

# Deep learning summer camp

## 4 Recurrent neural networks

Ole Winther

Dept for Applied Mathematics and Computer Science  
Technical University of Denmark (DTU)



July 4, 2016

# Objectives of talk

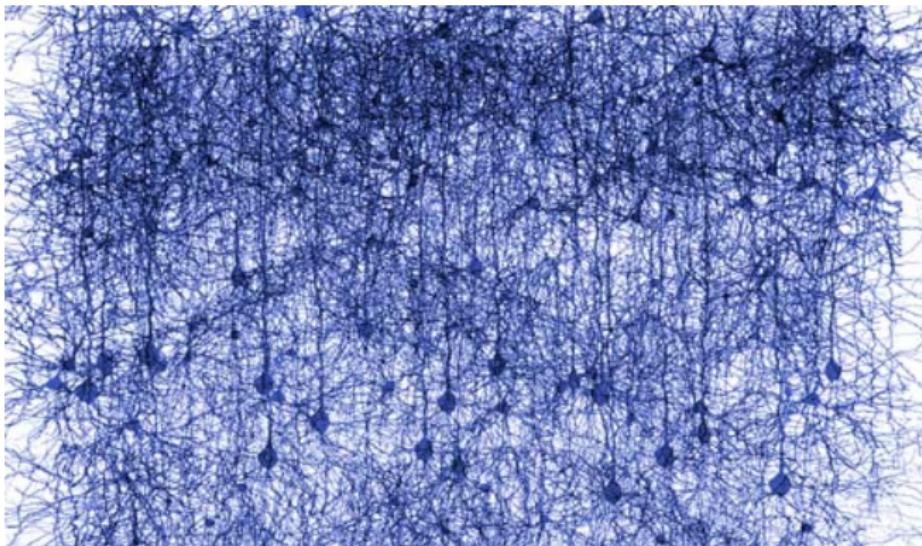
- Recurrent neural networks
- what they can be used for and
- how they are trained.
- Two example from my group's research:
  - Information retrieval
  - Biological sequence analysis
- Recap supervised learning



# Part 1: Recurrent neural networks

# Neural networks (NNs)

- Feedforward neural networks (FFNNs)
- Convolutional neural networks (CNNs)
- Recurrent Neural Networks (RNNs)
- Auto-encoders (AE)

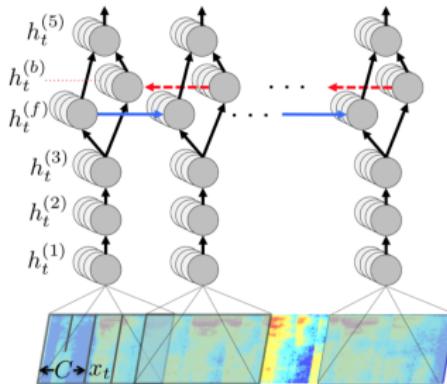


# Recurrent neural networks – DeepSpeech

## DeepSpeech: Scaling up end-to-end speech recognition

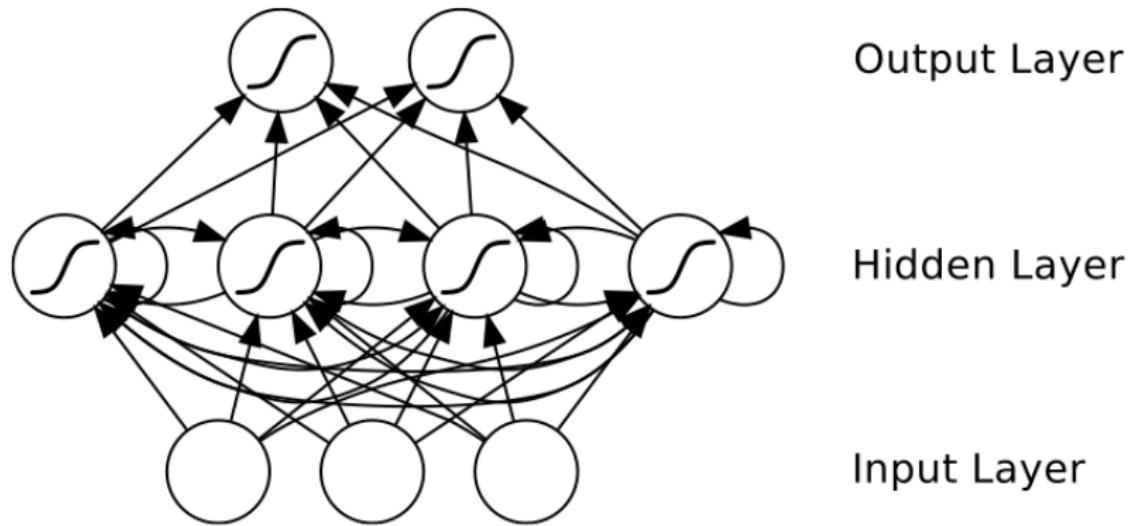
Awni Hannun\*, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen,  
Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng

Baidu Research – Silicon Valley AI Lab



Deep speech 2: ~human level performance + realtime server

# Recurrent neural networks

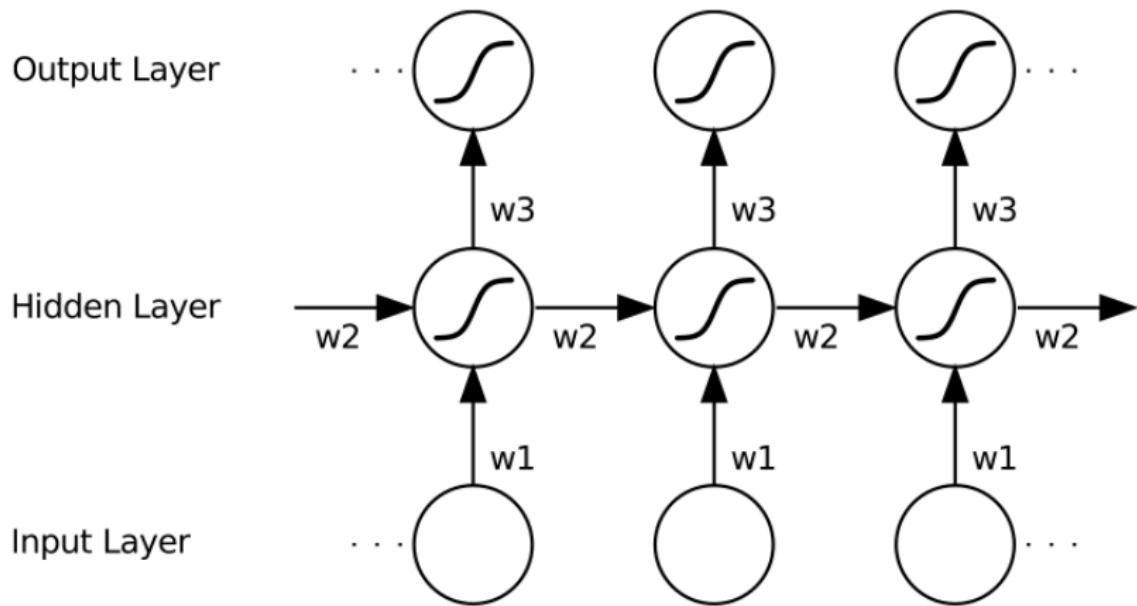


$$\mathbf{h}_t^{(1)} = f_1 \left( W^{(1)} \mathbf{x}_t + W^{(\rightarrow)} \mathbf{h}_{t-1}^{(1)} \right)$$

$$\mathbf{h}_t^{(2)} = f_2 \left( W^{(2)} \mathbf{h}_t^{(1)} \right)$$

[www.cs.toronto.edu/~graves/preprint.pdf](http://www.cs.toronto.edu/~graves/preprint.pdf)

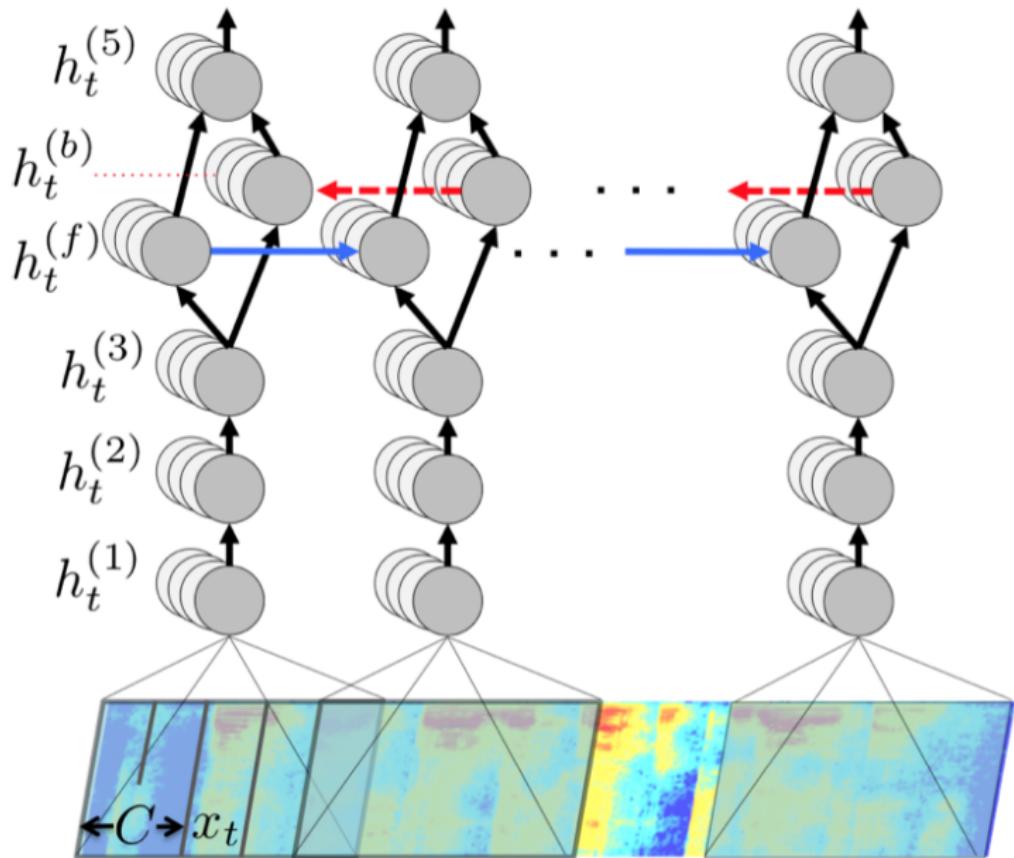
# Recurrent neural networks unrolled



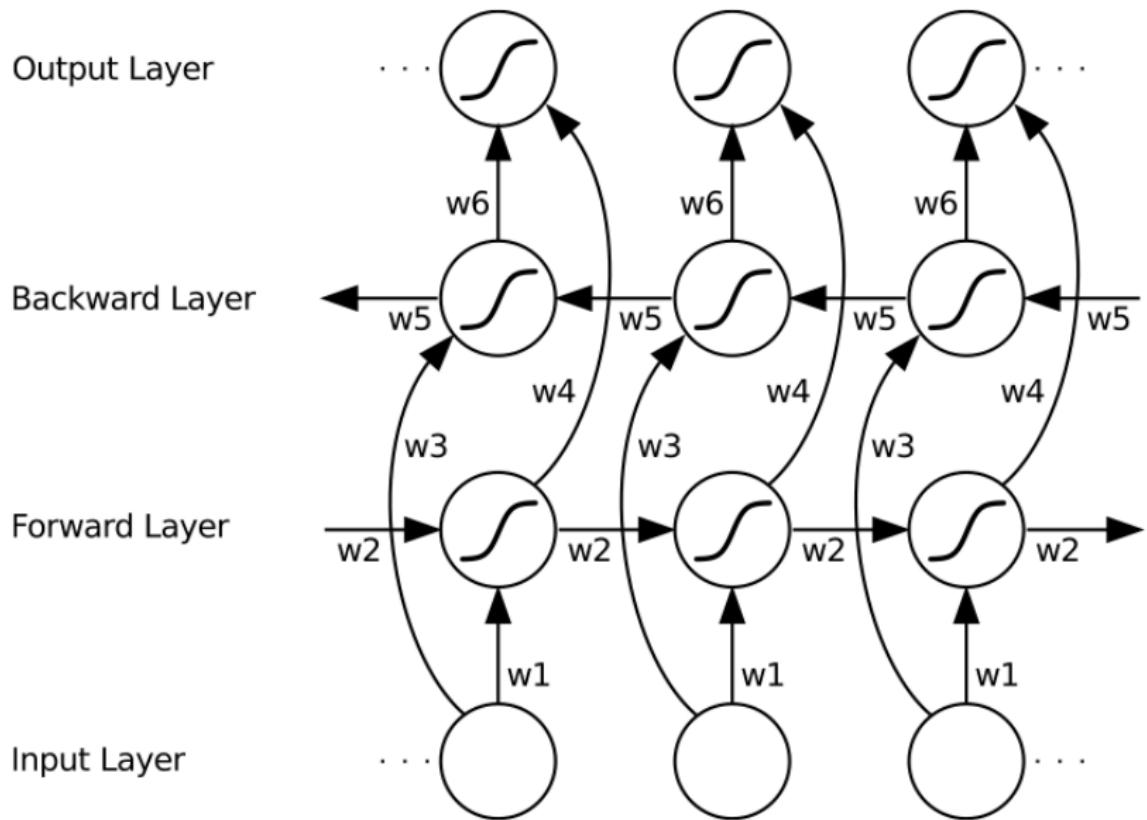
$$\mathbf{h}_t^{(1)} = f_1 \left( W^{(1)} \mathbf{x}_t + W^{(\rightarrow)} \mathbf{h}_{t-1}^{(1)} \right)$$

$$\mathbf{h}_t^{(2)} = f_2 \left( W^{(2)} \mathbf{h}_t^{(1)} \right)$$

# Bidirectional recurrent neural networks



# Recurrent neural networks unrolled bidirectional



# Encoder-decoder - machine translation

---

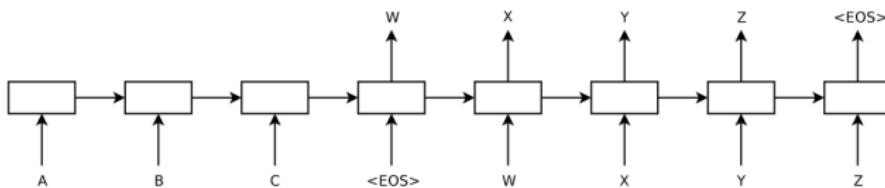
## Sequence to Sequence Learning with Neural Networks

---

**Ilya Sutskever**  
Google  
[ilyasu@google.com](mailto:ilyasu@google.com)

**Oriol Vinyals**  
Google  
[vinyals@google.com](mailto:vinyals@google.com)

**Quoc V. Le**  
Google  
[qvl@google.com](mailto:qvl@google.com)



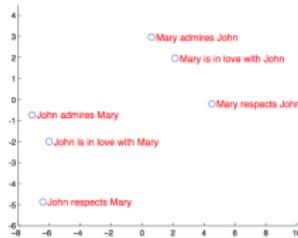
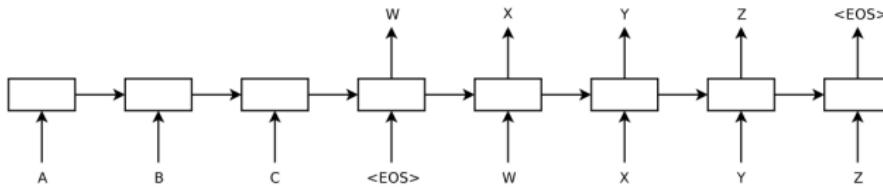
# Encoder-decoder - machine translation

## Sequence to Sequence Learning with Neural Networks

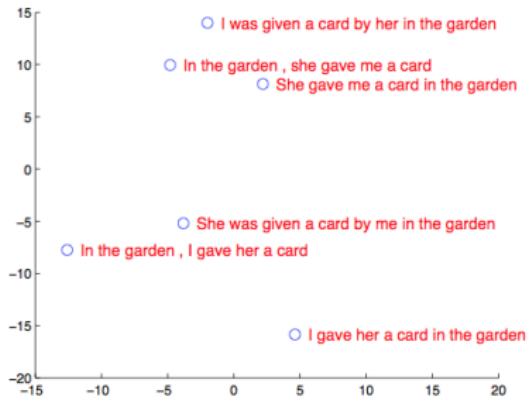
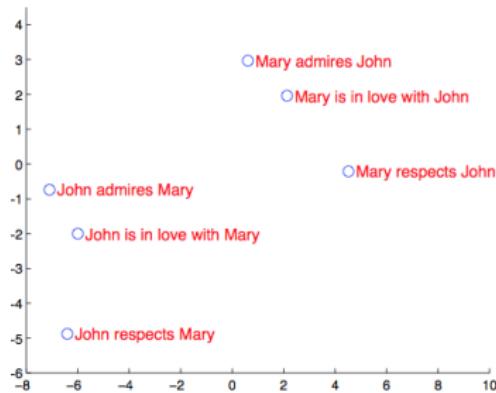
**Ilya Sutskever**  
Google  
[ilyasu@google.com](mailto:ilyasu@google.com)

**Oriol Vinyals**  
Google  
[vinyals@google.com](mailto:vinyals@google.com)

**Quoc V. Le**  
Google  
[qvl@google.com](mailto:qvl@google.com)



# Machine translation - visualising latent space



# Python interpretation with LSTM

---

## Learning to Execute

---

**Wojciech Zaremba**

Google & New York University

WOJ.ZAREMBA@GMAIL.COM

**Ilya Sutskever**

Google

ILYASU@GOOGLE.COM

# Python interpretation with LSTM

## Learning to Execute

Wojciech Zaremba

Google & New York University

WOJ.ZAREMBA@GMAIL.COM

Ilya Sutskever

Google

ILYASU@GOOGLE.COM

**Input:**

```
j=8584  
for x in range(8):  
    j+=920  
    b=(1500+j)  
    print((b+7567))
```

**Target:** 25011.**Model prediction:** 23011.**Input:**

```
c=445  
d=(c-4223)  
for x in range(1):  
    d+=5272  
    print((8942 if d<3749 else 2951))
```

**Target:** 8942.**Model prediction:** 8942.

# Dialogue system

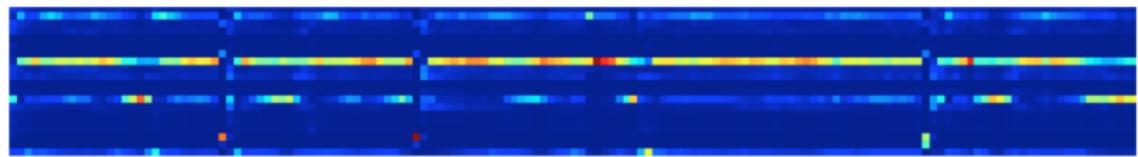
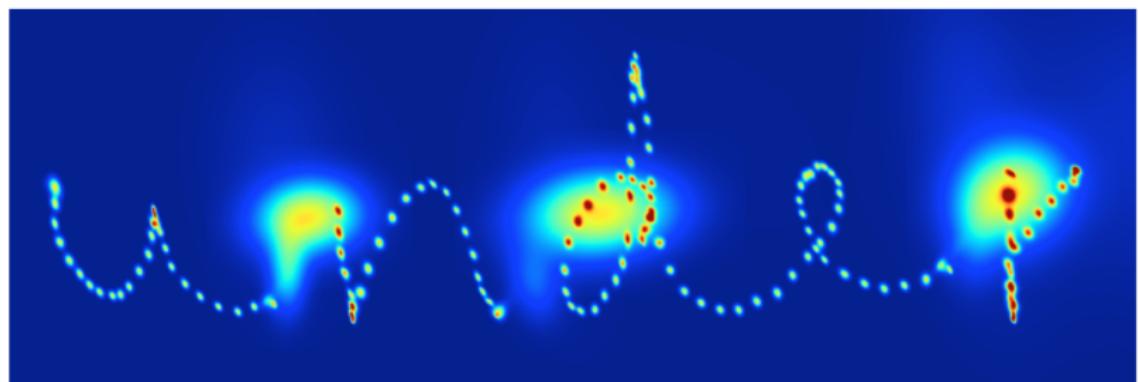
**Machine:** *hi*  
**Human:** *hello*  
**Machine:** *could you please let me know what are the operating systems you are using ?*  
**Human:** *i am using linux*  
**Machine:** *ok it was in the ssl vpn on which you stated it was broken*  
**Machine:** *please check if you are logged in to corp network*  
**Machine:** *what is the error that you are running please*  
**Human:** *i am seeing an error related to vpn*  
**Machine:** *what is the error message that you are getting when connecting to vpn using network connect ?*  
**Human:** *connection refused or something like that*  
**Machine:** *may i know the version of network connect you connect ?*  
**Human:** *i am not sure i know that*  
**Machine:** *is the network connect prompting for any user certificate*  
**Machine:** *?*  
**Human:** *yes*  
**Machine:** *are you at home or at work at the moment ?*  
**Machine:** *office ?*  
**Human:** *i am at home*  
**Machine:** *try this*  
**Machine:** *goto <URL> page and enter your name and password*

Many thanks to Pyry for sharing slides!

## Generating sequences - training data ( $x(t), y(t)$ )

would find the bus safe and sound  
As for Clark, unless it were a  
canvasser at the age of fifty-five  
Editorial. Dilemma of  
the the tides in the affairs of men;

## Generating sequences - uncertainty of pen position



Alex Graves, Generating Sequences With Recurrent Neural Networks

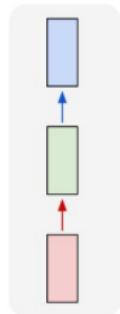
## Generating sequences - dreaming up new

purple misty dawn low lined  
borders of cold piney firs wine curvy  
height. Y Geeks the gather in  
cycle satet down in swing Te a  
over & high earne, Tend., madp

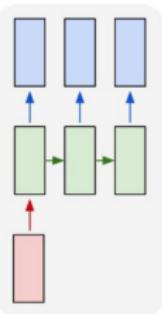
Alex Graves, Generating Sequences With Recurrent Neural Networks  
Demo

# Overview recurrent architectures

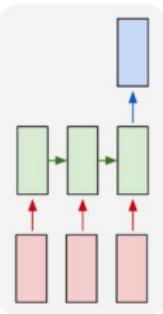
one to one



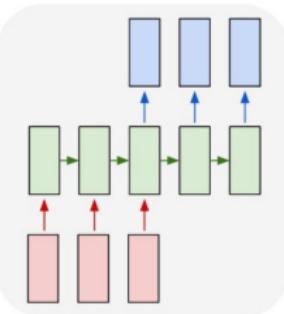
one to many



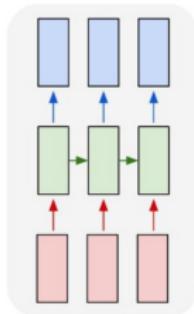
many to one



many to many



many to many



- Vanilla mode, no RNN.
- E.g. image classification

- Sequence output
- E.g. image captioning

- Sequence input
- E.g. sentiment analysis

- Sequence input and output (*encoder-decoder, sequence-to-sequence*)
- E.g. translation, question answering

- Synced sequence input and output
- E.g. label each video frame

From Andrej Karpathy blog: The Unreasonable Effectiveness of Recurrent Neural Networks

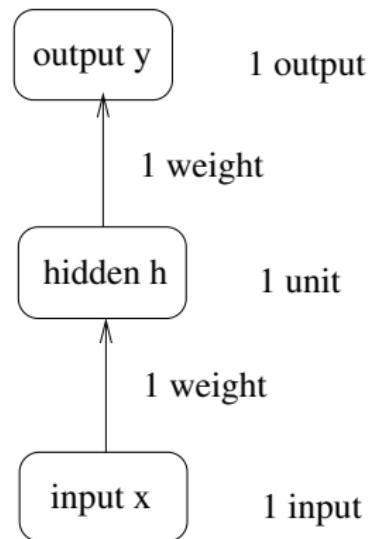
# Part 2:

## RNN training

### Backpropagation through time

# Recap backpropagation (Linnainmaa, 1970)

Computing gradients in a network.



- First with scalars. Use chain rule:

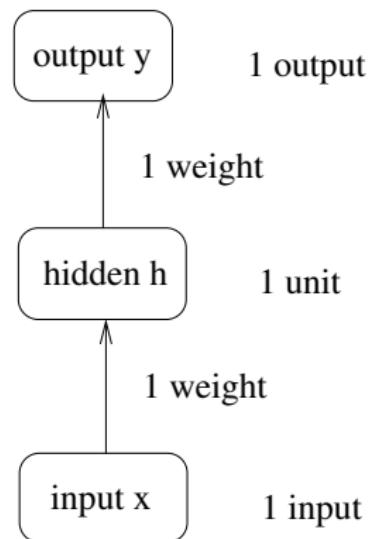
$$\frac{\partial C}{\partial w_2} = \frac{\partial C}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial w_2}$$

$$\frac{\partial C}{\partial w_1} = \frac{\partial C}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial h^{(1)}} \frac{\partial h^{(1)}}{\partial w_1}$$

- Chain rule:  $\frac{\partial h^{(2)}}{\partial x} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \frac{\partial h^{(1)}}{\partial x}$

# Recap backpropagation (Linnainmaa, 1970)

Computing gradients in a network.



- First with scalars. Use chain rule:

$$\frac{\partial C}{\partial w_2} = \frac{\partial C}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial w_2}$$

$$\frac{\partial C}{\partial w_1} = \frac{\partial C}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial h^{(1)}} \frac{\partial h^{(1)}}{\partial w_1}$$

- Chain rule:  $\frac{\partial h^{(2)}}{\partial x} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \frac{\partial h^{(1)}}{\partial x}$

# Back-propagation through time

- RNN definition:

$$\mathbf{h}_t^{(1)} = f_1 \left( W^{(1)} \mathbf{x}_t + W^{(\rightarrow)} \mathbf{h}_{t-1}^{(1)} \right)$$

$$\mathbf{h}_t^{(2)} = f_2 \left( W^{(2)} \mathbf{h}_t^{(1)} \right)$$

# Back-propagation through time

- RNN definition:

$$\mathbf{h}_t^{(1)} = f_1 \left( W^{(1)} \mathbf{x}_t + W^{(\rightarrow)} \mathbf{h}_{t-1}^{(1)} \right)$$

$$\mathbf{h}_t^{(2)} = f_2 \left( W^{(2)} \mathbf{h}_t^{(1)} \right)$$

- RNN weight derivatives:

$$\frac{\partial C}{\partial W^{(1,\rightarrow)}} = \sum_t \frac{\partial C}{\partial \mathbf{h}_t^{(1)}} \frac{\partial \mathbf{h}_t^{(1)}}{\partial W^{(1,\rightarrow)}}$$

$$\frac{\partial C}{\partial W^{(2)}} = \sum_t \frac{\partial C}{\partial \mathbf{h}_t^{(2)}} \frac{\partial \mathbf{h}_t^{(2)}}{\partial W^{(2)}}$$

# Back-propagation through time

- RNN definition:

$$\mathbf{h}_t^{(1)} = f_1 \left( W^{(1)} \mathbf{x}_t + W^{(\rightarrow)} \mathbf{h}_{t-1}^{(1)} \right)$$

$$\mathbf{h}_t^{(2)} = f_2 \left( W^{(2)} \mathbf{h}_t^{(1)} \right)$$

- RNN weight derivatives:

$$\frac{\partial C}{\partial W^{(1, \rightarrow)}} = \sum_t \frac{\partial C}{\partial \mathbf{h}_t^{(1)}} \frac{\partial \mathbf{h}_t^{(1)}}{\partial W^{(1, \rightarrow)}}$$

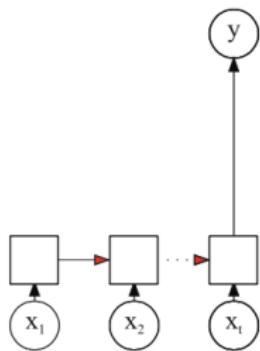
$$\frac{\partial C}{\partial W^{(2)}} = \sum_t \frac{\partial C}{\partial \mathbf{h}_t^{(2)}} \frac{\partial \mathbf{h}_t^{(2)}}{\partial W^{(2)}}$$

- Back-propagation through time:

$$\frac{\partial C}{\partial \mathbf{h}_t^{(1)}} = \frac{\partial C}{\partial \mathbf{h}_t^{(2)}} \frac{\partial \mathbf{h}_t^{(2)}}{\partial \mathbf{h}_t^{(1)}} + \frac{\partial C}{\partial \mathbf{h}_{t+1}^{(1)}} \frac{\partial \mathbf{h}_{t+1}^{(1)}}{\partial \mathbf{h}_t^{(1)}}$$

# Gradients vanishing and exploding

- Back-propagation through time:

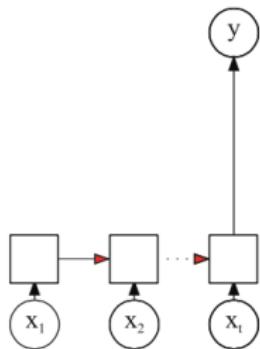


$$\frac{\partial C}{\partial \mathbf{h}_t^{(1)}} = \frac{\partial C}{\partial \mathbf{h}_t^{(2)}} \frac{\partial \mathbf{h}_t^{(2)}}{\partial \mathbf{h}_t^{(1)}} + \frac{\partial C}{\partial \mathbf{h}_{t+1}^{(1)}} \frac{\partial \mathbf{h}_{t+1}^{(1)}}{\partial \mathbf{h}_t^{(1)}}$$

- Scalars  $H = K = 1$  and length  $T$
- $\frac{\partial C}{\partial h_t^{(2)}} = 0$  for  $t < T$

# Gradients vanishing and exploding

- Back-propagation through time:



- Scalars  $H = K = 1$  and length  $T$
- $\frac{\partial C}{\partial h_t^{(2)}} = 0$  for  $t < T$
- Simplified back-propagation through time - ( $t < T$ ):

$$\frac{\partial C}{\partial h_T^{(1)}} = \frac{\partial C}{\partial h_T^{(2)}} f'_2 \left( W^{(2)} h_T^{(1)} \right) W^{(2)}$$

$$\frac{\partial C}{\partial h_t^{(1)}} = \frac{\partial C}{\partial h_{t+1}^{(1)}} f'_1 \left( a_t^{(1)} \right) W^{(\rightarrow)}$$

$$a_t^{(1)} \equiv W^{(1)} x_t + W^{(\rightarrow)} h_{t-1}^{(1)}$$

## Gradients vanishing and exploding continued

- Simplified back-propagation through time - ( $t < T$ ):

$$\frac{\partial C}{\partial h_t^{(1)}} = \frac{\partial C}{\partial h_{t+1}^{(1)}} f'_1 \left( a_t^{(1)} \right) W^{(\rightarrow)}$$
$$a_t^{(1)} \equiv W^{(1)} x_t + W^{(\rightarrow)} h_{t-1}^{(1)}$$

## Gradients vanishing and exploding continued

- Simplified back-propagation through time - ( $t < T$ ):

$$\frac{\partial C}{\partial h_t^{(1)}} = \frac{\partial C}{\partial h_{t+1}^{(1)}} f'_1 \left( a_t^{(1)} \right) W^{(\rightarrow)}$$
$$a_t^{(1)} \equiv W^{(1)} x_t + W^{(\rightarrow)} h_{t-1}^{(1)}$$

- Vanishing gradient:  $|f'_1 \left( a_t^{(1)} \right) W^{(\rightarrow)}| \ll 1$  - examples
- $|W^{(\rightarrow)}| \ll 1$

## Gradients vanishing and exploding continued

- Simplified back-propagation through time - ( $t < T$ ):

$$\frac{\partial C}{\partial h_t^{(1)}} = \frac{\partial C}{\partial h_{t+1}^{(1)}} f'_1 \left( a_t^{(1)} \right) W^{(\rightarrow)}$$
$$a_t^{(1)} \equiv W^{(1)} x_t + W^{(\rightarrow)} h_{t-1}^{(1)}$$

- Vanishing gradient:  $|f'_1 \left( a_t^{(1)} \right) W^{(\rightarrow)}| \ll 1$  - examples
- $|W^{(\rightarrow)}| \ll 1$
- $f(a) = \tanh(a)$ ,  $f'(a) = 1 - f^2(a) \ll 1$  for  $|a| \geq 3$ .

## Gradients vanishing and exploding continued

- Simplified back-propagation through time - ( $t < T$ ):

$$\frac{\partial C}{\partial h_t^{(1)}} = \frac{\partial C}{\partial h_{t+1}^{(1)}} f'_1(a_t^{(1)}) W^{(\rightarrow)}$$
$$a_t^{(1)} \equiv W^{(1)}x_t + W^{(\rightarrow)}h_{t-1}^{(1)}$$

- Vanishing gradient:  $|f'_1(a_t^{(1)}) W^{(\rightarrow)}| \ll 1$  - examples
- $|W^{(\rightarrow)}| \ll 1$
- $f(a) = \tanh(a)$ ,  $f'(a) = 1 - f^2(a) \ll 1$  for  $|a| \geq 3$ .
- Exploding gradient:  $|W^{(\rightarrow)}f'(a_t^{(1)})| > 1$  - example
- $f(a) = \text{relu}(a) = \max(0, a)$ ,  $f'(a) = (a > 0)$ ,  $a_t^{(1)} > 0$  and  $|W^{(\rightarrow)}| > 1$ .

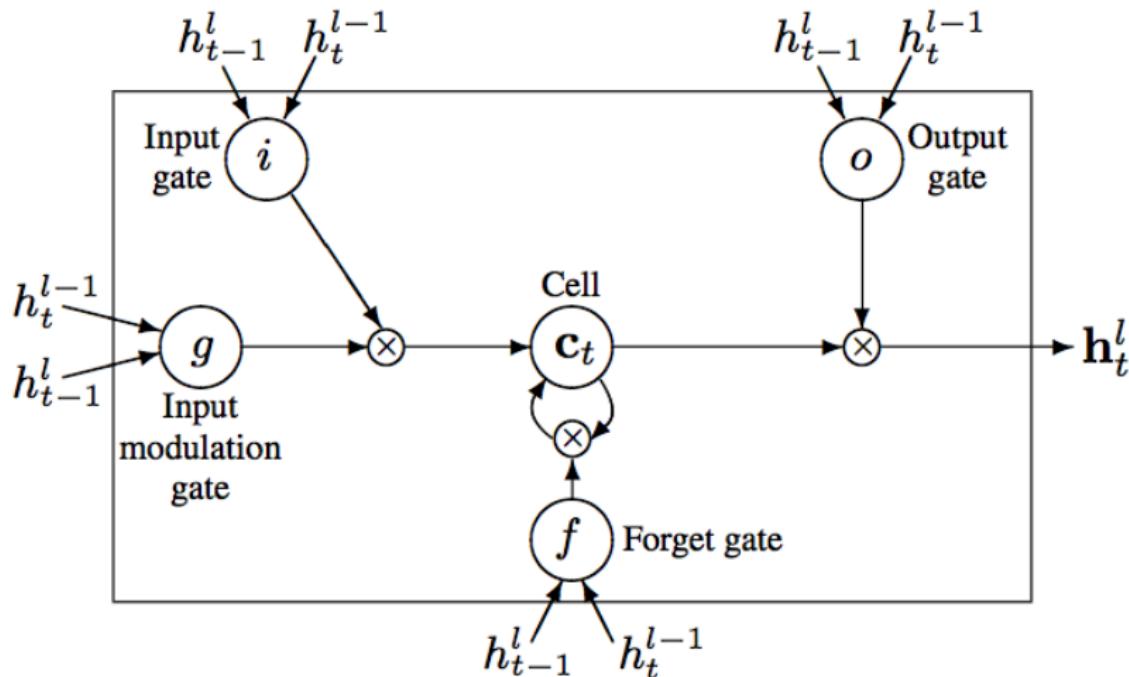
## Gradients vanishing and exploding continued

- Simplified back-propagation through time - ( $t < T$ ):

$$\frac{\partial C}{\partial h_t^{(1)}} = \frac{\partial C}{\partial h_{t+1}^{(1)}} f'_1(a_t^{(1)}) W^{(\rightarrow)}$$
$$a_t^{(1)} \equiv W^{(1)}x_t + W^{(\rightarrow)}h_{t-1}^{(1)}$$

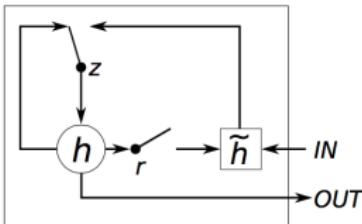
- Vanishing gradient:  $|f'_1(a_t^{(1)}) W^{(\rightarrow)}| \ll 1$  - examples
- $|W^{(\rightarrow)}| \ll 1$
- $f(a) = \tanh(a)$ ,  $f'(a) = 1 - f^2(a) \ll 1$  for  $|a| \geq 3$ .
- Exploding gradient:  $|W^{(\rightarrow)}f'(a_t^{(1)})| > 1$  - example
- $f(a) = \text{relu}(a) = \max(0, a)$ ,  $f'(a) = (a > 0)$ ,  $a_t^{(1)} > 0$  and  $|W^{(\rightarrow)}| > 1$ .
- Partial solution: gated units!

# Long short term memory cells



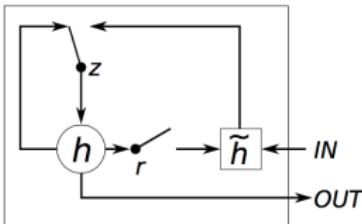
Hochreiter and Schmidhuber, 1997.

# Gated Feedback Recurrent Neural Networks



- Chung et al, 2015

# Gated Feedback Recurrent Neural Networks



- Chung et al, 2015

$$\mathbf{r}_t^{(1)} = \sigma \left( U^{(1)} \mathbf{x}_t + U^{(\rightarrow)} \mathbf{a}_{t-1}^{(1)} \right)$$

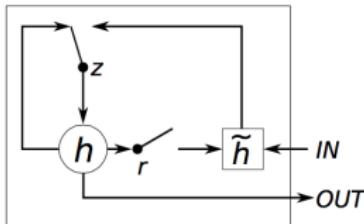
$$\mathbf{z}_t^{(1)} = \sigma \left( V^{(1)} \mathbf{x}_t + V^{(\rightarrow)} \mathbf{a}_{t-1}^{(1)} \right)$$

$$\tilde{\mathbf{a}}_t^{(1)} = W^{(1)} \mathbf{x}_t + W^{(\rightarrow)} (\mathbf{r}_t^{(1)} \circ \mathbf{a}_{t-1}^{(1)})$$

$$\mathbf{a}_t^{(1)} = (1 - \mathbf{z}_t^{(1)}) \circ \mathbf{a}_{t-1}^{(1)} + \mathbf{z}_t^{(1)} \circ f_1(\tilde{\mathbf{a}}_t^{(1)})$$

- $\sigma()$  is the logistic function  $\in [0, 1]$ .

# Gated Feedback Recurrent Neural Networks



- Chung et al, 2015

$$\mathbf{r}_t^{(1)} = \sigma \left( U^{(1)} \mathbf{x}_t + U^{(\rightarrow)} \mathbf{a}_{t-1}^{(1)} \right)$$

$$\mathbf{z}_t^{(1)} = \sigma \left( V^{(1)} \mathbf{x}_t + V^{(\rightarrow)} \mathbf{a}_{t-1}^{(1)} \right)$$

$$\tilde{\mathbf{a}}_t^{(1)} = W^{(1)} \mathbf{x}_t + W^{(\rightarrow)} (\mathbf{r}_t^{(1)} \circ \mathbf{a}_{t-1}^{(1)})$$

$$\mathbf{a}_t^{(1)} = (1 - \mathbf{z}_t^{(1)}) \circ \mathbf{a}_{t-1}^{(1)} + \mathbf{z}_t^{(1)} \circ f_1(\tilde{\mathbf{a}}_t^{(1)})$$

- $\sigma()$  is the logistic function  $\in [0, 1]$ .
- Note that if  $(\mathbf{z}_t^{(1)})_i = 0$  then  $(\mathbf{a}_t^{(1)})_i = (\mathbf{a}_{t-1}^{(1)})_i$

# Inside and LSTM cell

Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Cell that robustly activates inside if statements:

Karpathy et al, 2016

## Inside and LSTM cell 2

Cell that turns on inside comments and quotes:

```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
    struct audit_field *sf)
{
    int ret = 0;
    char *lsm_str;
    /* our own copy of lsm_str */
    lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
    if (unlikely(!lsm_str))
        return -ENOMEM;
    df->lsm_str = lsm_str;
    /* our own (refreshed) copy of lsm_rule */
    ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
        (void *)df->lsm_rule);
    /* Keep currently invalid fields around in case they
     * become valid after a policy reload. */
    if (ret == -EINVAL) {
        pr_warn("audit rule for LSM \\\"%s\\\" is invalid\n",
            df->lsm_str);
        ret = 0;
    }
    return ret;
}
```

# Part 3: Deep learning for information retrieval

## Ellen's case story



For 25 years, Ellen struggled to find a diagnosis for the multitude of debilitating symptoms that seemed to increase year after year.

- Her symptoms included muscle cramps, intense headaches, rapid weight gain, fatigue, edema, intolerance to heat, excessive sweating, joint pain, tingling in her hands and feet, frequent bone fractures, acid reflux, intense anxiety and panic attacks, high blood pressure, high cholesterol, high blood sugar, sleep apnea, menstrual irregularities, peripheral vision loss and double vision.
- Source: <http://www.uptodate.com/>
- Any suggestions?

“When you hear hoofbeats behind you, don’t expect to see a zebra”

Search

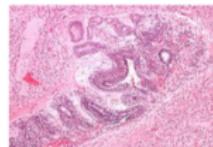
### Filters

Group by:  disease  gene

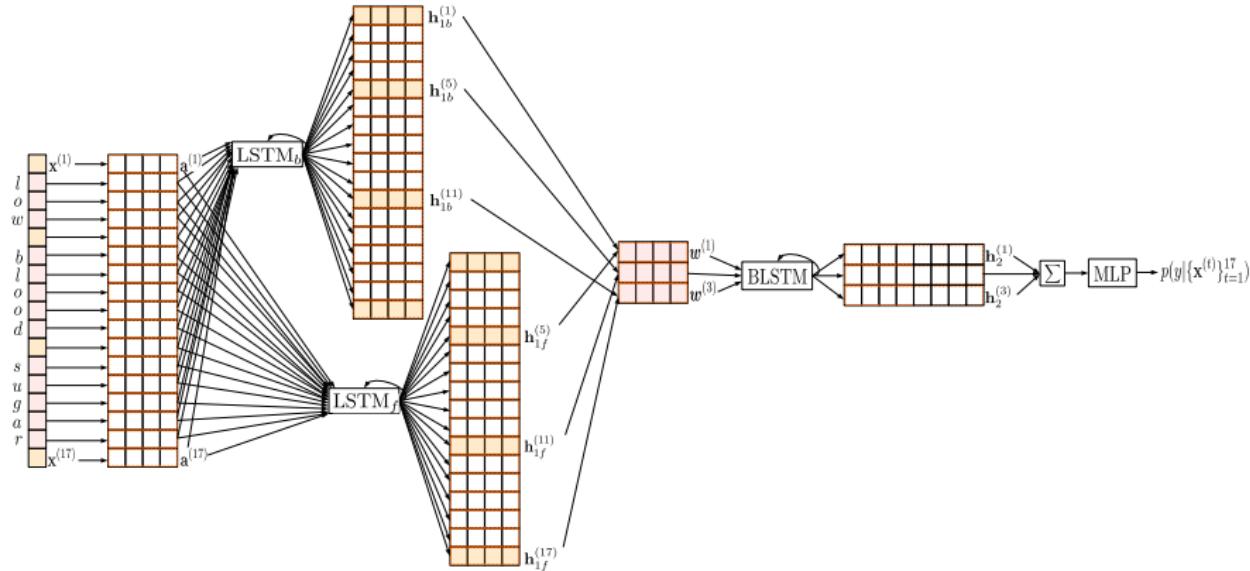
Mediastinal tumor	Mediastinal tumor
Desmoplastic small round cell ... (2)	Retrieved: 28-09-2014 Source: WIKIPEDIA ( <a href="#">Original article</a> )
Thymoma (2)	The mediastinum is the cavity that separates the lungs from the rest of the chest. It contains the heart, esophagus, trachea, thymus, and aorta. The mediastinum has three main parts: the <b>anterior</b> mediastinum (front), the middle mediastinum, and the posterior mediastinum (back). The most common <b>mediastinal</b> masses are neurogenic tumors (20% of <b>mediastinal</b> tumors), usually found in the posterior mediastinum, followed by thymoma (15-20%) located in the <b>anterior</b> mediastinum. Masses in the <b>anterior</b> portion of the mediastinum can include thymoma, lymphoma, pheochromocytoma, germ cell tumors including teratoma, thyroid tissue, and parathyroid lesions. Masses in this area are more likely to be malignant than those in other compartments. Masses in the posterior portion of the mediastinum tend to be neurogenic in origin, and in adults tend to be of neural sheath origin including neurilemmomas and neurofibromas. Lung cancer typically spreads to the lymph nodes in the mediastinum.
Adult-onset Still's disease (3)	
Large-cell lymphoma (3)	
Follicular lymphoma	
Periodic fever, familial, autos... (2)	
Thymic carcinoma	
TNF receptor associated periodic... (2)	
Japanese encephalitis	
Periodic fever, familial, autosomal... (2)	

### Diagnosis

In several editions of Physical Diagnosis, concerning **mediastinal** tumors the author writes: According to Christian1 the **mediastinal** neoplasms which are neither so rare nor so obscure as to make diagnosis practically impossible are:

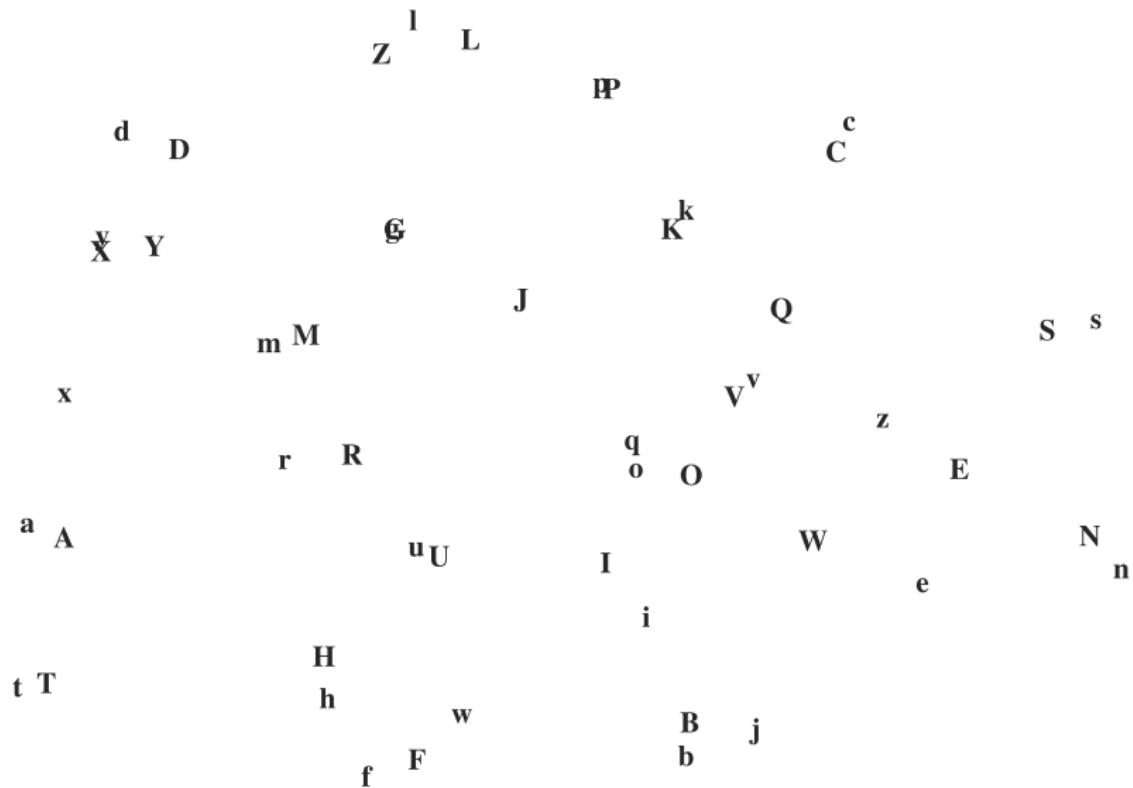


# Deep learning - for search

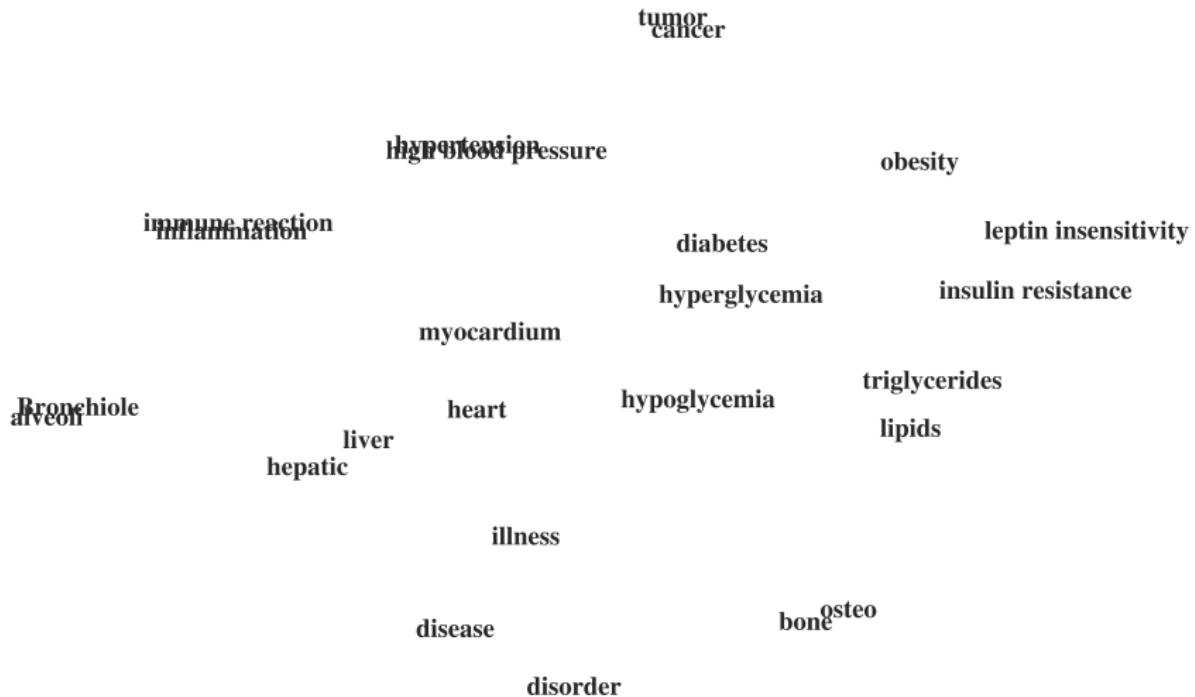


- $\sim 8260$  CUIs - unique disease labels
- $\sim 14261$  articles OMIM, Wikipedia, Orphanet, NIH sources
- Training: map 100 character article snippets to disease label
- Test: queries with diagnosis.

# Visualizing latent representation for characters



# Visualizing latent representation for words



# Example queries

## Query

- syndrome characterized by the following tumors: medullary thyroid, adrenomedullary, parathyroid
- diagnosis associated with abdominal obesity, elevated triglycerides, low hdl, hypertension, and elevated glucose
- cystic lung disease in women of childbearing age
- 8 months old, male, progressive signs of respiratory distress, tachypnea, pulmonary hypertension, tortuosity of aortic arch, facial dysmorphisms
- eponym for acute pericarditis 2 weeks after myocardial infarction
- diagnosis for otalgia, ear discharge, and pruritis
- hemolytic anemia associated with IgM antibodies directed against RBC i/i antigens
- recurrent unilateral retro-orbital headache with remissions and relapses
- most likely diagnosis causing generalized pruritus in the setting of inflammatory bowel disease
- diagnosis associated with anion gap acidosis, osmolar gap, and urinary calcium oxalate crystals
- diagnosis when platelet count > 600,000/ $\mu$ l in absence reactive thrombocytosis
- autism spectrum disorder characterized by social impairment and repetitive behaviors but normal language and cognition development (final jeopardy)
- cancer associated with hepatitis C
- diagnosis for regular tachycardia associated with intermittent jugular cannon waves
- myocardial complication in patients who receive mantle-field radiation for Hodgkin's disease
- most common cause of goiter worldwide
- personality disorder characterized by unstable interpersonal relationships, impulsiveness, and self-destructive behaviors
- most common type of paroxysmal reentrant supraventricular tachycardia
- tropical zoonosis associated with hemorrhage, abdominal pain, hepatosplenomegaly and this finding
- most consistent radiographic finding associated with ankylosing spondylitis
- signature injury of soldiers deployed to Iraq and Afghanistan
- periodic headache classically found in male smokers
- agent causing conjunctivitis and umbilicated papular lesions on lid margins
- syndrome characterized by acute swelling and pain in the face and arms that evolves into a brawny, indurated edema in patients who have performed unaccustomed physical exertion
- hereditary condition suggested by genital edema in a child riding a bicycle
- characterized by lack of expression of CD55 and CD59 on erythrocytes
- gene mutation associated with familial adenomatous polyposis
- eponym applied to a painless nodule in the palmar aponeurosis
- diagnosis classically associated with low-back pain and stiffness that worsens at night and improves with physical activity or a hot shower
- diagnosis associated with hepatitis, negative viral serology, psychiatric symptoms and family history of hepatitis
- most common cause of edema and hypothyroidism in USA
- name of childhood illness caused by parvovirus B19

Disease	Combined Rank	Chars2WordRank	Words2DocRank	Solr Rank
men type 2	1	1	1	1
metabolic syndrome	1	1	1	1
lymphangiomyomatosis	1	1	1	1
arterial tortuosity syndrome	1	1	1	1
dressler's syndrome	1	2	1	1
otitis externa	1	2	1	1
cold agglutinin disease	1	2	1	1
cluster headache	1	2	1	1
primary sclerosing cholangitis	1	7	1	1
ethylene glycol poisoning	1	9	1	1
essential thrombocythemia	2	1	1	2
asperger syndrome	2	1	1	2
hepatocellular carcinoma	2	1	1	10
ventricular tachycardia	2	1	1	21
myocardial infarction	2	1	1	23
iodine deficiency	3	4	2	2
borderline personality	3	5	2	2
atrioventricular nodal reentrant tachycardia	3	5	3	3
leptospirosis	3	20	2	2
sacroiliitis	3	46	2	2
traumatic brain injury	3	69	2	2
cluster headache	3	408	2	2
molluscum contagiosum	3	774	2	2
eosinophilic fascitis	3	1009	2	2
hereditary angioedema	4	2	40	40
paroxysmal nocturnal hemoglobinuria	5	53	3	3
apc gene	6	4	17	17
dupuytren contracture	7	7	4	4
ankylosing spondylitis	8	4	26	26
wilson's disease	8	7	19	19
hashimoto disease	10	6	7	7
erythema infectiosum	14	7	14	14

# Results

<i>Metric</i> <i>Model</i>	MRR	Recall@10	Recall@20
Chars2Doc CNN (Small)	0.195	0.319	0.387
Chars2Doc BLSTM (Small)	0.174	0.343	0.407
Chars2Words2Doc BLSTM (Large)	0.327	0.548	<b>0.665</b>
Solr	<b>0.355</b>	<b>0.585</b>	0.657

# Results

<i>Metric</i> <i>Model</i>	MRR	Recall@10	Recall@20
Chars2Doc CNN (Small)	0.195	0.319	0.387
Chars2Doc BLSTM (Small)	0.174	0.343	0.407
Chars2Words2Doc BLSTM (Large)	0.327	0.548	<b>0.665</b>
Solr	<b>0.355</b>	<b>0.585</b>	0.657

- SOLR and C2W2D make different errors (Recall@20)

		Chars2Words2Doc	
		+	-
Solr	+	0.540	0.117
	-	0.125	0.218

# Results

<i>Metric</i> <i>Model</i>	MRR	Recall@10	Recall@20
Chars2Doc CNN (Small)	0.195	0.319	0.387
Chars2Doc BLSTM (Small)	0.174	0.343	0.407
Chars2Words2Doc BLSTM (Large)	0.327	0.548	<b>0.665</b>
Solr	<b>0.355</b>	<b>0.585</b>	0.657

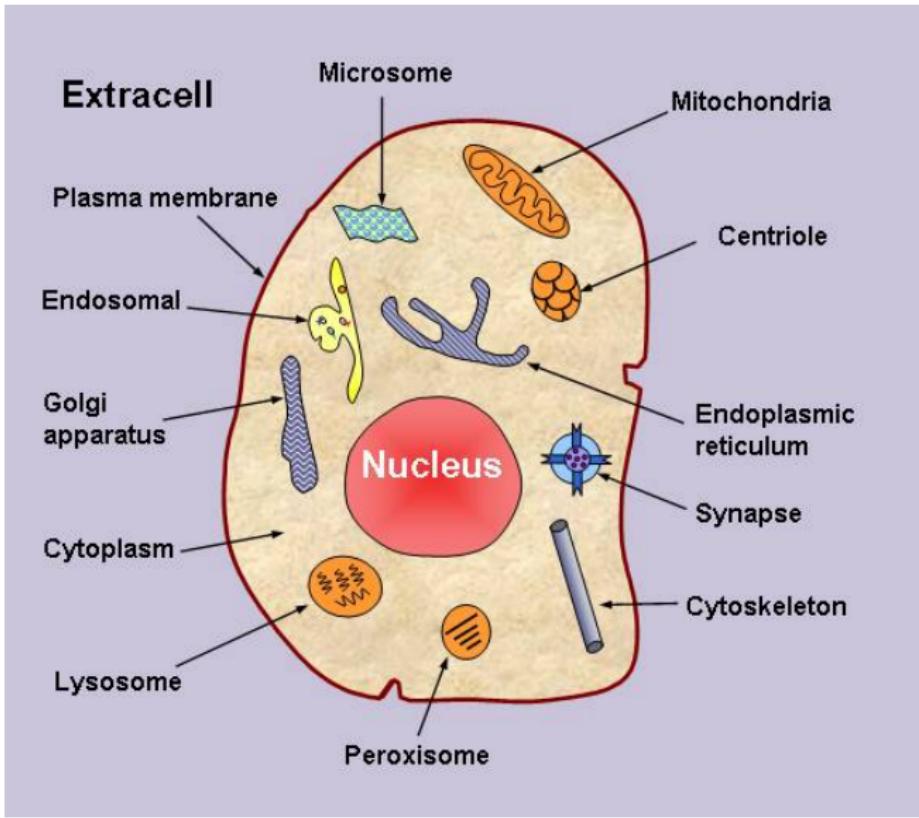
- SOLR and C2W2D make different errors (Recall@20)

		Chars2Words2Doc	
		+	-
Solr	+	0.540	0.117
	-	0.125	0.218

- Simple combination:  
MRR=0.373, Recall@10=0.657 and Recall@20=0.738

# Part 4: Deep learning for biological sequence analysis

# Sub-cellular localization



# Sub-cellular localization machine learning set-up

- Use (one-hot encoded) protein sequence as input

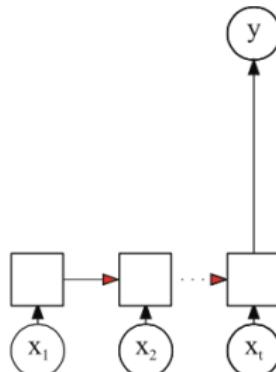
```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILEPFTMVALAGVHLTFLHETGSNNPLGLTSDDSKIPFHPYYTIKDFLG
LLILLLLLLLALLSPDMLGDPDNHMPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLLTWIGSQPVVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

# Sub-cellular localization machine learning set-up

- Use (one-hot encoded) protein sequence as input

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPIYGITNLV
EWIWGGFSVDKATLNRFCAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSSDKIPFHPYYTIKDFLG
LLLILLLLLLALLSPDMLGDPDNHMPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMRLPLSQALFWTLTMDLLLTWIGSQPVVEPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

- 11 localization classes softmax output
- RNN architecture - needs to memorize to the end!

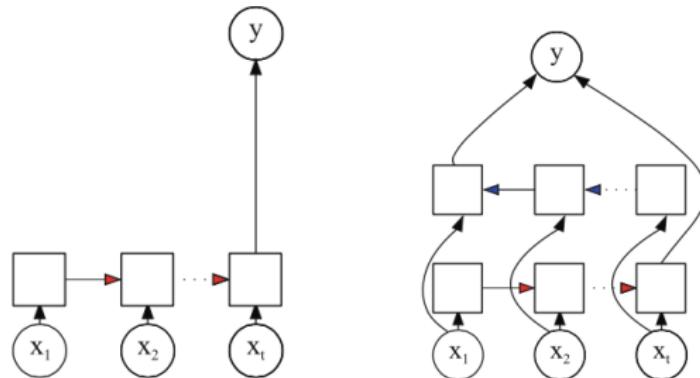


# Sub-cellular localization machine learning set-up

- Use (one-hot encoded) protein sequence as input

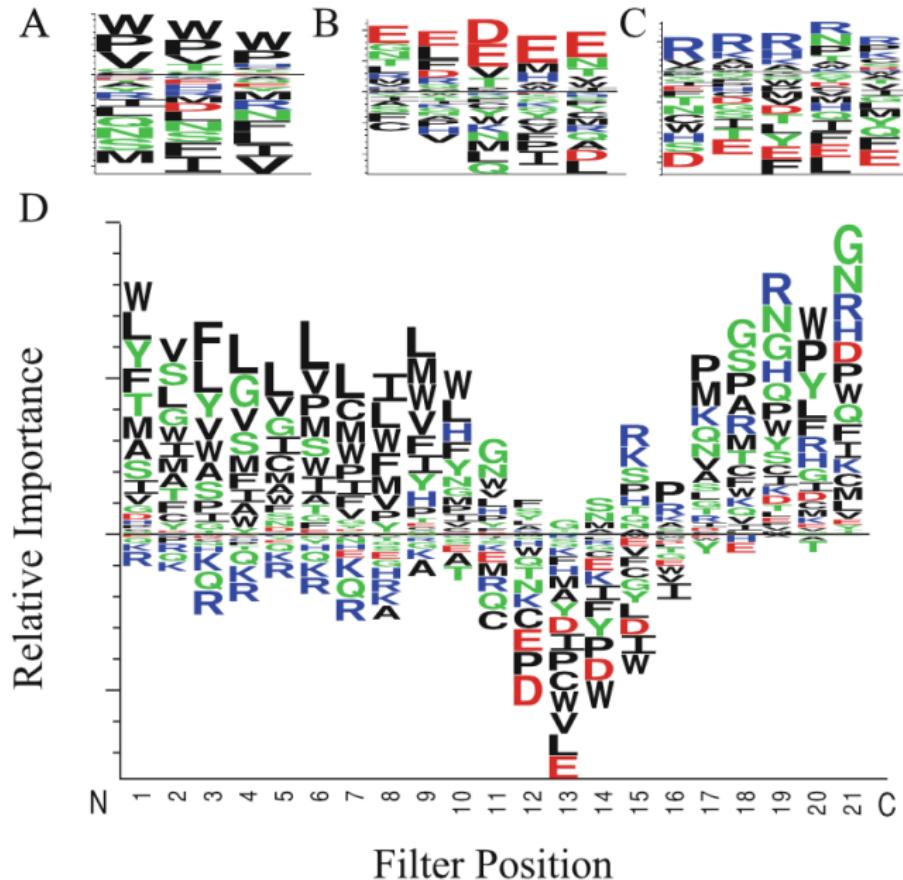
```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPIYGITNLV
EWIWGGFSVDKATLNRFCAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDDSKIPFHPYYTIKDFLG
LLLILLLLLLALLSPDMLGDPDNHMPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMRLPLSQALFWTLTMDLLLTWTIGSQPVVEPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

- 11 localization classes softmax output
- RNN architecture - needs to memorize to the end!



- First layer is a convolutional layer
- Sonderby et. al. <http://arxiv.org/abs/1503.01919>

# Visualizing learned convolutions



# Advanced feature - attention

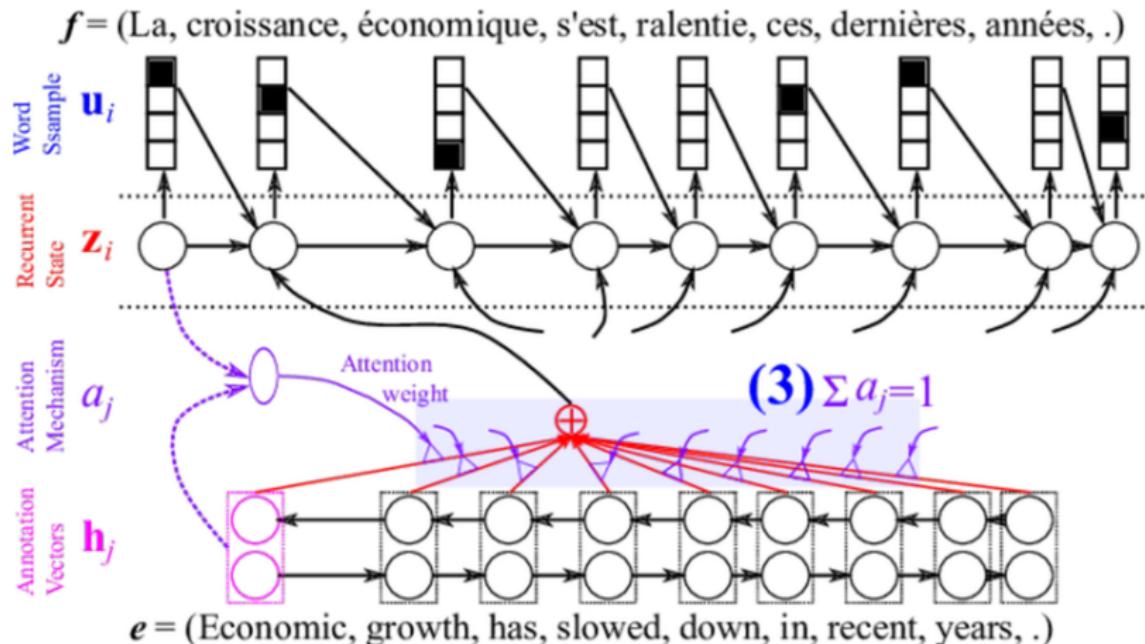


Figure 5. The relevance scores returned by the attention mechanism are normalized to sum to 1, which helps us interpret them as probabilities. From this probabilistic perspective, we compute the expectation of the annotation vectors under this distribution.

# Advanced feature - attention

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



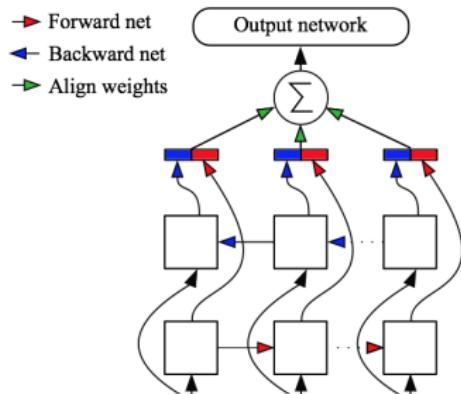
A woman is throwing a frisbee in a park.



Figure 6. Sample translations made by the neural machine translation model with the soft-attention mechanism.  
Edge thicknesses represent the attention weights found by the attention model.

# Attention mechanism

- Attention mechanism (adapted from Bahdanau et al, 2014)



- Compute context vector  $c$ ,  $\dim(c) = \dim(h_t)$ ,  $h_t = \begin{bmatrix} h_t^{(f)} \\ h_t^{(b)} \end{bmatrix}$

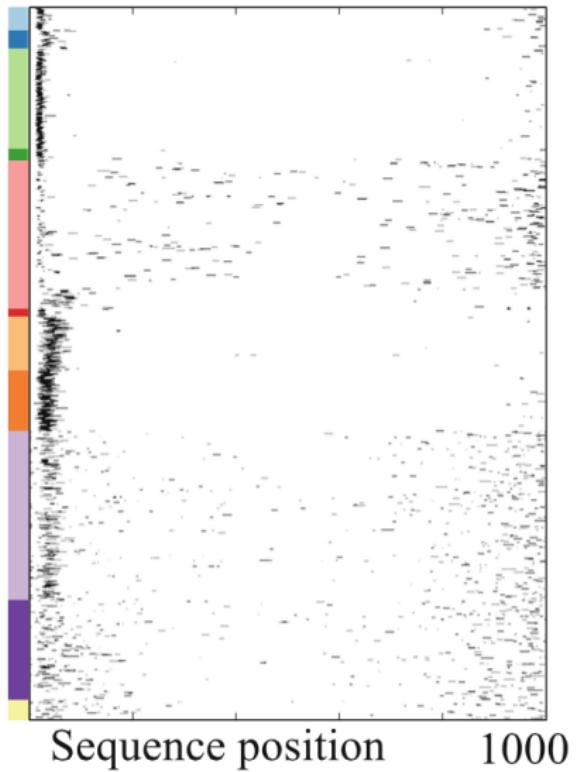
$$c = \sum_{t=1}^{T_x} a_t h_t$$

$$a_t = \frac{\exp(f(\textcolor{green}{h}_t; W))}{\sum_{t'=1}^{T_x} \exp(f(\textcolor{green}{h}_{t'}; W)))}$$

that are fed into feed-forward network.

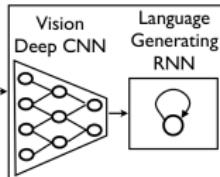
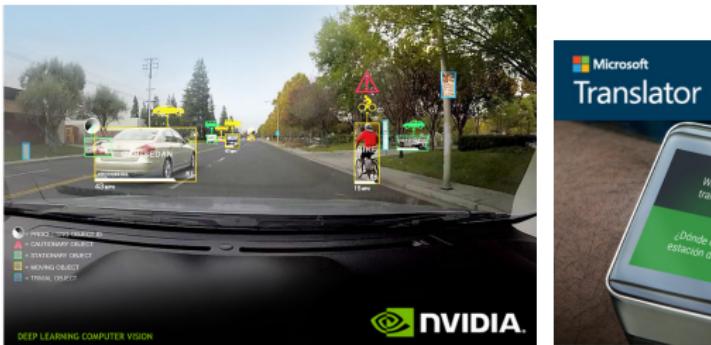
# Where does the network pay attention?

- █ ER
- █ Golgi
- █ Extracellular
- █ Lysosomal
- █ Plasma membrane
- █ Vacuolar
- █ Chloroplast
- █ Mitochondrial
- █ Cytoplasmic
- █ Nuclear
- █ Peroxisomal



# Part 5: Supervised learning recap

# Deep learning is changing the world



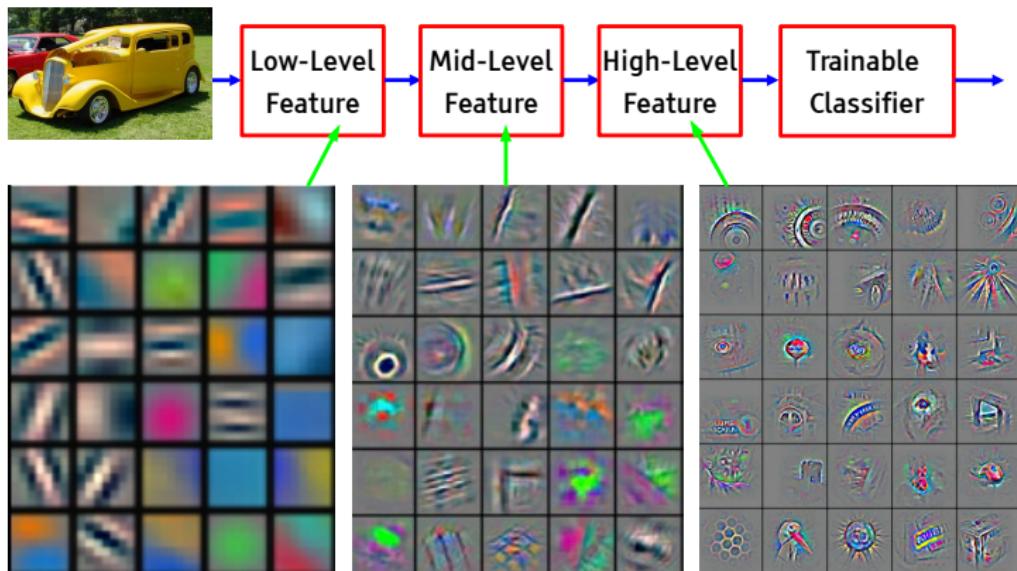
**A group of people shopping at an outdoor market.**

**There are many vegetables at the fruit stand.**

# Deep Learning = Learning Hierarchical Representations

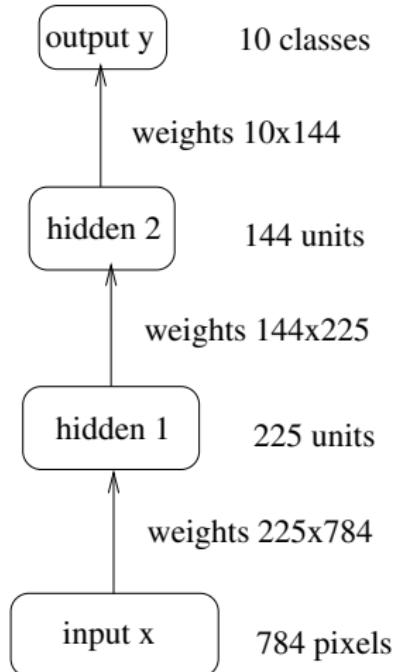
Y LeCun

It's deep if it has more than one stage of non-linear feature transformation



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

# Example Network



$$\mathbf{h}^{(3)} = \text{softmax}(\mathbf{W}^{(3)}\mathbf{h}^{(2)} + \mathbf{b}^{(3)})$$

$$\mathbf{h}^{(2)} = \text{relu}(\mathbf{W}^{(2)}\mathbf{h}^{(1)} + \mathbf{b}^{(2)})$$

$$\mathbf{h}^{(1)} = \text{relu}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)})$$

$$\text{softmax}(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

$$\text{relu}(z) = \max(0, z)$$

# Training criterion

Find parameters

$$\theta = \{\mathbf{W}^{(L)}, \mathbf{b}^{(L)}\}$$

that minimize expected negative log-likelihood:

$$C = \mathbb{E}_{\text{data}} [-\log P(\mathbf{y}|\mathbf{x}, \theta)].$$

⇒ Learning becomes optimization.

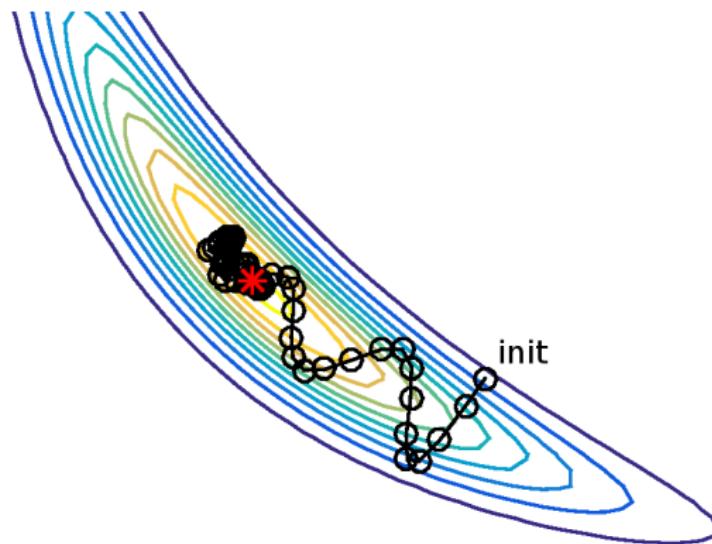
# Backpropagation

- For computing the gradient  $\mathbf{g} = \nabla_{\theta} C(\theta) = \begin{pmatrix} \frac{\partial C}{\partial \theta_1} \\ \vdots \\ \frac{\partial C}{\partial \theta_n} \end{pmatrix}$ 
  - Chain rule** for derivatives
  - Dynamic programming** avoids exponential complexity
- Modern libraries do automatic derivations

## Momentum method, mini-batch training

$$\mathbf{m}_{k+1} = \alpha \mathbf{m}_k - \eta_k \mathbf{g}_k$$

$$\theta_{k+1} = \theta_k + \mathbf{m}_{k+1}$$



Use only a small batch of training data at once,  
noisy estimate of gradient  $\mathbf{g}$  is ok.

## Sensitivity and initialisation

- Signals in deep models tend to exponentially decay/explode
- Solution 1: Careful initialisation
- Solution 2: Batch normalisation  
(also speeds up and regularises)

# Goal of Regularisation

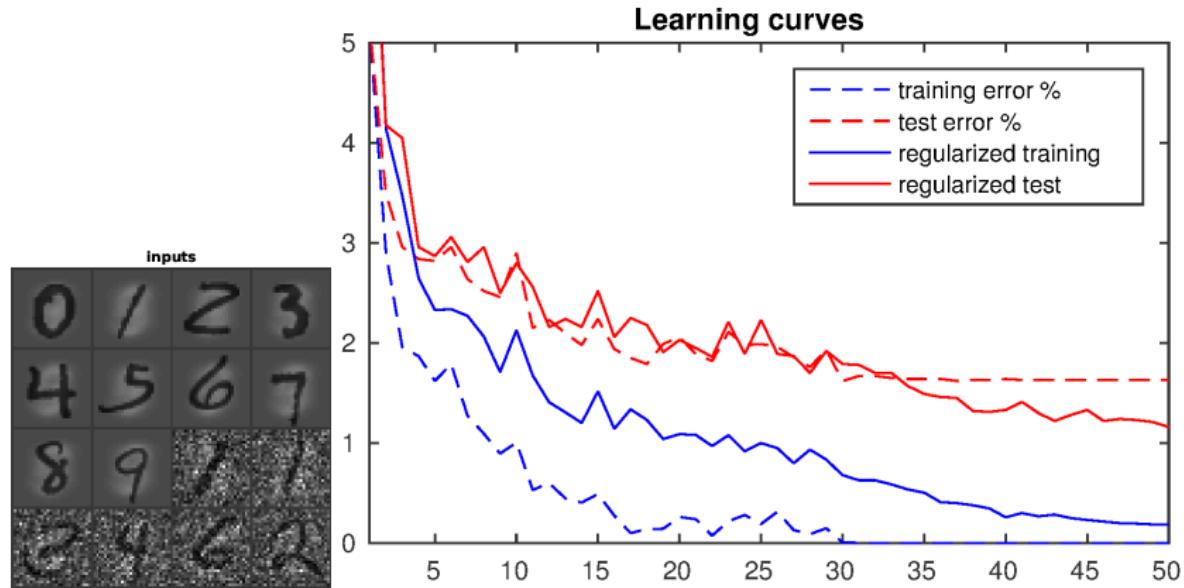
- Easy to perform great on the training set (overfitting).
- **Regularisation** improves generalisation to new data.
- Use held-out validation data to choose hyperparameters (e.g. regularisation strength).
- Use held-out test data to evaluate performance.

# Regularization methods

[www.deeplearningbook.org/contents/regularization.html](http://www.deeplearningbook.org/contents/regularization.html)

- Limited size of network
- Early stopping
- Weight decay
- Data augmentation
- Injecting noise
- Parameter sharing (e.g. convolutional)
- Sparse representations
- Ensemble methods
- Auxiliary tasks (e.g. unsupervised)
- Probabilistic treatment (e.g. variational methods)
- Adversarial training, ...

## Example: Injecting noise



Without regularisation **training error** goes to zero and learning stops.

With noise regularisation, **test error** keeps dropping.

## References

- <https://github.com/kjw0612/awesome-rnn>
- <http://arxiv.org/pdf/1507.01273.pdf>
- <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>
- <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- <http://arxiv.org/pdf/1211.5063.pdf>
- <http://arxiv.org/abs/1506.02078>
- <http://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/>
- <http://arxiv.org/pdf/1506.03340.pdf>
- <http://arxiv.org/abs/1502.03044>
- <http://arxiv.org/abs/1410.3916>
- <http://arxiv.org/abs/1410.5401>
- <http://arxiv.org/abs/1410.4615>
- <http://arxiv.org/pdf/1506.05869.pdf>



Thanks!  
Ole Winther