From Deep Learning to Neural Programs

Roland Memisevic
University of Montreal // Twenty Billion Neurons
November 26, 2015

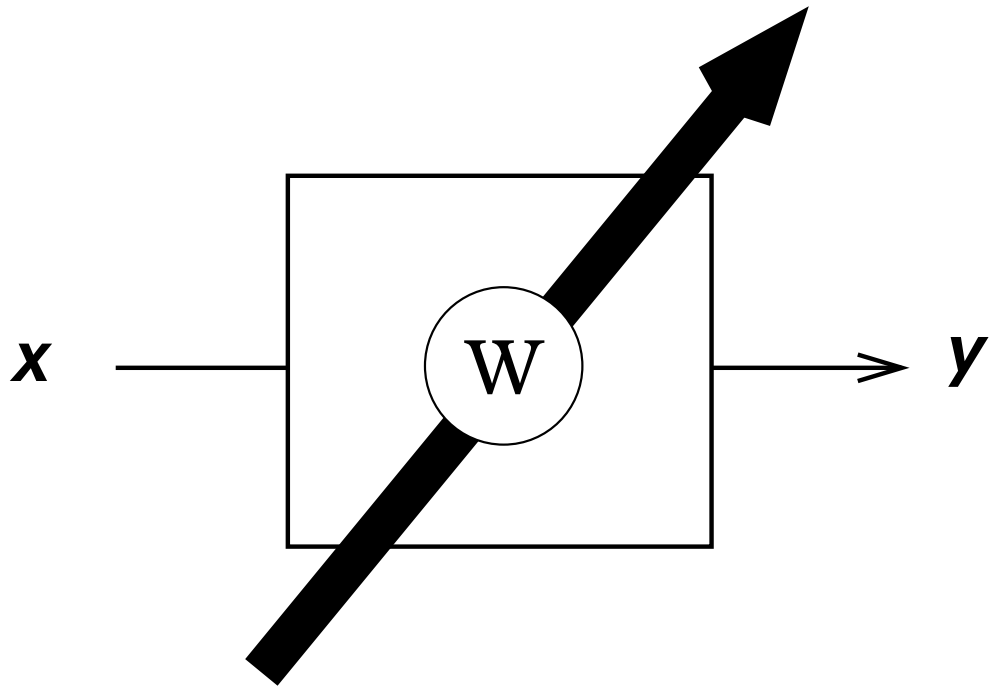
The three major thrusts in Deep Learning

- *The role of fast hardware*
- *The importance of data pooling*
- *The transition to "neural programs"*

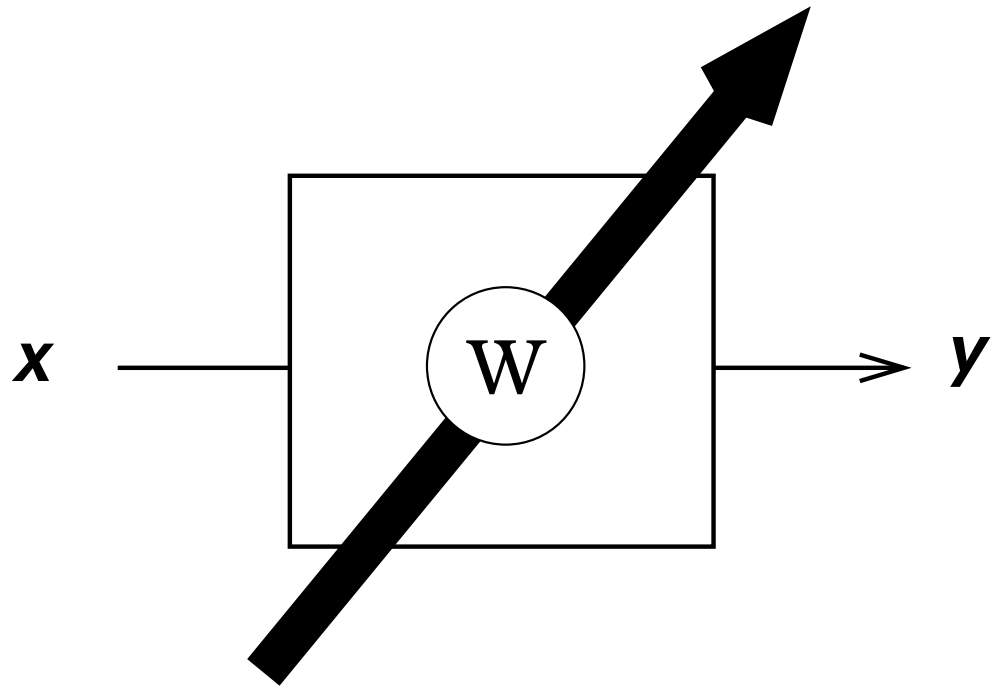
Deep learning is enabled by parallel hardware

Parallel hardware is enabled by deep learning

Machine Learning

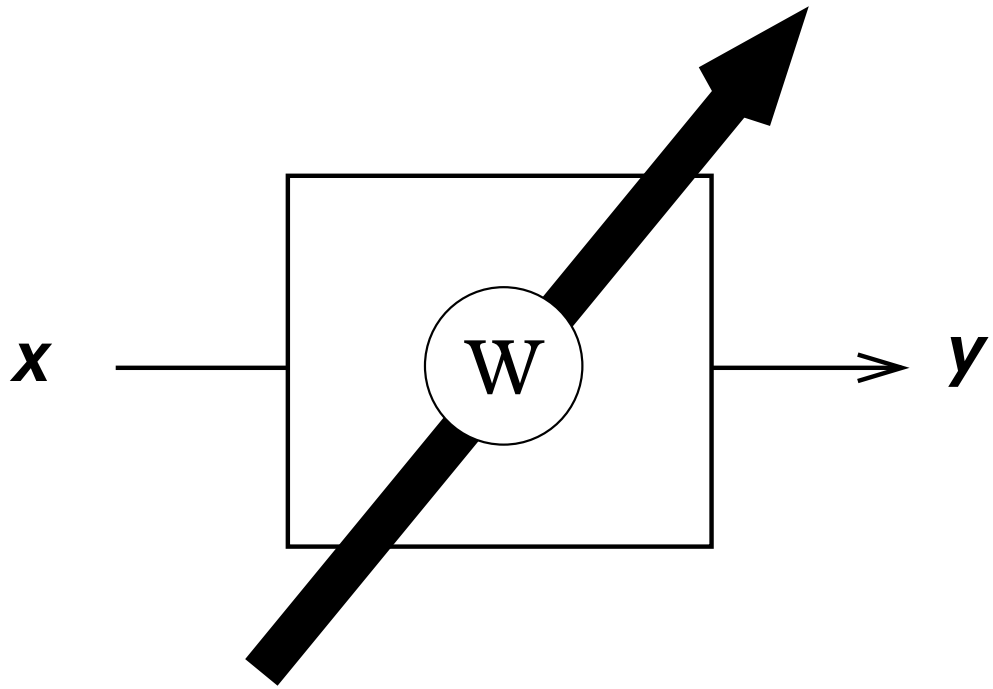


Machine Learning



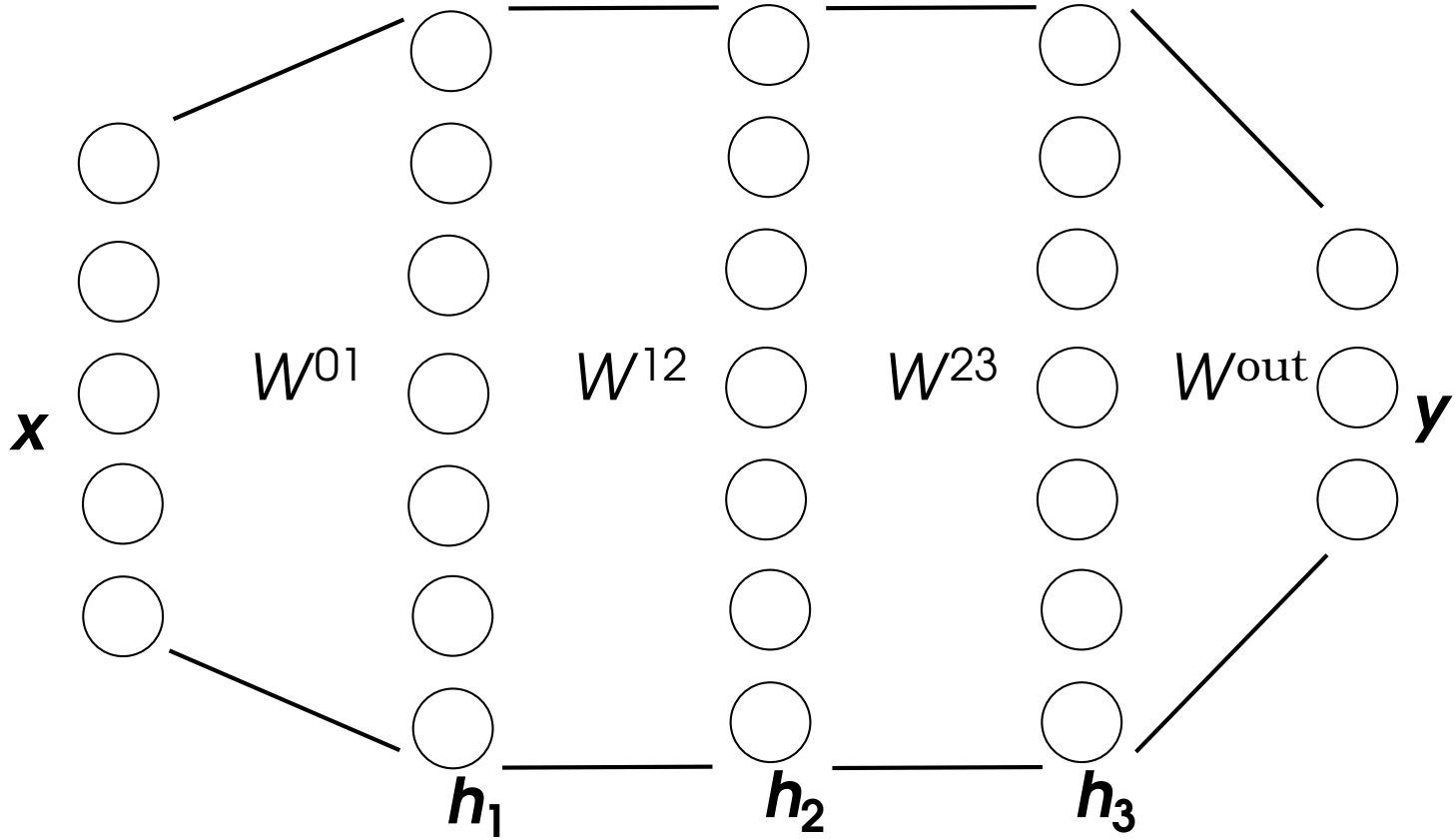
- ML allows us to harness **training data** $\left(\boldsymbol{x}_n, \boldsymbol{t}_n\right)_{n=1\ldots N}$

Machine Learning



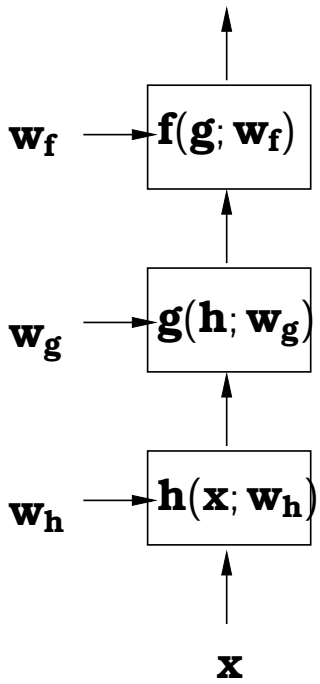
- ML allows us to harness **training data** $\left(\boldsymbol{x}_n, \boldsymbol{t}_n\right)_{n=1\ldots N}$
- ML allows us to harness **parallelization**

Neural networks



$$\boldsymbol{y}(\mathbf{x}) = W^{\text{out}} h(W^{23} h(W^{12} h(W^{01} \boldsymbol{x})))$$

Backprop



$\mathbf{f}$ $\frac{\partial \mathbf{f}}{\partial \mathbf{g}}$ $\frac{\partial \mathbf{f}}{\partial \mathbf{w_f}}$

fprop bprop grad

- Backprop prescribes how to compose end-to-end trainable systems from differentiable components:
- Use components which provide the methods **fprop**, **bprop** and **grad**. Now gradient computations can be completely automated using the chain-rule.
- Well-suited for support by software frameworks.

DL impact in speech recognition

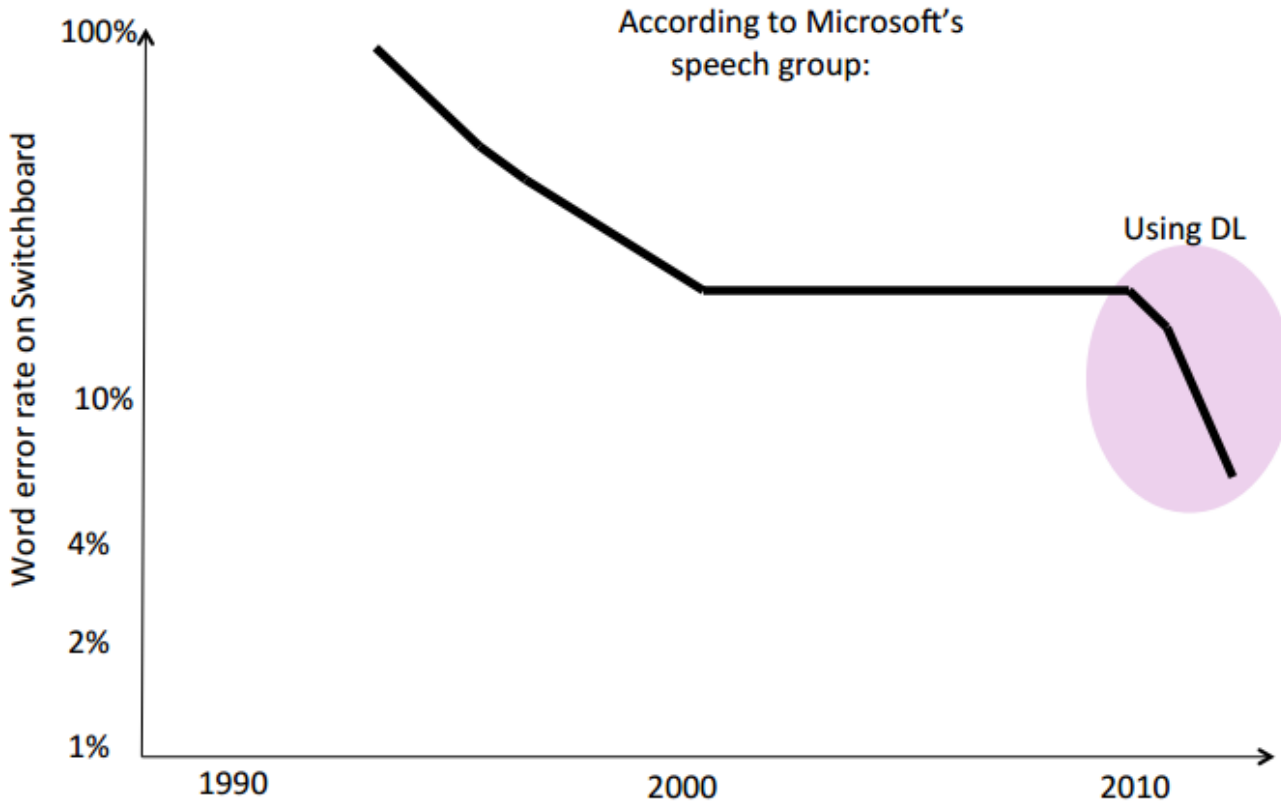
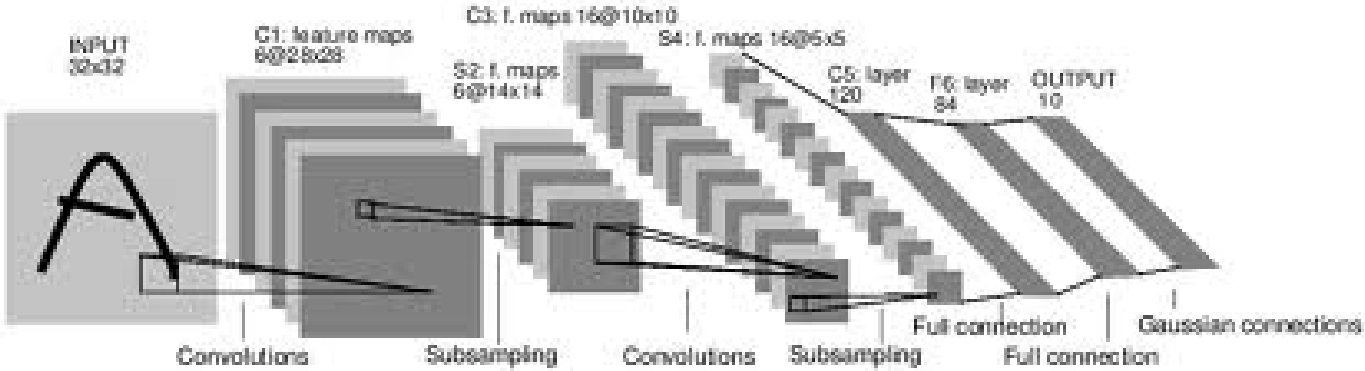


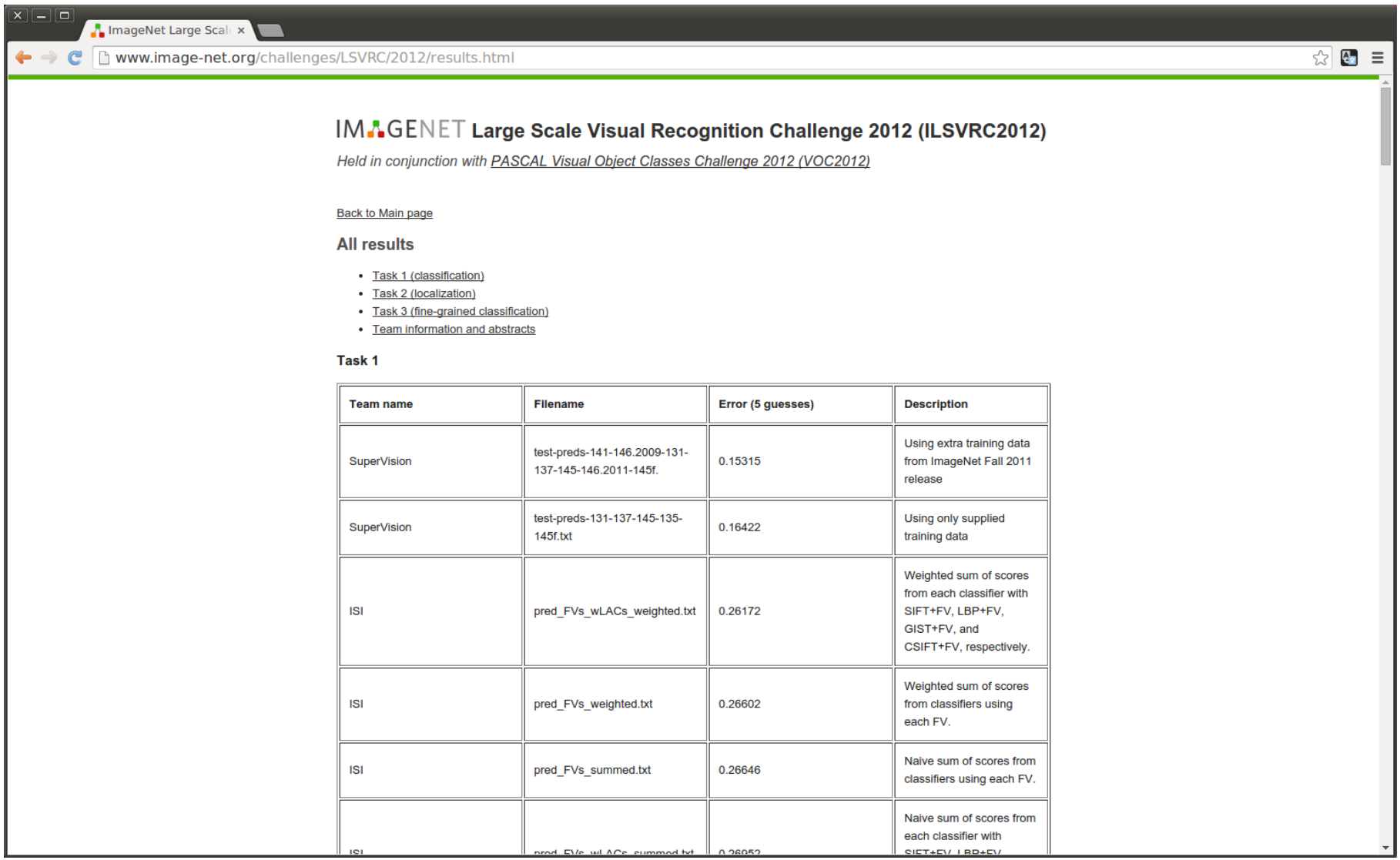
figure from Yoshua Bengio

Convolutional networks (CNN)



- LeCun et al. 1998

ImageNet challenge 2012



ImageNet Large Scal ×

www.image-net.org/challenges/LSVRC/2012/results.html

IMAGENET **Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)**

Held in conjunction with PASCAL Visual Object Classes Challenge 2012 (VOC2012)

Back to Main page

All results

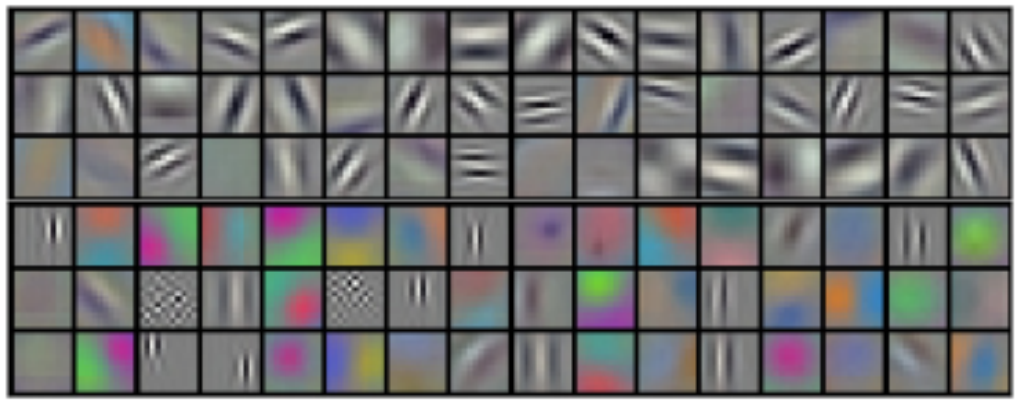
- Task 1 (classification)
- Task 2 (localization)
- Task 3 (fine-grained classification)
- Team information and abstracts

Task 1

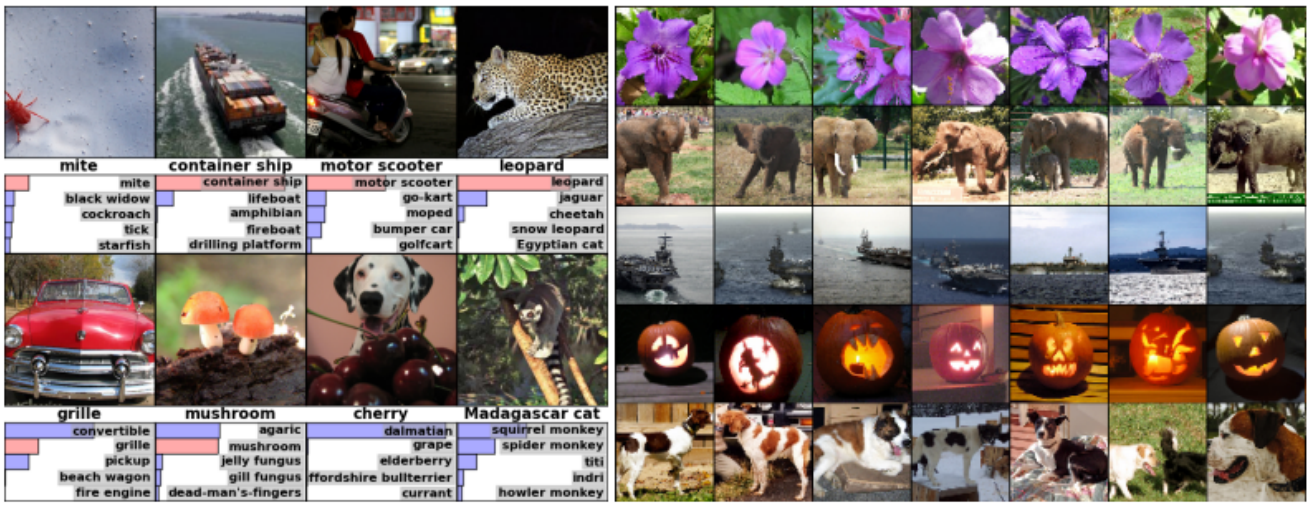
Team name	Filename	Error (5 guesses)	Description
SuperVision	test-preds-141-146.2009-131-137-145-146.2011-145f.	0.15315	Using extra training data from ImageNet Fall 2011 release
SuperVision	test-preds-131-137-145-135-145f.txt	0.16422	Using only supplied training data
ISI	pred_FVs_wLACs_weighted.txt	0.26172	Weighted sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.
ISI	pred_FVs_weighted.txt	0.26602	Weighted sum of scores from classifiers using each FV.
ISI	pred_FVs_summed.txt	0.26646	Naive sum of scores from classifiers using each FV.
ISI	pred_FVs_wLACs_summed.txt	0.26952	Naive sum of scores from each classifier with SIFT+FV, LBP+FV

ImageNet challenge 2012

some first-layer features

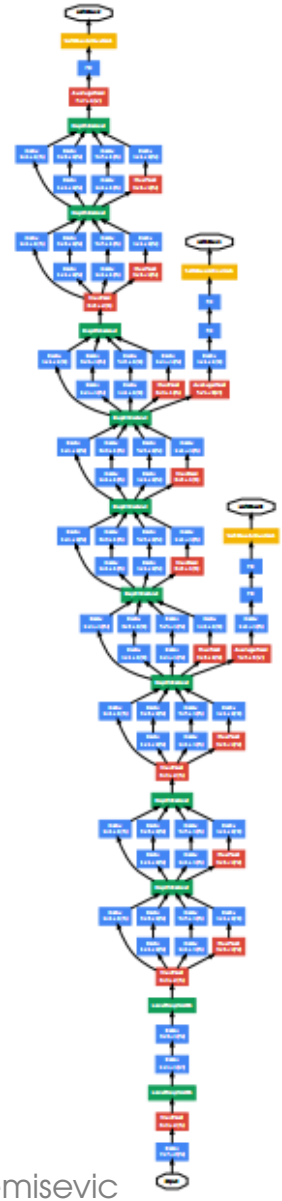


some results



Krizhevsky, Sutskever, Hinton; 2012

GoogLeNet (Szegedy et al. 2014)



- Exercise in (a) scaling up, (b) unconventional architectures
- Won ImageNet 2014 with **6.66%** top-5 error rate
- A variation of this network including BatchNormalization (Ioffe, Szegedy, 2015) achieves **4.8%** top-5 error rate, surpassing the accuracy of human raters

Where to get the labelled data?

High-level features

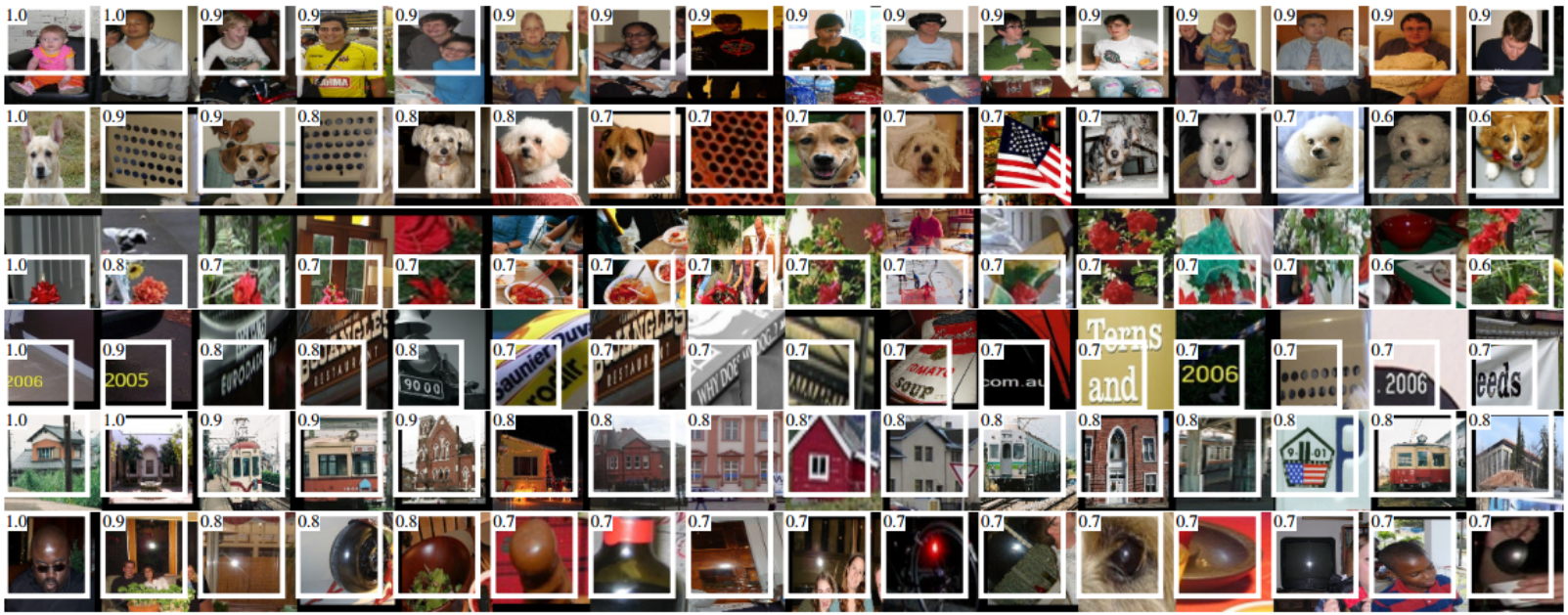
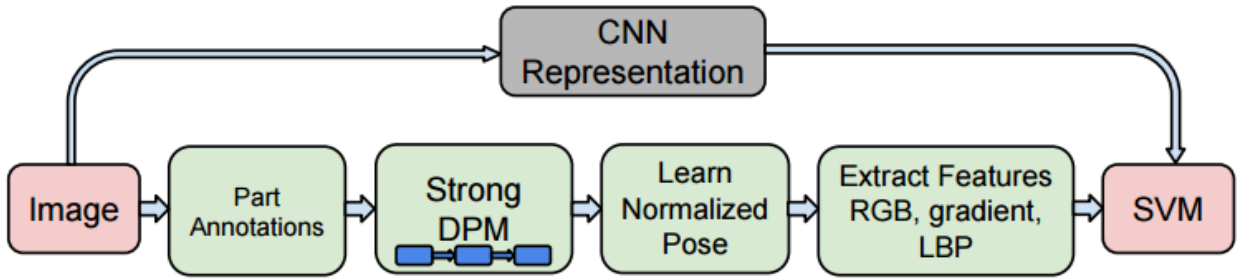
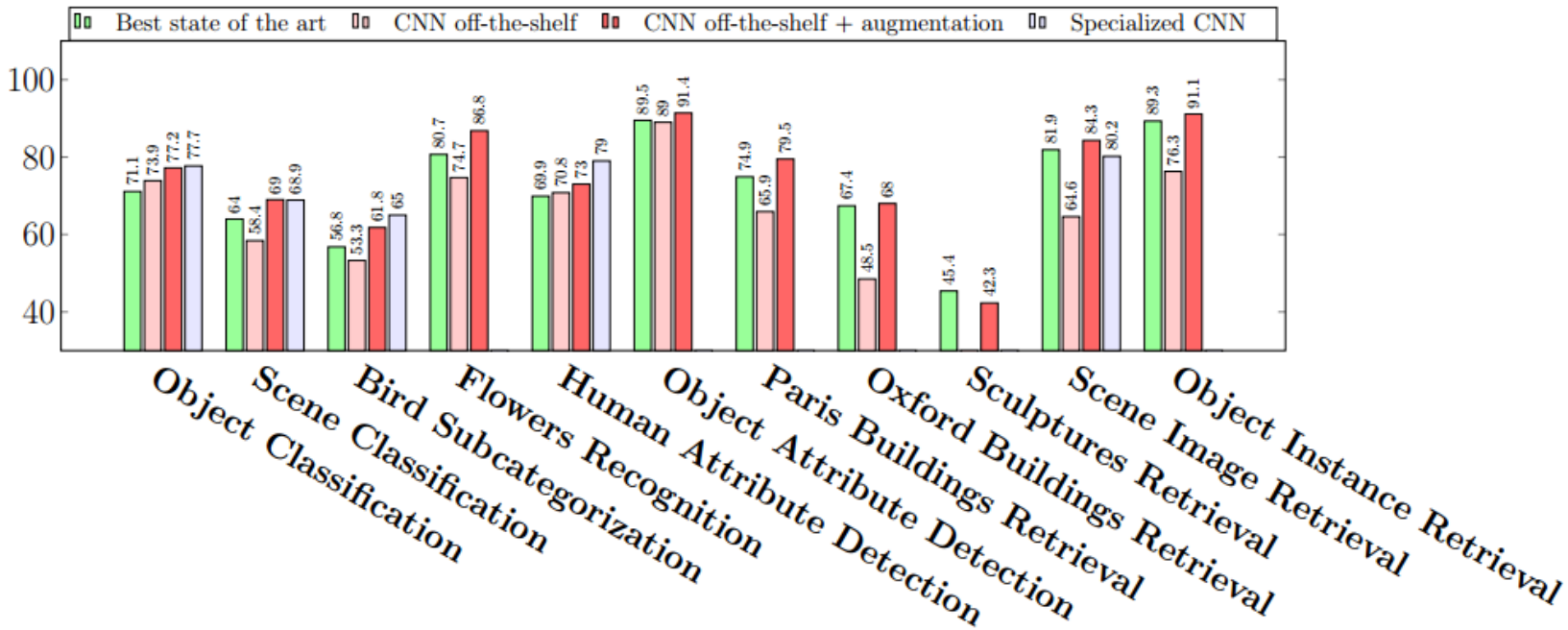


Figure 4: Top regions for six pool$_5$ units. Receptive fields and activation values are drawn in white. Some units are aligned to concepts, such as people (row 1) or text (4). Other units capture texture and material properties, such as dot arrays (2) and specular reflections (6).

Girshick et al., 2014

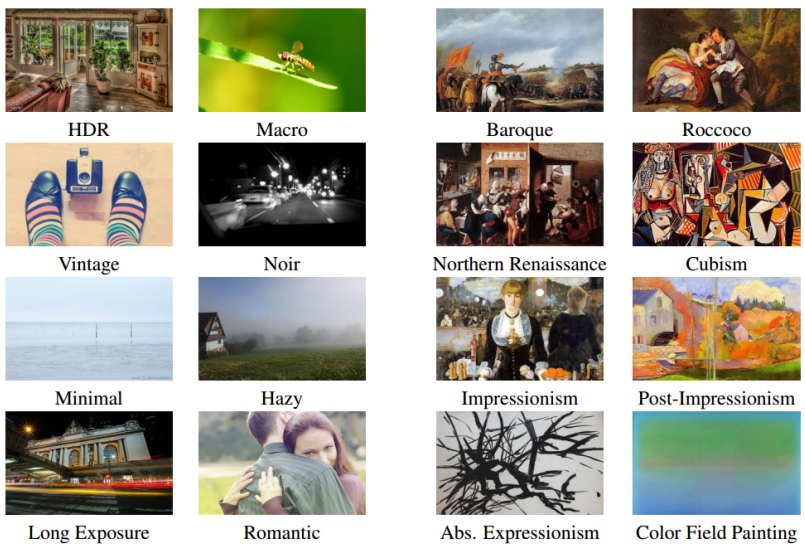
Conv-nets learn good *generic* features





(Razavian, Azizpour, Sullivan, Carlsson; 2014)

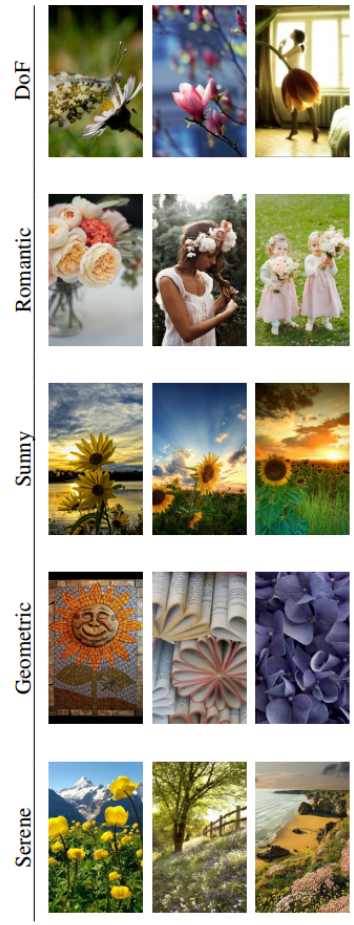
Karayev et al 2014: Recognizing Image Style



Flickr Style: 80K images covering 20 styles.

Wikipaintings: 85K images for 25 art genres.

Query: "dress".



Query: "flower".

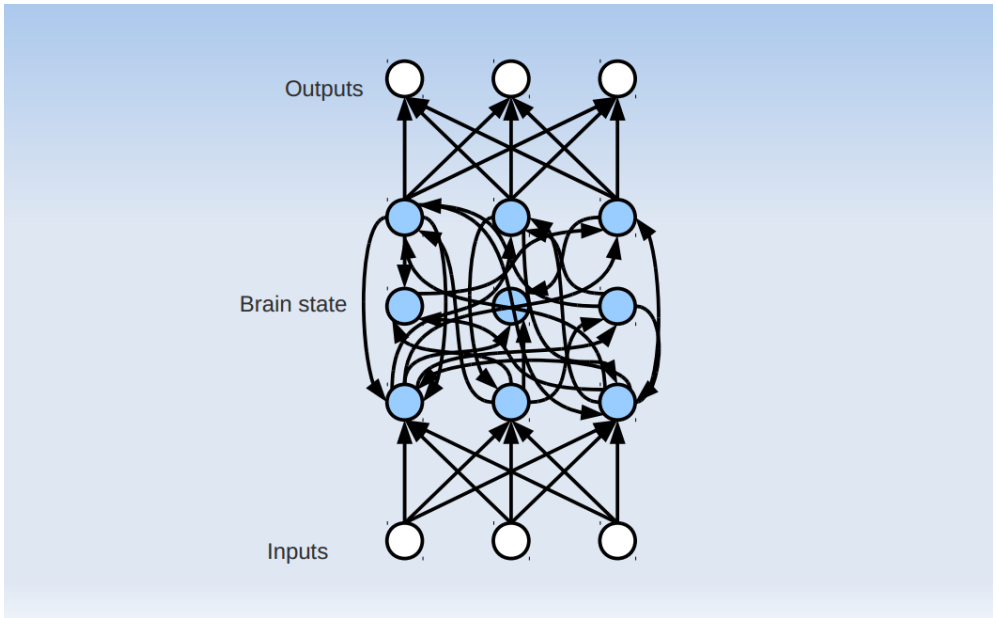
The disappearance of the data dilemma

- The success of generic imagenet-features shows that **a task** may be a better way to create features for other tasks (transfer learning) than **unlabelled data** (unsupervised learning) would.
- **Pooling** together many datasets and training one model to solve many task helps deal with the data scarcity, too. This is also how children get many labels.
- It resolves an old AI mystery called **grounding**.

There is no reason not to harness imagenet (and many other datasets) to improve, say, machine translation.

Where to get all the tasks?

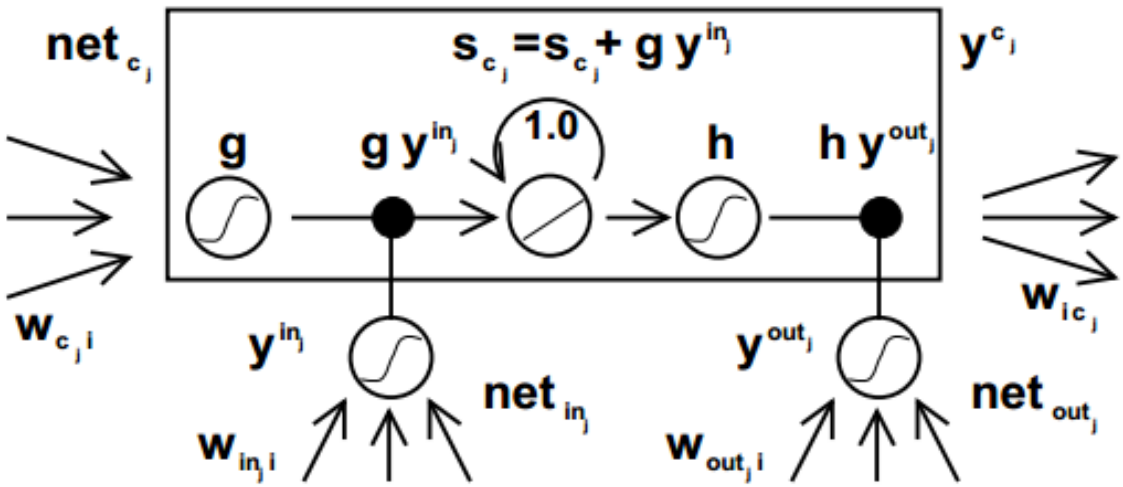
Recurrent networks (RNN)



picture from http://www.cs.toronto.edu/ asamir/cifar/Ilya_slides.pdf

- Stepping the network T time steps yields the equivalent of a T-layer feedforward net with weights that are shared between layers.
- Training the network by unrolling it in time is called back-prop-through-time (BPTT).

Long-Short Term Memory (LSTM)



(Hochreiter, Schmidthuber; 1997)

Generating text

```
Naturalism and decision for the majority of Arab countries' capitalide was grounded
by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated
with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal
in the [[Protestant Immineners]], which could be said to be directly in Cantonese
Communication, which followed a ceremony and set inspired prison, training. The
emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom
of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known
in western [[Scotland]], near Italy to the conquest of India with the conflict.
Copyright was the succession of independence in the slop of Syrian influence that
was a famous German movement based on a more popular servicious, non-doctrinal
and sexual power post. Many governments recognize the military housing of the
[[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]],
that is sympathetic to be to the [[Punjab Resolution]]
(PJS)[http://www.humah.yahoo.com/guardian.
cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery
was swear to advance to the resources for those Socialism's rule,
was starting to signing a major tripad of aid exile.]]
```

from: Andrej Kaparthy:
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Generating text

```
/*
 * 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;
}
```

from: Andrej Kaparthy:
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Generating text

For $\bigoplus_{n=1,\ldots,m}$ 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 compariciy 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.

Lemma 0.1. *Assume (3) and (3) by the construction in the description.*

Suppose $X = \lim |X|$ *(by the formal open covering* X *and a single map* $\underline{Proj}_X(\mathcal{A}) = \mathrm{Spec}(B)$ *over* U *compatible with the complex*

$$Set(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \to \mathcal{C}_{Z/X}$ *is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition ?? (without element is when the closed subschemes are catenary. If* T *is surjective we may assume that* T *is connected with residue fields of* S*. Moreover there exists a closed subspace* $Z \subset X$ *of* X *where* U *in* X' *is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem*

(1) f *is locally of finite type. Since* $S = \mathrm{Spec}(R)$ *and* $Y = \mathrm{Spec}(R)$.

Proof. This is form all sheaves of sheaves on X. But given a scheme U and a surjective étale morphism $U \to X$. Let $U \cap U = \coprod_{i=1,\ldots,n} U_i$ be the scheme X over S at the schemes $X_i \to X$ and $U = \lim_i X_i$. □

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X},\ldots,0}$.

Lemma 0.2. *Let* X *be a locally Noetherian scheme over* S, $E = \mathcal{F}_{X/S}$. *Set* $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}'_n$. *Since* $\mathcal{I}^n \subset \mathcal{I}^n$ *are nonzero over* $i_0 \leq \mathfrak{p}$ *is a subset of* $\mathcal{J}_{n,0} \circ \overline{A}_2$ *works.*

Lemma 0.3. *In Situation ??. Hence we may assume* $\mathfrak{q}' = 0$.

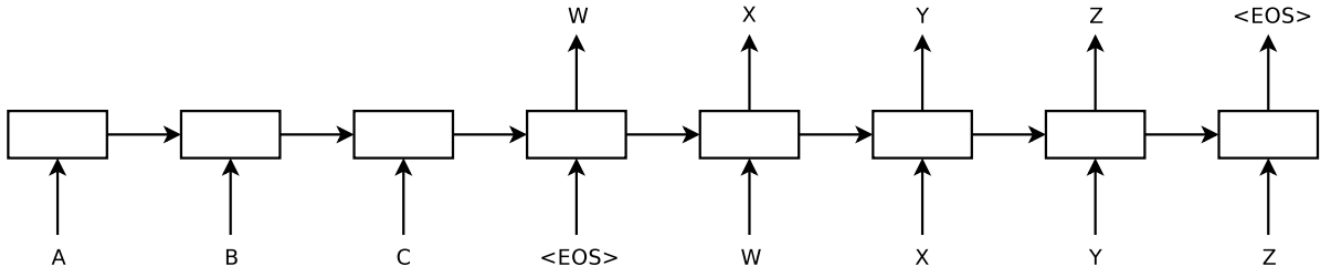
Proof. We will use the property we see that $\mathfrak{p}$ is the mext functor (??). On the other hand, by Lemma ?? we see that

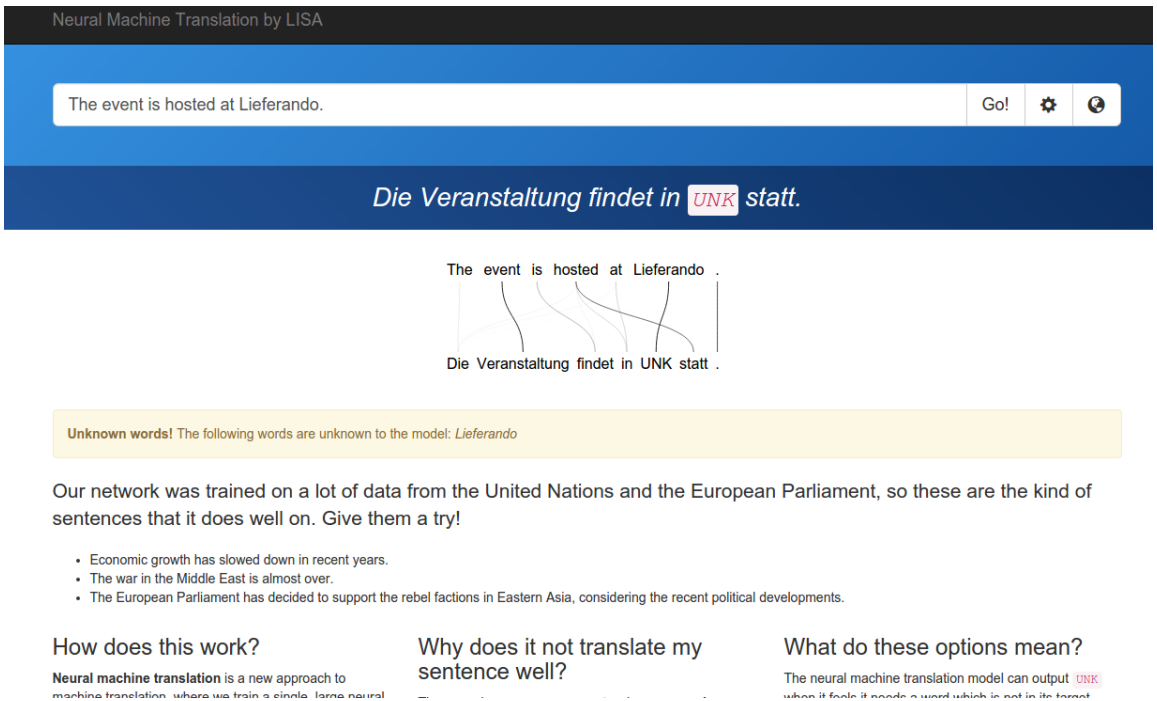
$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

where K is an F-algebra where δ_{n+1} is a scheme over S. □

from: Andrej Kaparthy:
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

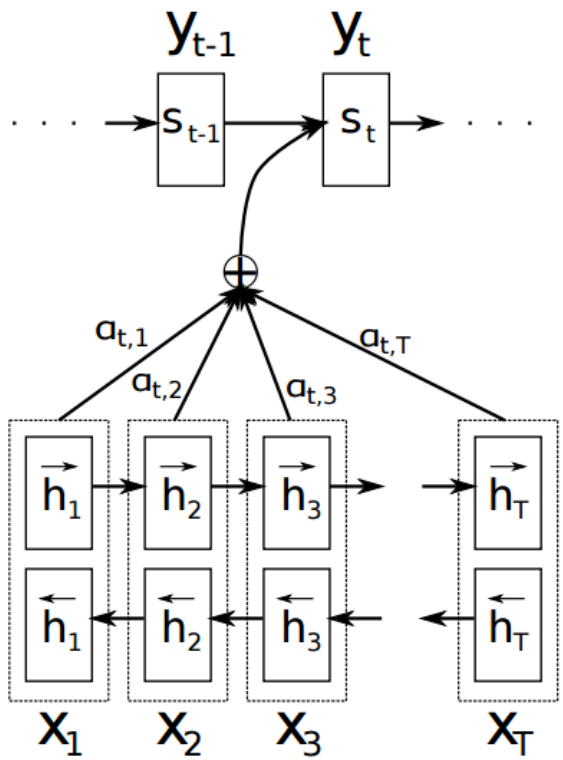
Encoder-decoder for machine translation





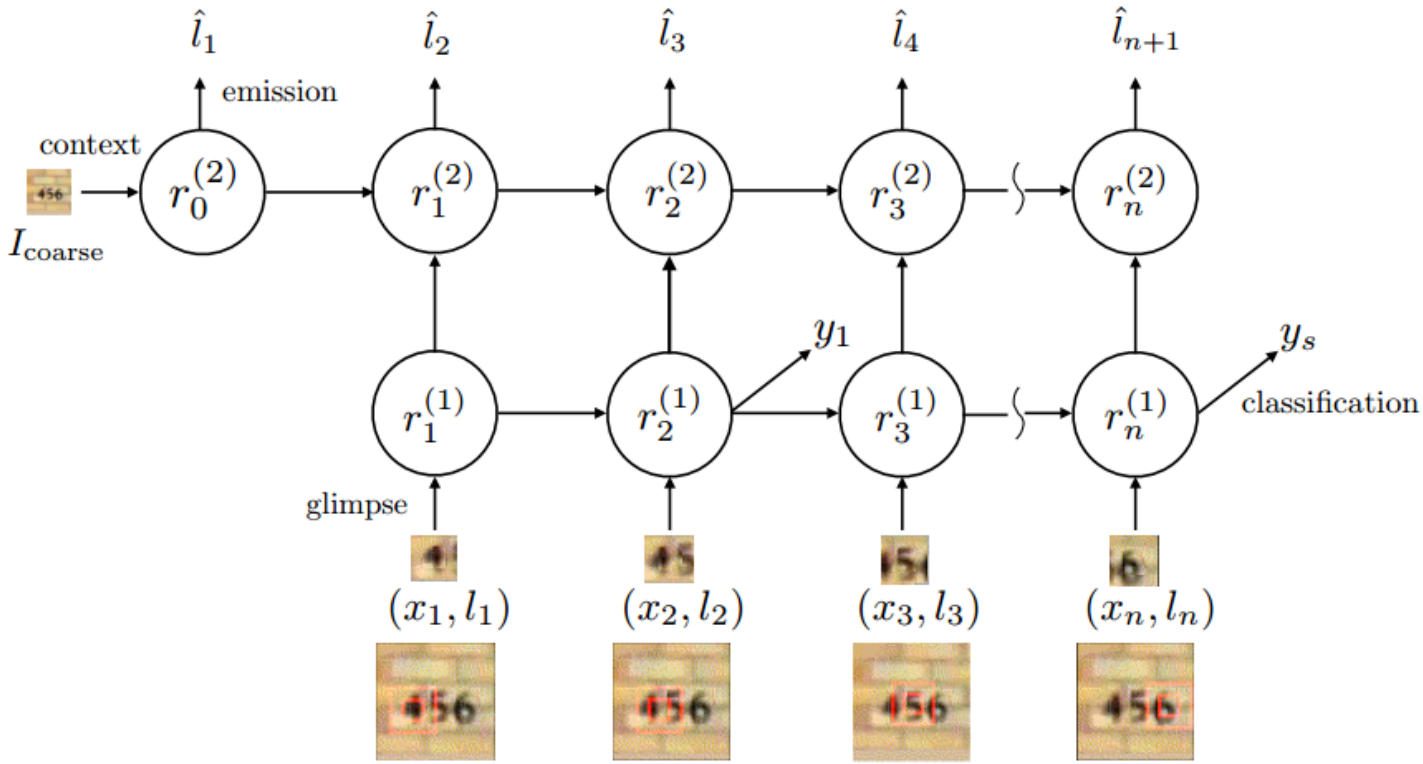
Sutskever et al. NIPS 2014; Bahdanau et al. 2014

(Soft) attention in the encoder-decoder model



Bahdanau et al. 2014

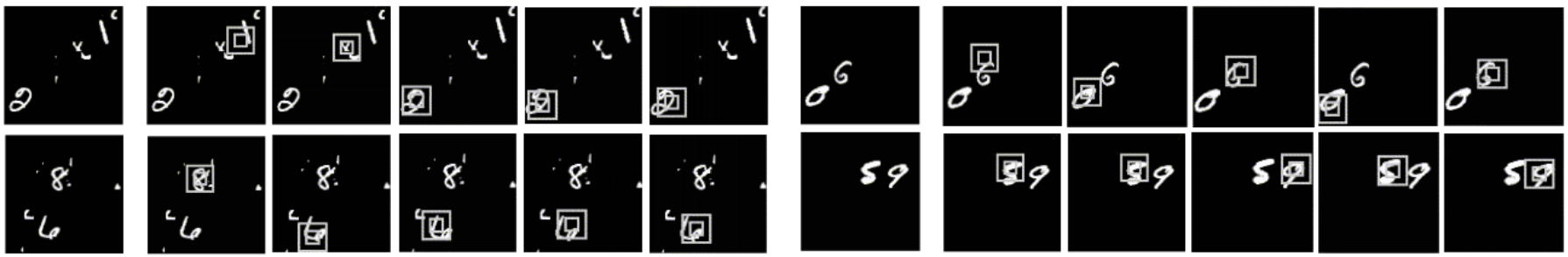
Visual (hard) attention



Ba et al. 2015, Mnih et al. 2013

- Back-prop through the gaze-decision not possible! $\rightarrow$ Use sampling

Visual (hard) attention



SVHN:

Model	Test Err.
11 layer CNN Goodfellow et al. (2013)	3.96%
10 layer CNN	4.11%
Single DRAM	5.1%
Single DRAM MC avg.	4.4%
forward-backward DRAM MC avg.	**3.9%**

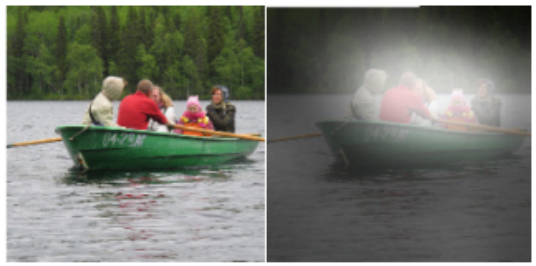
Visual (soft) attention (Xu et al 2015)

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

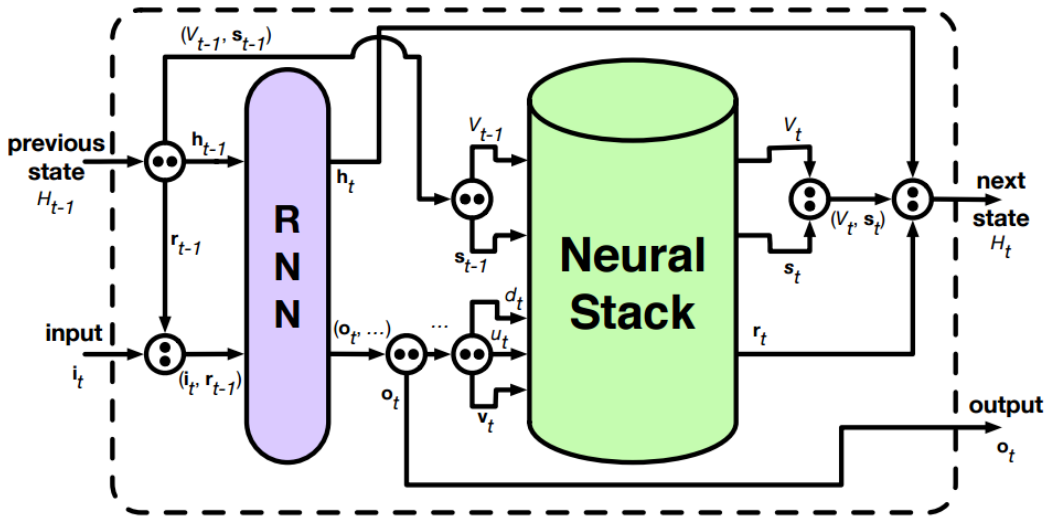


A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Differentiable models of computation

- Neural Turing Machine (Graves et al, 2014)
- Memory Networks (Weston et al, 2014)
- Learning to Transduce with Unbounded Memory (Grefenstette et al. 2015)



- Base everything on differentiable operations, or use sampling

Learning to execute (Zaremba, Sutskever; 2014)

Input:

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

Target: 25011.

Input:

```
i=8827
c=(i-5347)
print((c+8704) if 2641<8500 else 5308)
```

Target: 12184.

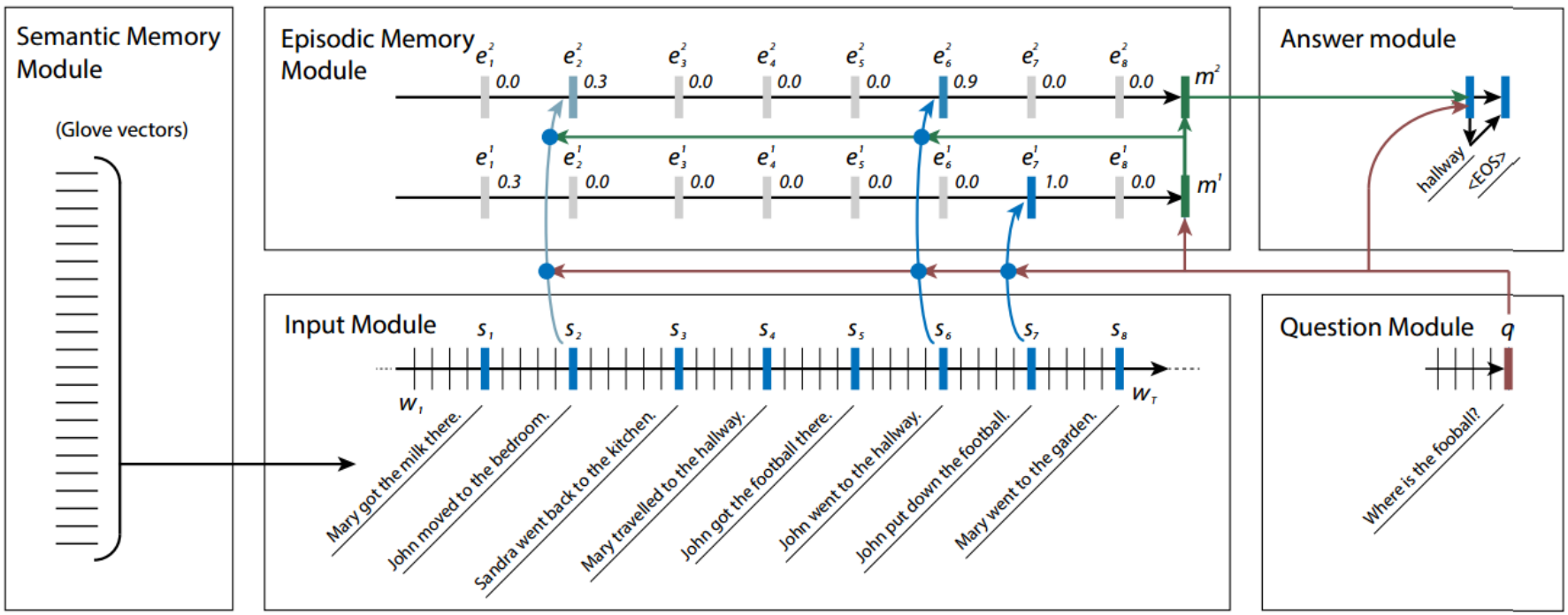
From neural networks to “neural programs”

I:	Jane went to the hallway.
I:	Mary walked to the bathroom.
I:	Sandra went to the garden.
I:	Daniel went back to the garden.
I:	Sandra took the milk there.
Q:	Where is the milk?
A:	garden

I:	Everybody is happy.
Q:	What’s the sentiment?
A:	positive
Q:	What are the POS tags?
A:	NN VBZ JJ .
I:	The answer is far from obvious.
Q:	In French?
A:	La réponse est loin d’être évidente.

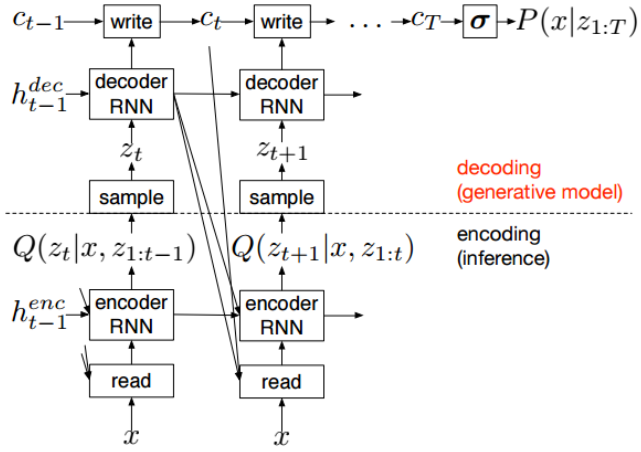
Kumar et al 2015

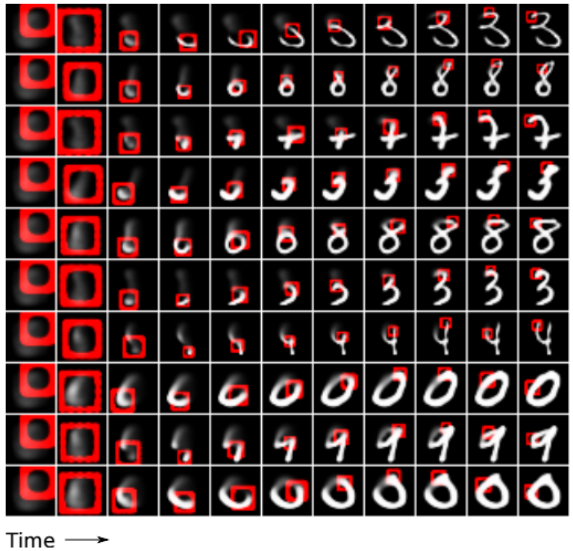
From neural networks to “neural programs”

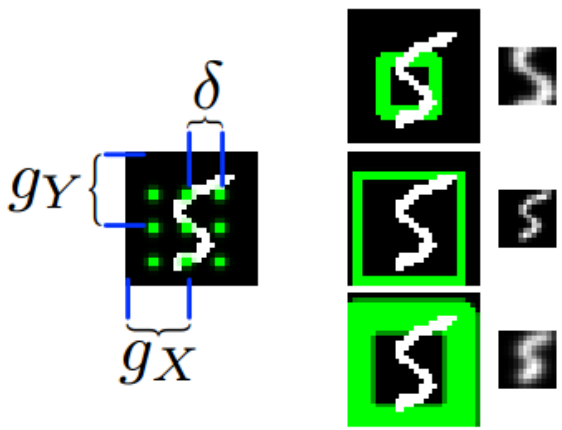


Kumar et al 2015

Operating on a buffer (Gregor et al., 2015)







Drawing on a buffer using a pre-trained convnet

Gatys, Ecker, Bethge (2015)

- Optimize pixels so as to

  (i) match the hidden layer activations of the content image

  (ii) match a non-linear function of the hidden layer activations of the style image

- This is DRAW in disguise!
- (And it works so well mainly because it uses a pre-trained convnet!)

Some neural program applications

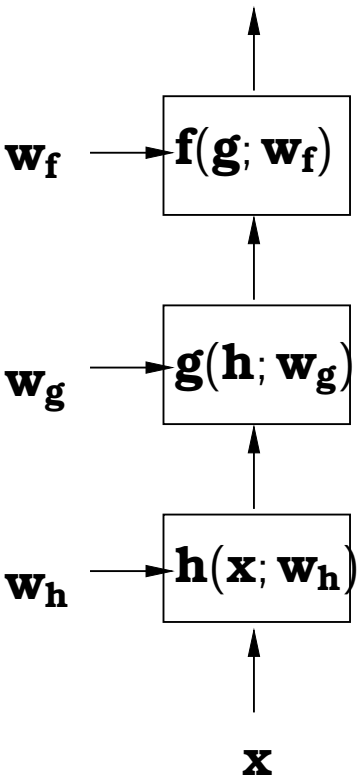
- Drawing/painting
- Generate, *then edit*, text (eg. for translation)
- Generating mocap sequences
- Speed up complicated models or ensembles (Bucila et al 2006)
- Speed up routines (eg. Esmaeilzadeh et al. 2012)

Von Neumann via Deep Learning

- In the past, we simulated neural nets on classic hardware and it didn't work
- Today, we simulate classic hardware on neural nets and it works beautifully

The benefit of today's way: Add parallelization and your "program" may run faster and faster and faster...

Deep Learning as a compute paradigm



- solve a task by performing a *series of operations*
- use *learning* to define the computations
- make the constituent computations *parallel*

Other deep learning directions

- Scaling up, **hardware**
- Applications
- **Data-pooling**/grounding/transfer learning: use one model to solve many tasks
- Reinforcement learning
- Architectures/circuitry/**neural programs**

Thank you

Questions?

www.twentybn.com