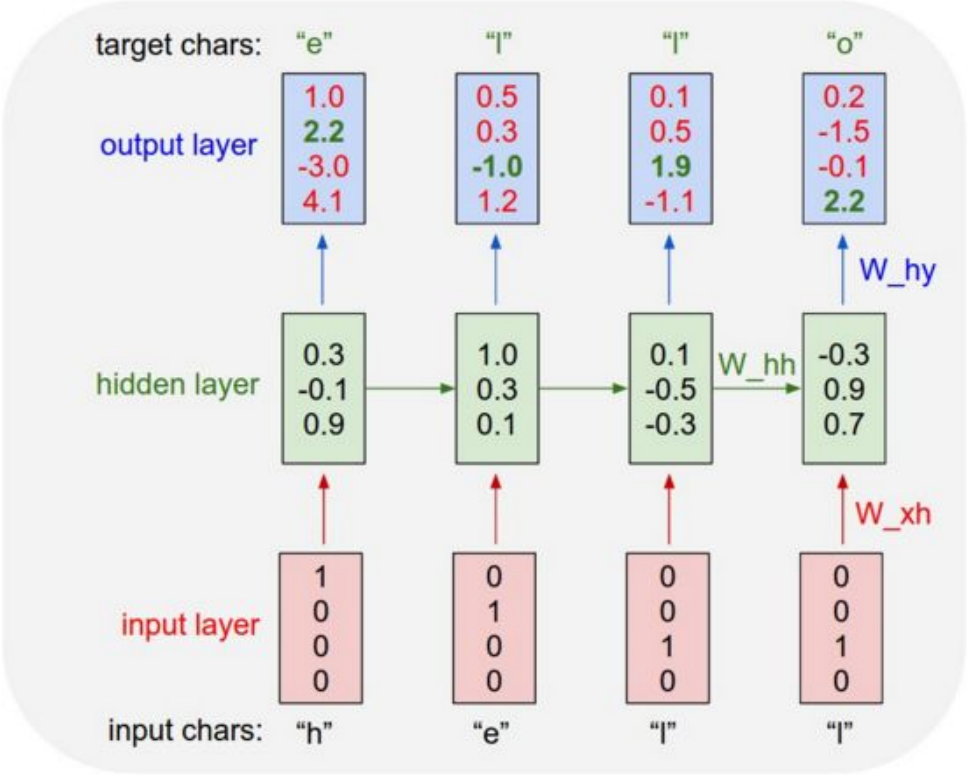
$$\tilde{c}_t = tanh(\mathbf{x}_t\mathbf{W}_{xc} + \mathbf{h}_{t-1}\mathbf{W}_{hc} + \mathbf{b}_c),$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{c}_t,$$

$$\mathbf{o}_t = \sigma_o(\mathbf{x}_t\mathbf{W}_{xo} + \mathbf{h}_{t-1}\mathbf{W}_{ho} + \mathbf{w}_{co} \odot \mathbf{c}_t + \mathbf{b}_o),$$

$$\mathbf{h}_t = \mathbf{o}_t \odot tanh(\mathbf{c}_t),$$

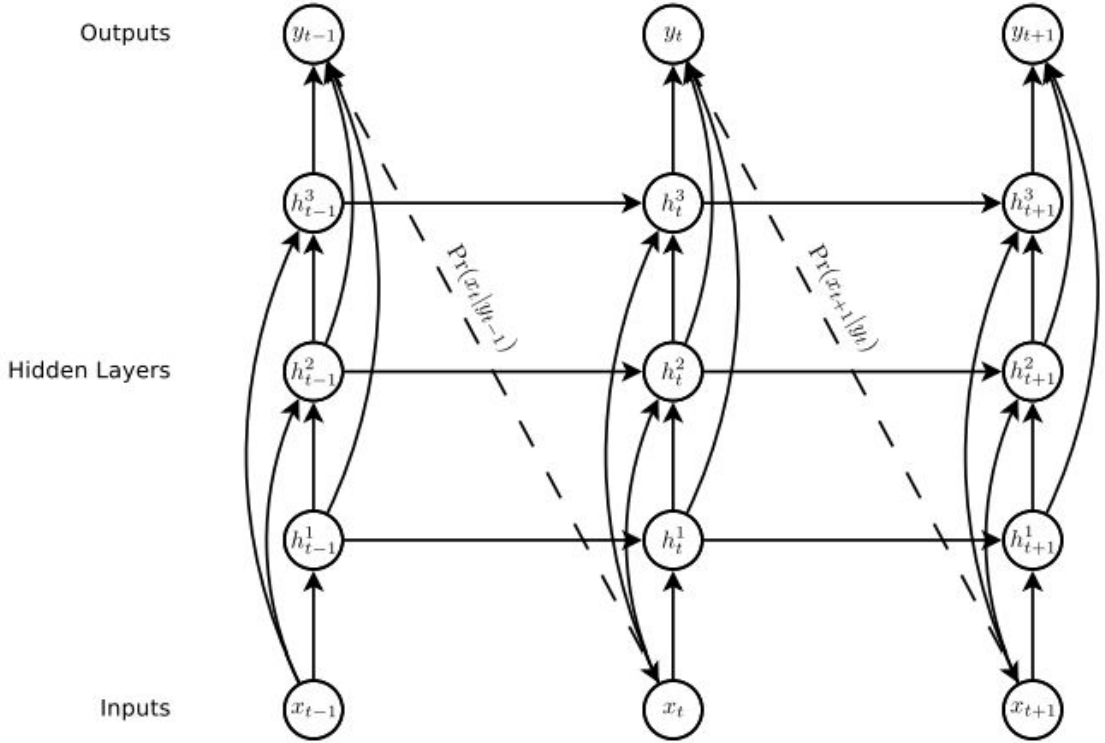
Example: Predicting next char with RNN / LSTM



How about hallucinating new text char by char

- Encode each character as one-hot vector
- Train RNN / LSTM using your favourite corpus to predict the next character.
- Test time, input one character and predict new one.
- Feed predicted character as a new input. Repeat.
- Use special STOP char to detect when to stop generating.

Example: Generating speech

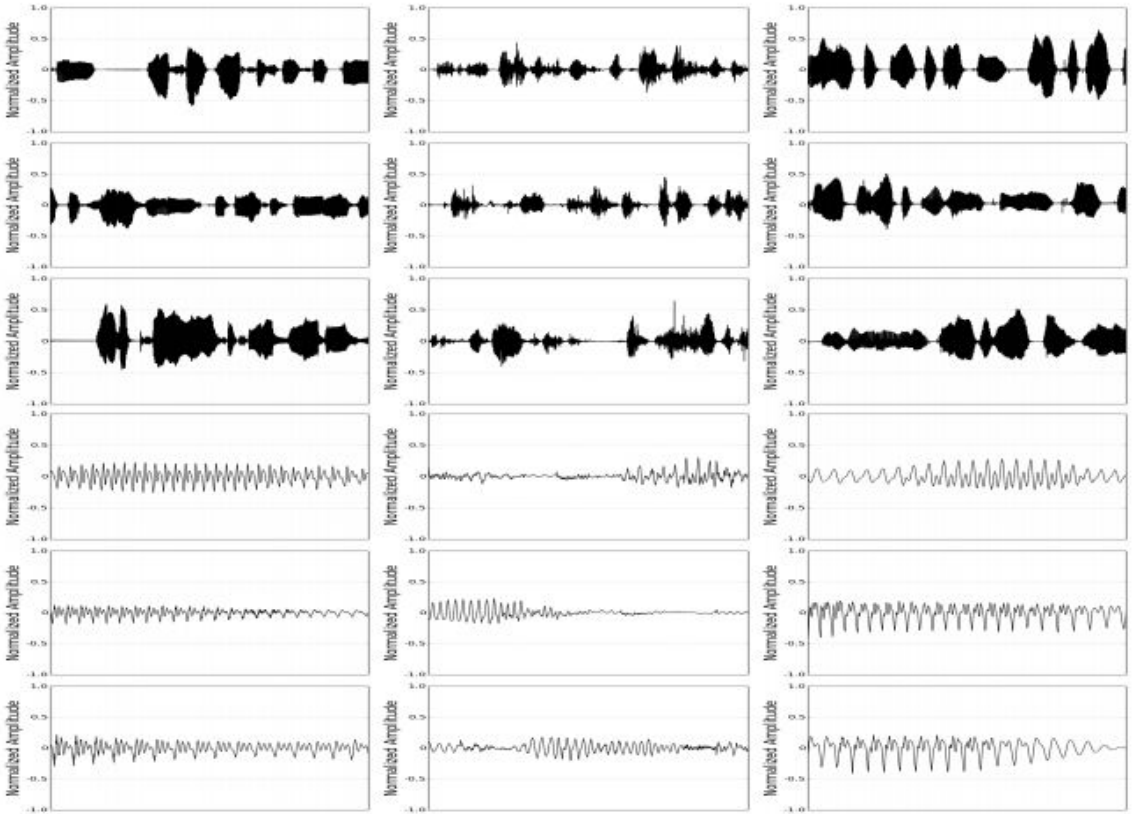


$$\Pr(\mathbf{x}) = \prod_{t=1}^{T} \Pr(x_{t+1}|y_t)$$

$$\mathcal{L}(\mathbf{x}) = -\sum_{t=1} \log \Pr(x_{t+1}|y_t)$$

This is the discrete case (such as chars.) for real valued case replace Pr() with GMM. So RNN is going to predict GMM parameters which will then generate MFCC vectors and voiced/unvoiced decision per time step. This is called mixture density network (MDM), for example Edward has an implementation.

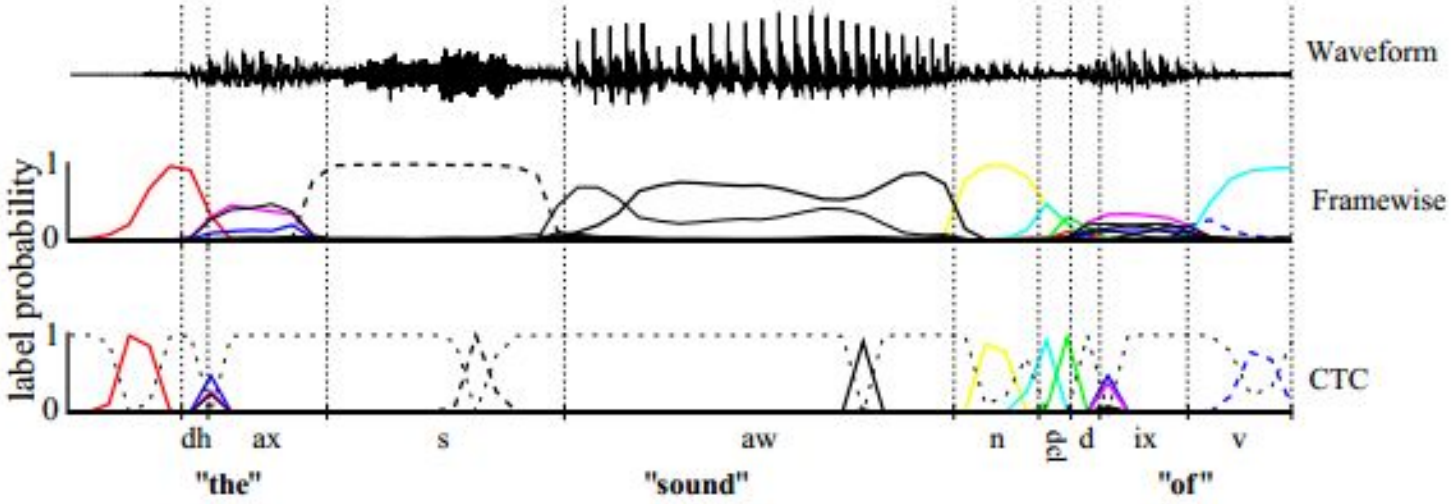
What the results look like



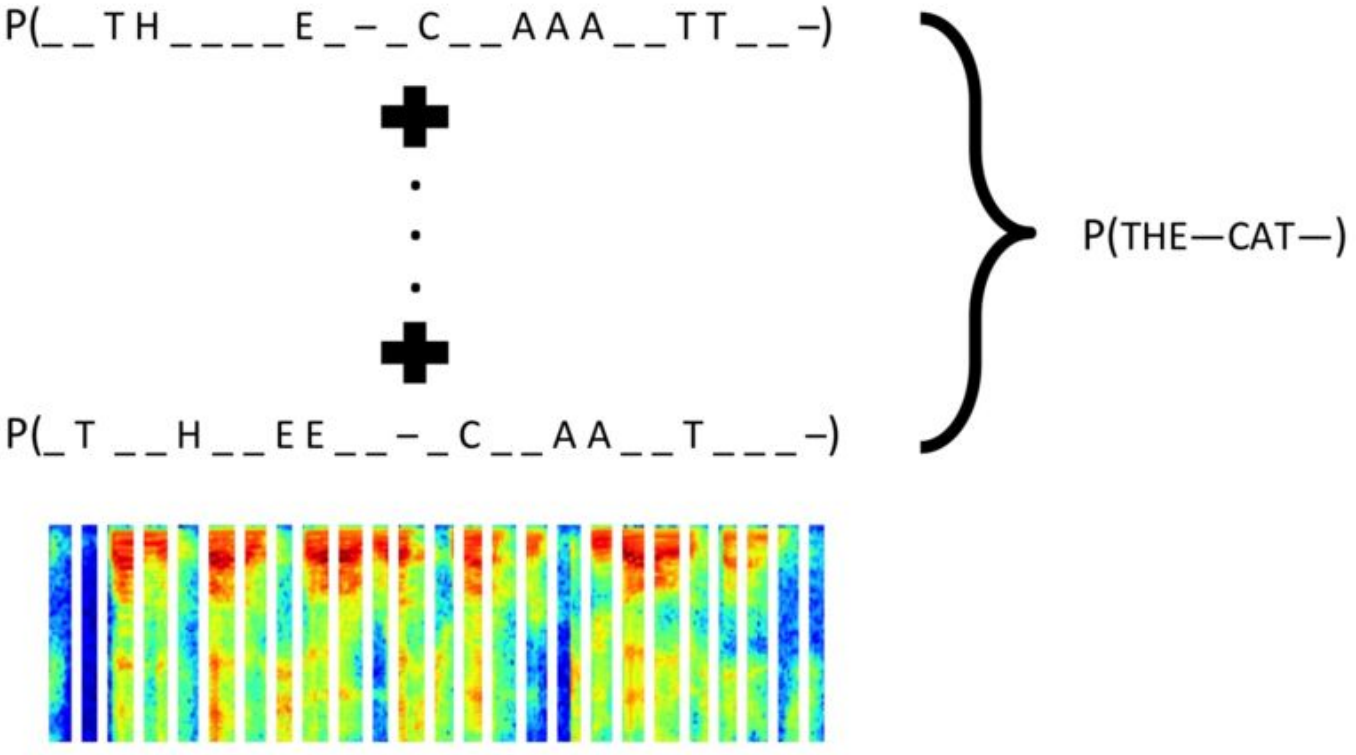
(a) Ground Truth (b) RNN-GMM (c) VRNN-Gauss

Example: Use LSTM to map features to words

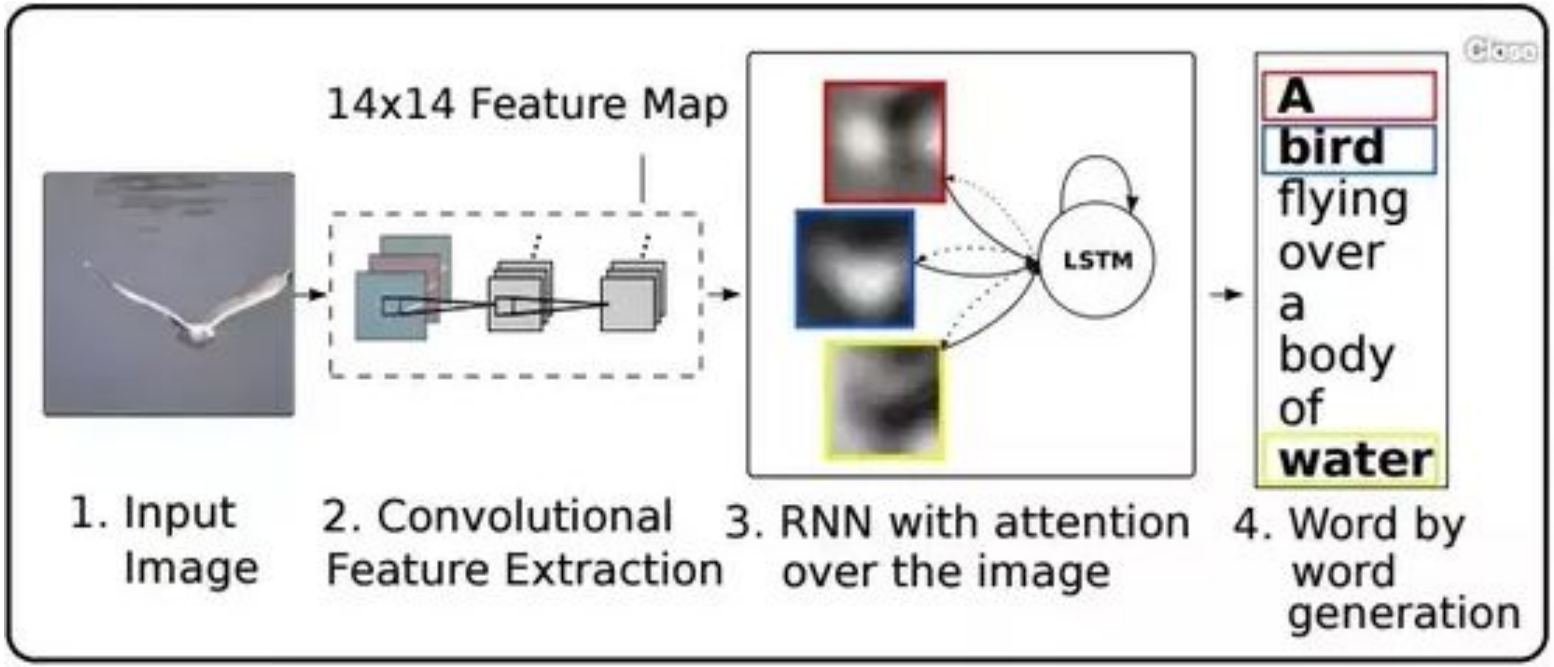
- Technique is called connectionist temporal classification (CTC).
- Allows end-to-end speech recognition.
- Reduces human effort in developing ASR system for new language.



CTC does not require knowledge of the alignments



LSTM is so 90’s, attention modeling is modern stuff



Attention is a way for network to focus

by *ent423* , *ent261* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* , *ent85* , near *ent312* , a *ent119* official told *ent261* on wednesday . he was identified thursday as special warfare operator 3rd class *ent23* , 29 , of *ent187* , *ent265* . `` *ent23* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused

. . .

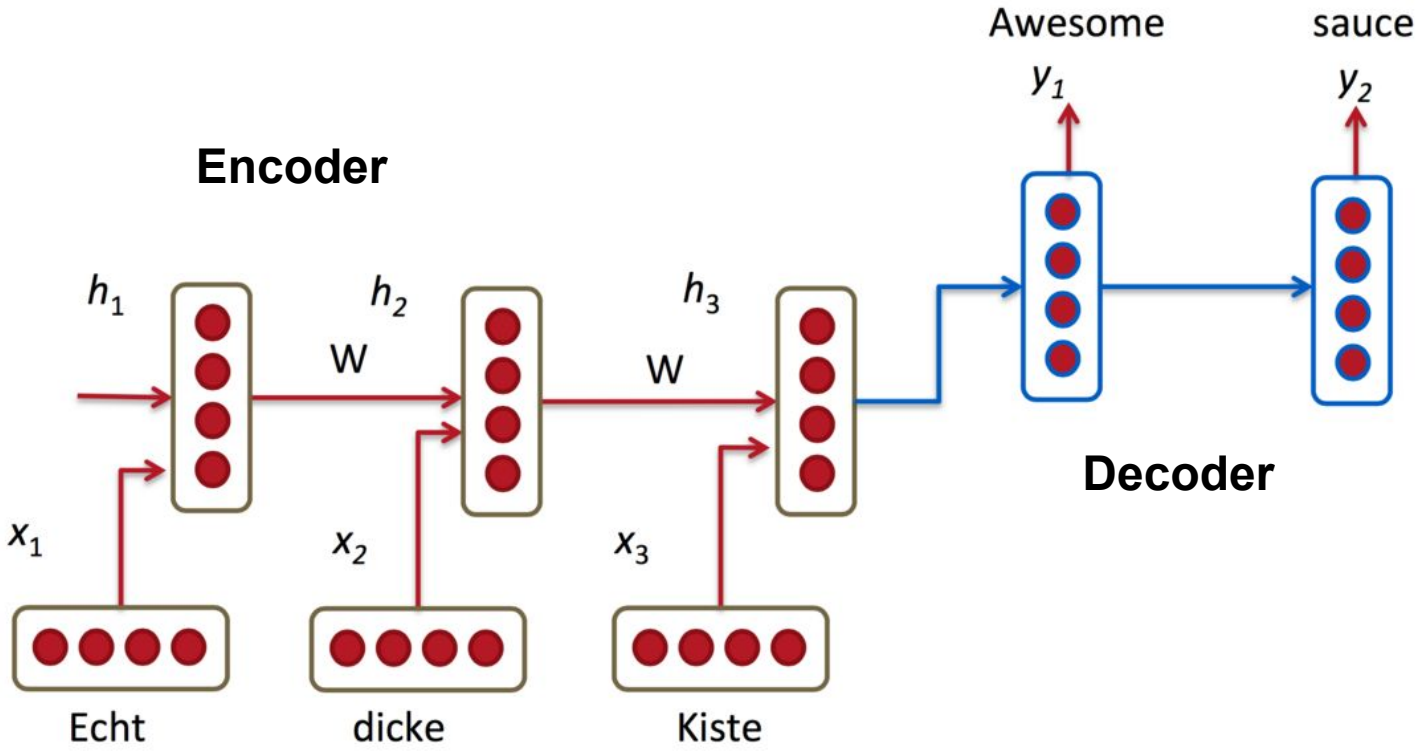
ent119 identifies deceased sailor as **X** , who leaves behind a wife

by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

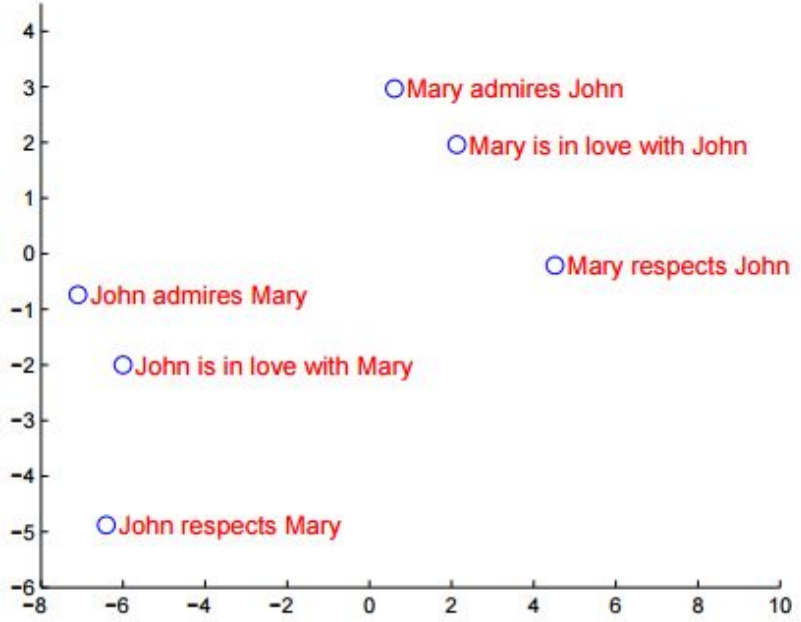
. . .

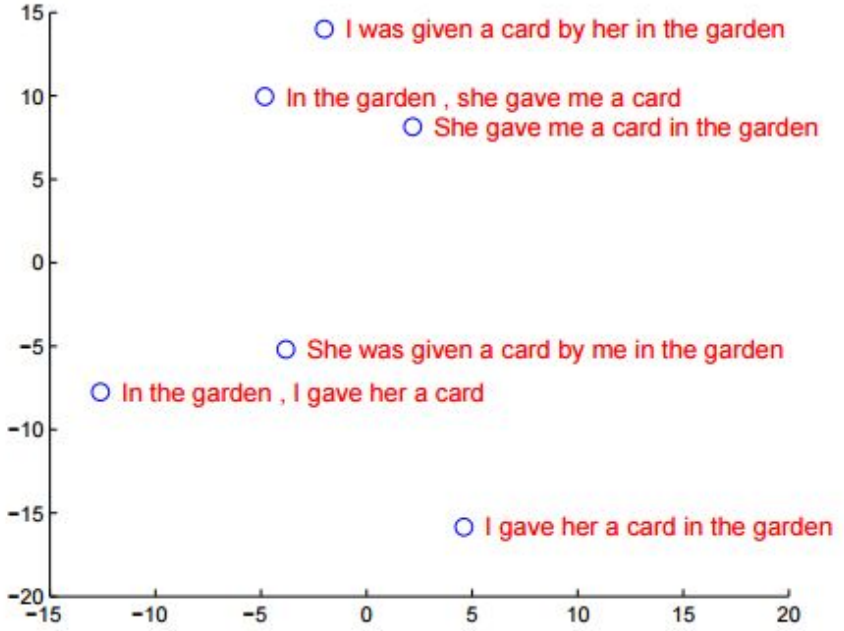
X dedicated their fall fashion show to moms

Example: Neural machine translation (RNN)

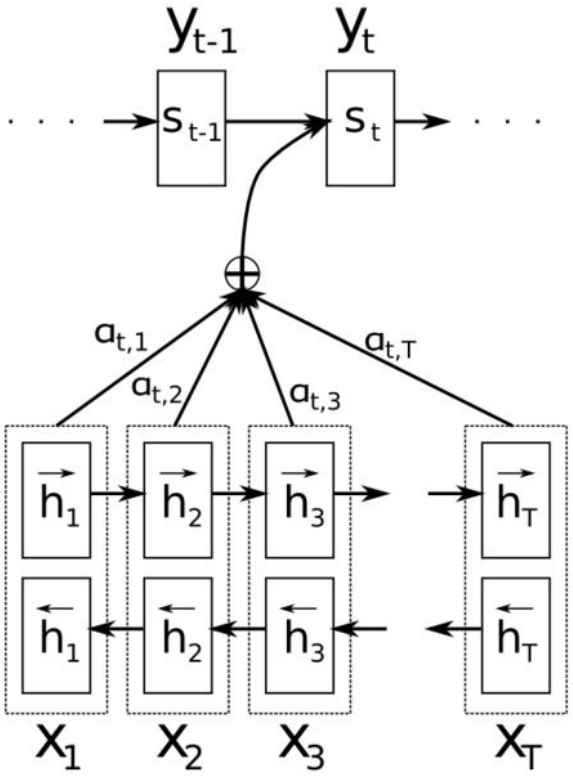


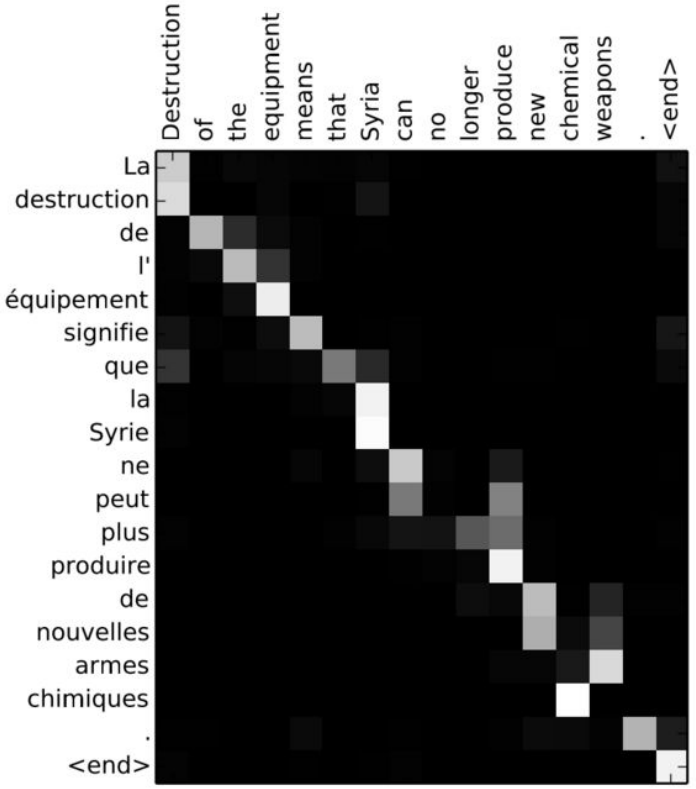
Final hidden state encodes the *whole* source sentence



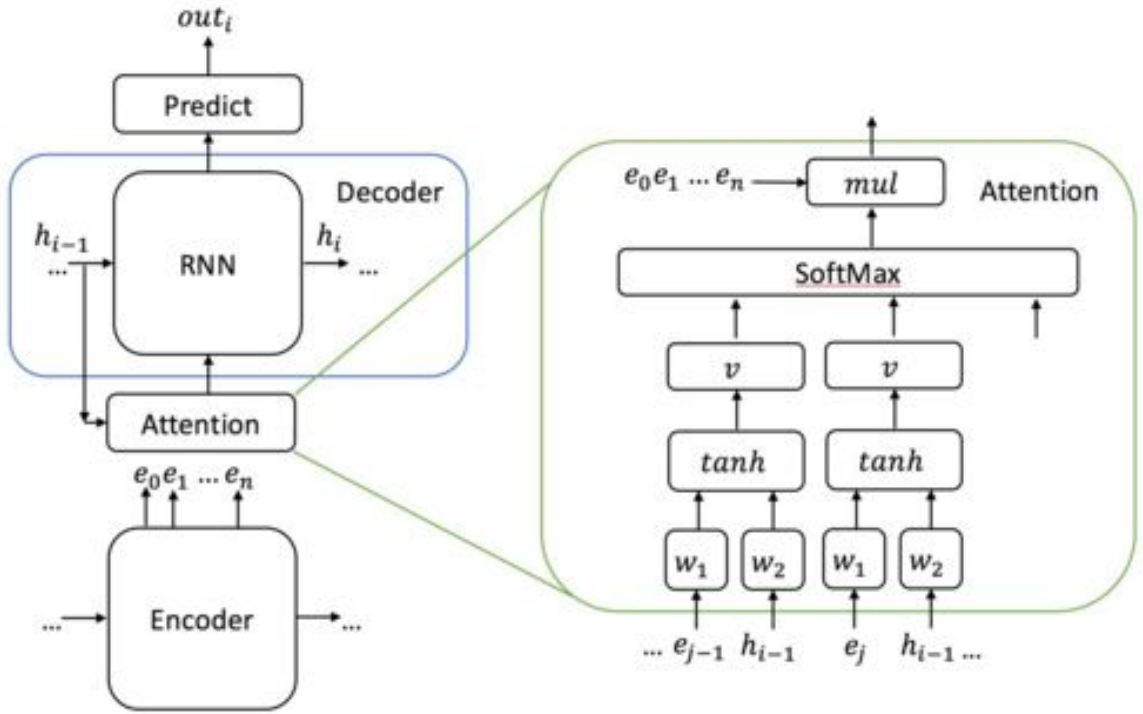


But it is better if all source hidden states affect target





More details about attention mechanism



Example: neural summarization

Source (First Sentence)

Russian Defense Minister Ivanov called Sunday for the creation of a joint front for combating global terrorism.

Target (Title)

Russia calls for joint front against terrorism.

Example: Grammar correction

Source (Original Sentence)

There is no a doubt, tracking systems has brought many benefits in this information age .

Target (Corrected Sentence)

There is no doubt, tracking systems have brought many benefits in this information age .