DLSG #3 - Gradient Descent Algos + Intro to ConvNets

DLSG

November 11, 2016

Agenda

1. Admin
2. Gradient Descent Algos
3. ConvNets

Housekeeping

- Numerous new members!
- Does the mailing list actually work?
- New website!
  - Most of the materials there
  - One GitHub repo
  - News as PRs?

Big Schedule Review

- Introduction to Keras
- (Artificial) Neural Networks and their training

- Convolutional Neural Netowrks
- Recurrent Neural Networks
- Neural Networks in Computer Vision
- Neural Networks in Natural Language Processing

News

- Character Sequence Models for ColorfulWords (`http://colorlab.us`)

News

- Character Sequence Models for ColorfulWords (`http://colorlab.us`)
- Deep Learning 'ahem' detector (`https://github.com/worldofpiggy/deeplearning-ahem-detector`)

News

- Character Sequence Models for ColorfulWords (`http://colorlab.us`)
- Deep Learning 'ahem' detector (`https://github.com/worldofpiggy/deeplearning-ahem-detector`)
- Judging a Book By its Cover (`https://arxiv.org/abs/1610.09204`)
- LipNet: Sentence-Level Lipreading (`http://openreview.net/forum?id=BkjLkSqxg`)

News

- Character Sequence Models for ColorfulWords (http://colorlab.us)
- Deep Learning 'ahem' detector (https://github.com/worldofpiggy/deeplearning-ahem-detector)
- Judging a Book By its Cover (https://arxiv.org/abs/1610.09204)
- LipNet: Sentence-Level Lipreading (http://openreview.net/forum?id=BkjLkSqxg)
  *93% accuracy at reading sentences of the form "set|lay|place <colour> in|on|with <letter> <number> soon|again".

News

- Character Sequence Models for ColorfulWords (http://colorlab.us)
- Deep Learning 'ahem' detector (https://github.com/worldofpiggy/deeplearning-ahem-detector)
- Judging a Book By its Cover (https://arxiv.org/abs/1610.09204)
- LipNet: Sentence-Level Lipreading (http://openreview.net/forum?id=BkjLkSqxg)
  *93% accuracy at reading sentences of the form "set|lay|place <colour> in|on|with <letter> <number> soon|again".
- ...

Overview

- Gradient Descent Variants
- Memntum
- Nesterov Accelerated Gradient
- Adagrad
- RMSProp
- Notes, Tips & Tricks

Gradient Descent Variants

Core idea of gradient descent

Minimize $J(\theta)$ parametrized by $\theta \in \mathbb{R}^d$ by updating θ in the opposite direction of the gradient $\nabla_\theta J(\theta)$.

Batch Gradient Descent

aka Vanilla Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_{\theta_t} J(\theta_t)$$

- Might be very slow
- No-go for big datasets
- Impossible to update "online" (new examples on-the-fly)
- Guaranteed to converge to the global minimum for convex error surfaces and to a local minimum for non-convex surfaces

```
for i in range(nb_epochs):
  params_grad = evaluate_gradient(loss_function, data, params)
  params = params - learning_rate * params_grad
```

Stochastic Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_{\theta_t} J(\theta_t; x^{(i)}; y^{(i)})$$

- Usually faster convergence
- Where batch gradient descent does redundant computation, SGD updates frequently and creates fluctiations.
- When slowly decreasing the learning rate, SGD shows the same convergence behaviour as batch gradient descent

```
for i in range(nb_epochs):
  np.random.shuffle(data)
  for example in data:
    params_grad = evaluate_gradient(loss_function, example, params)
    params = params - learning_rate * params_grad
```

Mini-batch Gradient Descent

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i:i+n)}; y^{(i:i+n)})$$

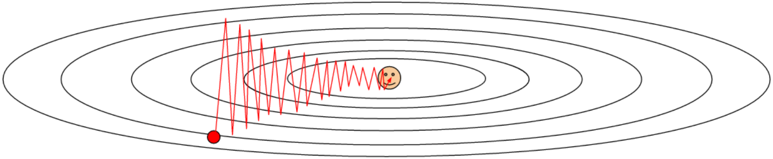
- Best of both worlds
- Reduced variance of parameter updates – more stable convergence
- Batch size usually $n = [50, 250]$
- SGD and Mini-batch Gradient Descent are used interchangebly

```
for i in range(nb_epochs):
  np.random.shuffle(data)
  for batch in get_batches(data, batch_size=50):
    params_grad = evaluate_gradient(loss_function, batch, params)
    params = params - learning_rate * params_grad
```

Gradient Descent – Challenges

- Choosing a proper learning rate is difficult (too small, too large, too steady...)
- Learning rate schedules help, but still need to be pre-defined in advance
- Same learning rate for all parameter updates (larger updates to more infrequent features might be more desirable)
- Ending up trapped in suboptimal local optima

Stochastic Gradient Descent



Momentum

$$v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta)$$

$$\theta_{t+1} = \theta_t - v_t$$

- Helps navigate SGD when one dimension curves more steeply than than the other (common around local optima)
- Basically fights against oscillations
- Momentum term γ is usually set to 0.9
- "Pushing a ball down a hill" metaphor

Nesterov Accelerated Gradient

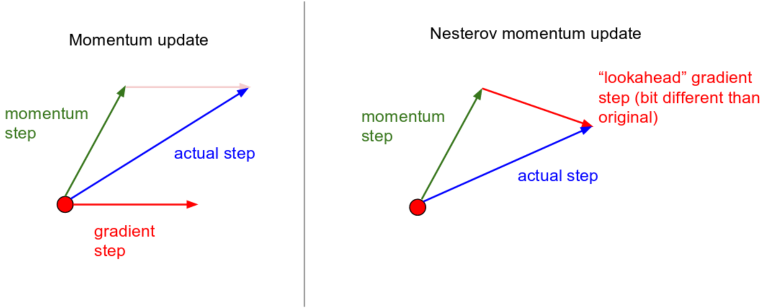
Let's not blindly trust gravity

$$v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta - \gamma v_{t-1})$$

$$\theta_{t+1} = \theta_t - v_t$$

- Give the moving ball some notion of where it is going
- $\theta - \gamma v_{t-1}$ approximates (gives a rough idea of) the next position of the parameters
- "Update with anticipation" prevents the ball from going too fast
- Significantly improved performance of RNNs on numerous tasks
- Is able to adapt updates to the slope – we'd like to also adapt updates to "parameter importance"

Nesterov Momentum



Adagrad

$$g_{t,i} = \nabla_\theta J(\theta_i)$$

$$\theta_{t+1,i} = \theta_{t,i} - \eta \cdot g_{t,i}$$

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i}$$

- $G_t \in \mathbb{R}^{d \times d}$ – diagonal metrix where $G_{t,ii}$ is the sum of the squares of the gradients w.r.t θ_i up to time t.
- ϵ helps to avoid division-by-zero issues (usually on the order of $1e-8$)
- Main benefit: no need for tuning the learning rate
- Main weakness: accumulation of squared gradients in the denominator
- Learning rate will shrink (sometimes way too much)

RMSProp

RMSProp update

[Tieleman and Hinton, 2012]

```
# Adagrad update
cache += dx**2
x += - learning_rate * dx / (np.sqrt(cache) + 1e-7)
```

```
# RMSProp
cache = decay_rate * cache + (1 - decay_rate) * dx**2
x += - learning_rate * dx / (np.sqrt(cache) + 1e-7)
```

RMSProp

rmsprop: A mini-batch version of rprop

- rprop is equivalent to using the gradient but also dividing by the size of the gradient.
  - The problem with mini-batch rprop is that we divide by a different number for each mini-batch. So why not force the number we divide by to be very similar for adjacent mini-batches?
- rmsprop: Keep a moving average of the squared gradient for each weight

$$MeanSquare(w,t) = 0.9\ MeanSquare(w,\ t-1) + 0.1\left(\partial E / \partial w\,(t)\right)^2$$

- Dividing the gradient by $\sqrt{MeanSquare(w,\ t)}$ makes the learning work much better (Tijmen Tieleman, unpublished).

Introduced in a slide in Geoff Hinton's Coursera class, lecture 6

Cited by several papers as:

[52] T. Tieleman and G. E. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude., 2012.

Adam

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

Both m_t and v_t are initialized as 0s, so they need to be bias-corrected.

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$$

Adam

Adam update

[Kingma and Ba, 2014]

```
# Adam
m,v = #... initialize caches to zeros
for t in xrange(1, big_number):
  dx = # ... evaluate gradient
  m = beta1*m + (1-beta1)*dx # update first moment
  v = beta2*v + (1-beta2)*(dx**2) # update second moment
  mb = m/(1-beta1**t) # correct bias
  vb = v/(1-beta2**t) # correct bias
  x += - learning_rate * mb / (np.sqrt(vb) + 1e-7)
```

momentum

bias correction
(only relevant in first few iterations when t is small)

RMSProp-like

The bias correction compensates for the fact that m,v are initialized at zero and need some time to “warm up”.

Visual Demo

So what should one use?

- RMSProp and Adam are very similar
- Bias-correction in Adam has been shown to outpreform RMSProp slightly towards the end
- **Adam** is usually a good default choice for CNNs, RMSProp might be worth considering for big RNNs

Other improvements to consider

- Shuffling and Curriculum Learning (pretty big result with LSTMs)
- Early stopping & Ensemble averaging
- Gradient noise

$$g_{t,i} = g_{t,i} + N(0, \sigma_t^2)$$

$$\sigma_t^2 = \frac{\eta}{(1+t)^\gamma}$$

Makes networks more robust to poor initialization – authors 'suspect' the added noise gives model more chances to escape and find local minima.

ConvNets

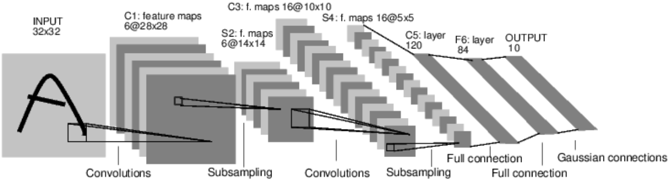
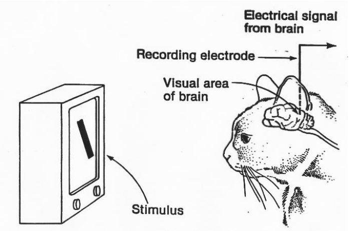
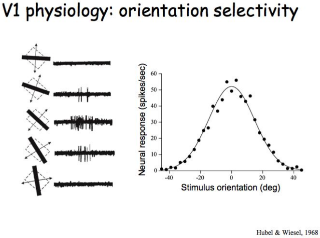


Figure: LeNet [LeCun et al., 1998]

Hubel & Wiesel

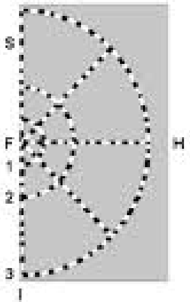
1959 - Receptive fields of single neurones in the cat's striate cortex
1962 - Receptive fields, binocular interaction and functional architecture in the cat's visual cortex

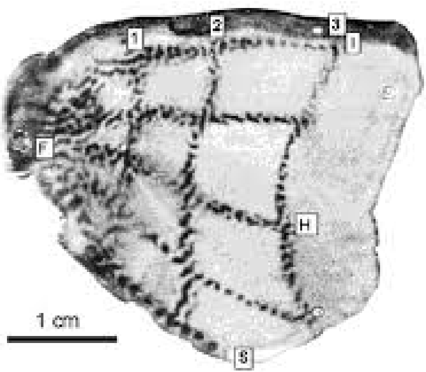




A bit of history

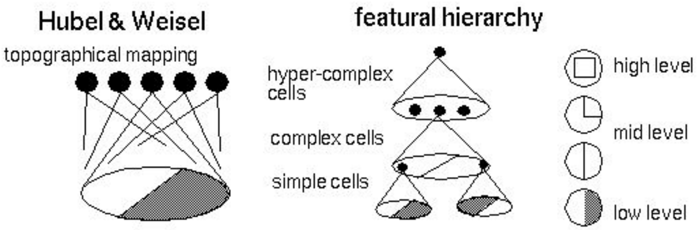
Topographical mapping in the cortex: nearby cells in cortex represented nearby regions in the visual field





A bit of history

Hierarchical organization



A bit of history

Neurocognitron [Fukushima 1980]: "sandwich" architecture (SCSCSC...) simple cells: modifiable parameters complex cells: perform pooling

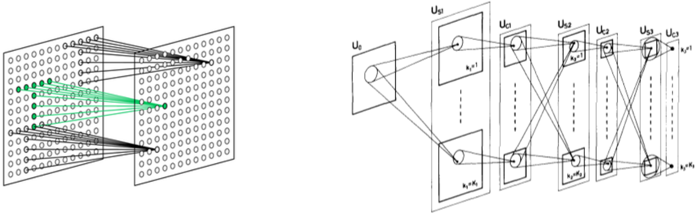


Figure: Neurocognitron [Fukushima 1980]

LeNet

Gradient-based learning applied to document recognition

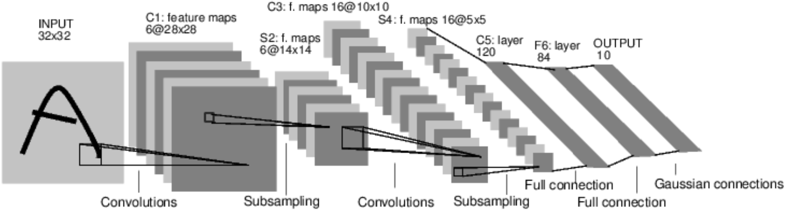


Figure: LeNet [LeCun, Bottou, Bengio, Haffner 1998]

AlexNet

ImageNet Classification with Deep Convolutional Neural Networks

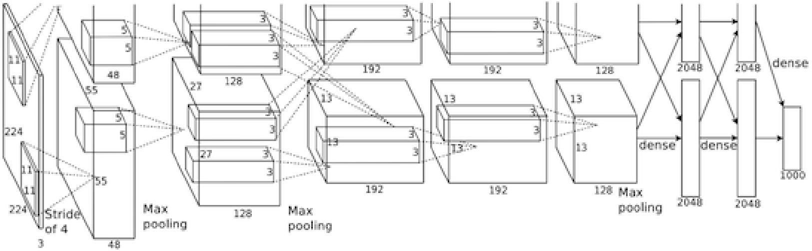


Figure: AlexNet [Krizhevsky, Sutskever, Hinton, 2012]

ResNet

Depth Revolution

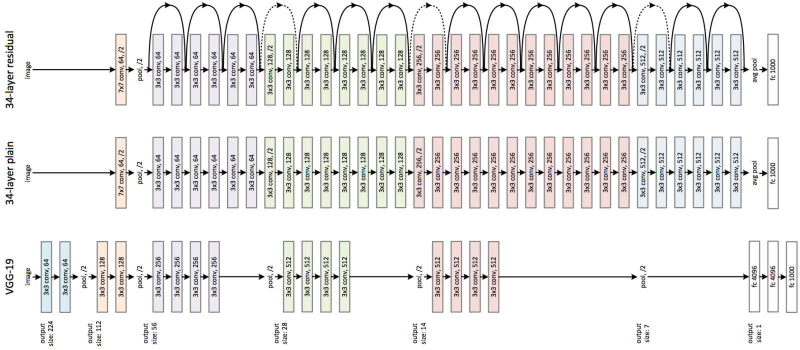


Figure: ResNet [Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015]

ConvNets today

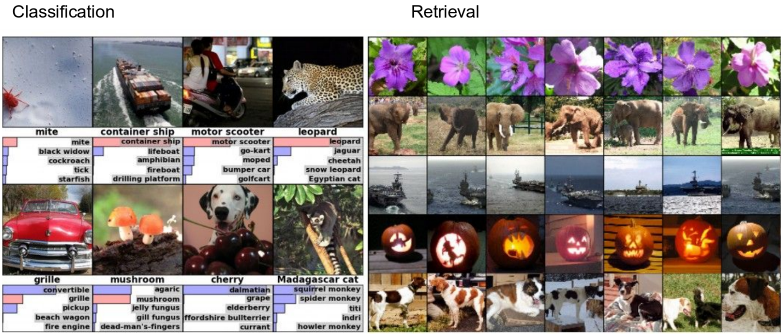


Figure: [Krizhevsky 2012]

ConvNets today

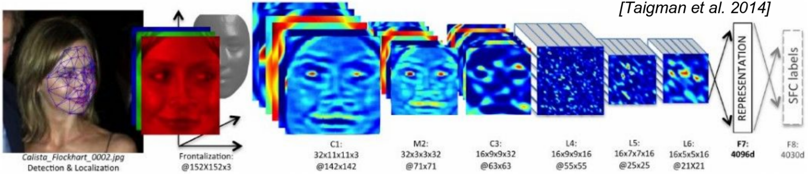
Figure: [Faster R-CNN: Ren, He, Girshick, Sun 2015] Detection Segmentation & [Farabet et al., 2012]

ConvNets today

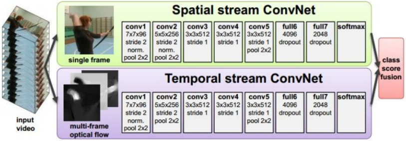
NVIDIA Tegra X1

Figure: Self driving cars

ConvNets today



[Taigman et al. 2014]



[Simonyan et al. 2014]

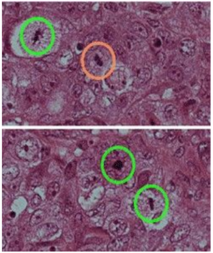
[Goodfellow 2014]

ConvNets today

[Toshev, Szegedy 2014]

[Mnih 2013]

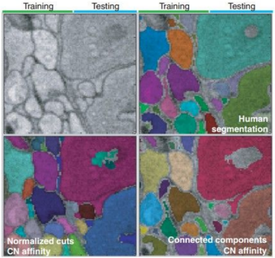
ConvNets today



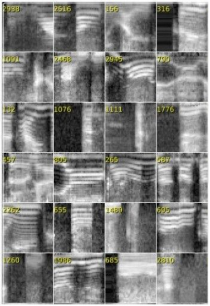
[Ciresan et al. 2013]

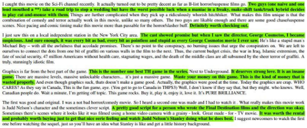
[Sermanet et al. 2011]
[Ciresan et al.]

ConvNets today



[Turaga et al., 2010]





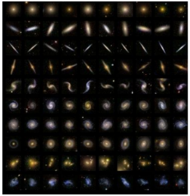
I caught this movie on the Sci-Fi channel recently. It actually turned out to be pretty decent as far as B-list horror/suspense films go. **Two guys (one naive and one loud mouthed a **) take a road trip to stop a wedding but have the worst possible luck when a maniac in a freaky, make-shift tank/truck hybrid decides to play cat-and-mouse with them.** Things are further complicated when they pick up a ridiculously whorish hitchhiker. What makes this film unique is that the combination of comedy and terror actually work in this movie, unlike so many others. The two guys are likable enough and there are some good chase/suspense scenes. Nice pacing and comic timing make this movie more than passable for the horror/slasher buff. **Definitely worth checking out.**

I just saw this on a local independent station in the New York City area. **The cast showed promise but when I saw the director, George Cosmotos, I became suspicious. And sure enough, it was every bit as bad, every bit as pointless and stupid as every George Cosmotos movie I ever saw.** He's like a stupid man's Michael Bey – with all the awfulness that accolade promises. There's no point to the conspiracy, no burning issues that urge the conspirators on. We are left to ourselves to connect the dots from one bit of graffiti on various walls in the film to the next. Thus, the current budget crisis, the war in Iraq, Islamic extremism, the fate of social security, 47 million Americans without health care, stagnating wages, and the death of the middle class are all subsumed by the sheer terror of graffiti. A truly, stunningly idiotic film.

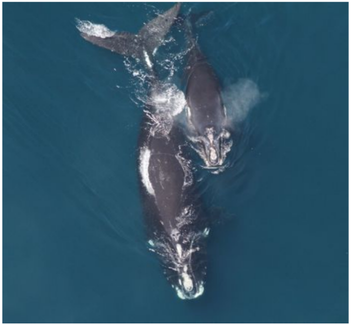
Graphics is far from the best part of the game. **This is the number one best TH game in the series.** Next to Underground. **It deserves strong love. It is an insane game.** There are massive levels, massive unlockable characters... it's just a massive game. **Waste your money on this game. This is the kind of money that is wasted properly.** And even though graphics suck, thats doesn't make a game good. Actually, the graphics were good at the time. Today the graphics are crap. WHO CARES? As they say in Canada, This is the fun game, aye. (You get to go to Canada in THPS3) Well, I don't know if they say that, but they might. who knows. Well, Canadian people do. Wait a minute, I'm getting off topic. This game rocks. Buy it, play it, enjoy it, love it. It's PURE BRILLIANCE.

The first was good and original. I was a not bad horror/comedy movie. So I heard a second one was made and I had to watch it . What really makes this movie work is Judd Nelson's character and the sometimes clever script. **A pretty good script for a person who wrote the Final Destination films and the direction was okay.** Sometimes there's scenes where it looks like it was filmed using a home video camera with a grainy - look. Great made - for - TV movie. **It was worth the rental and probably worth buying just to get that nice eerie feeling and watch Judd Nelson's Stanley doing what he does best.** I suggest newcomers to watch the first one before watching the sequel, just so you'll have an idea what Stanley is like and get a little history background.

[Denil et al. 2014]



ConvNets today



Whale recognition, Kaggle Challenge

Mnih and Hinton, 2010

ConvNets today

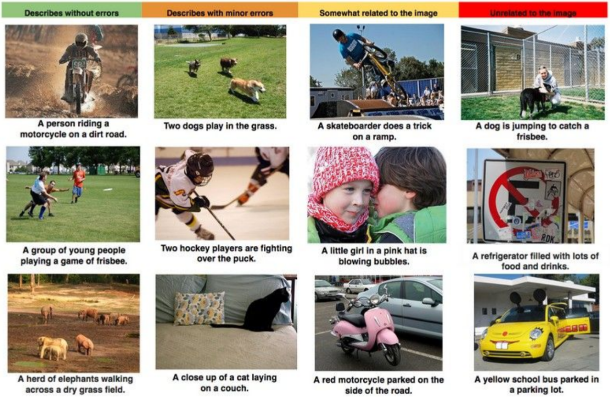
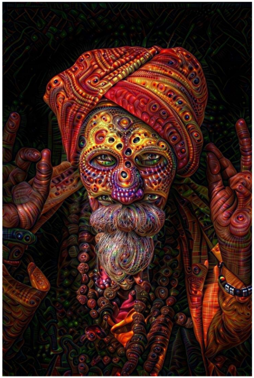
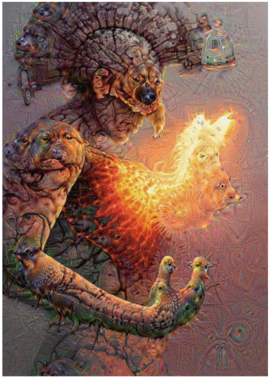
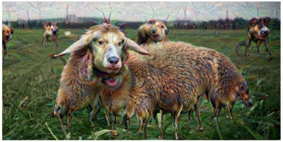


Image Captioning

[Vinyals et al., 2015]

ConvNets today



reddit.com/r/deepdream

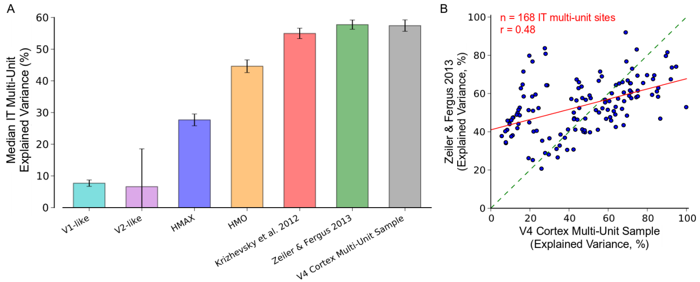


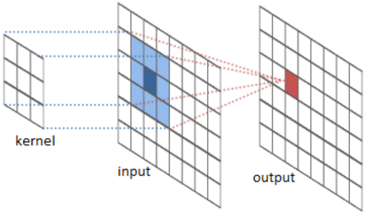
Figure: Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition [Cadieu et al., 2014]

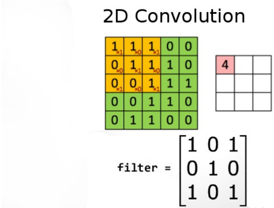
Convolution

So what is that "convolution"

Convolution

So what is that "convolution"





Convolution

$$\sum_{i=1}^{n} x_i w_i$$

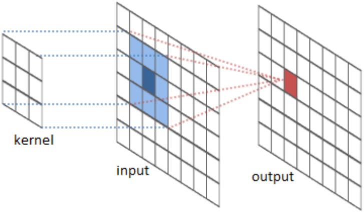
Convolution

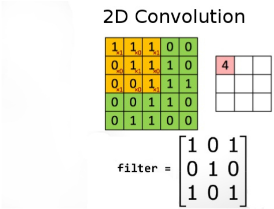
$$\sum_{i=1}^{n} x_i w_i = X \cdot W$$

Convolution

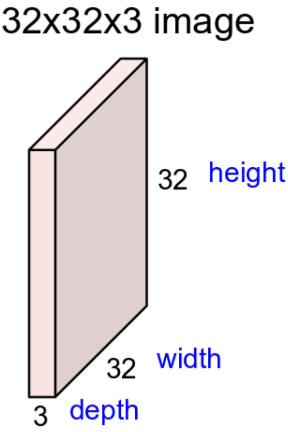
$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f(m)g(n-m)$$

Convolution

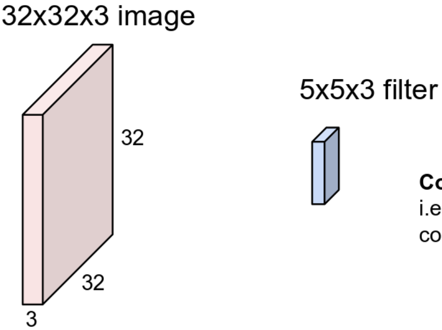




Convolution layer

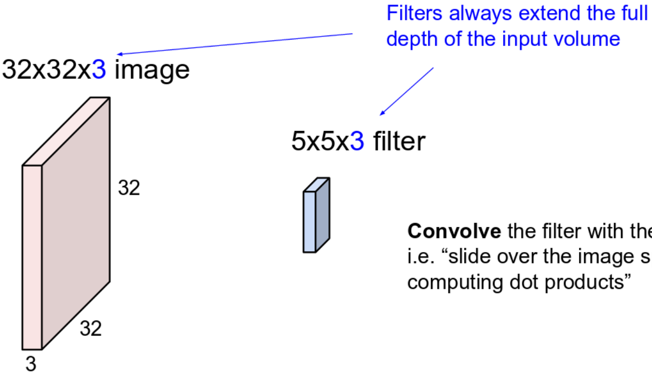


Convolution layer



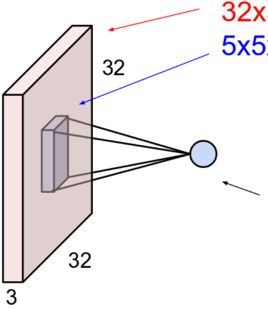
Convolve the filter with the image
i.e. "slide over the image spatially,
computing dot products"

Convolution layer



Convolve the filter with the image
i.e. "slide over the image spatially, computing dot products"

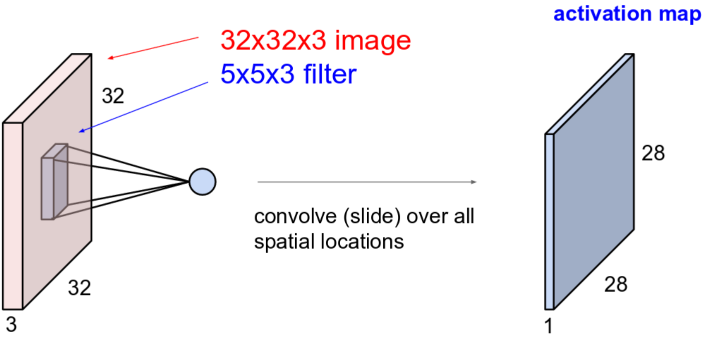
Convolution layer



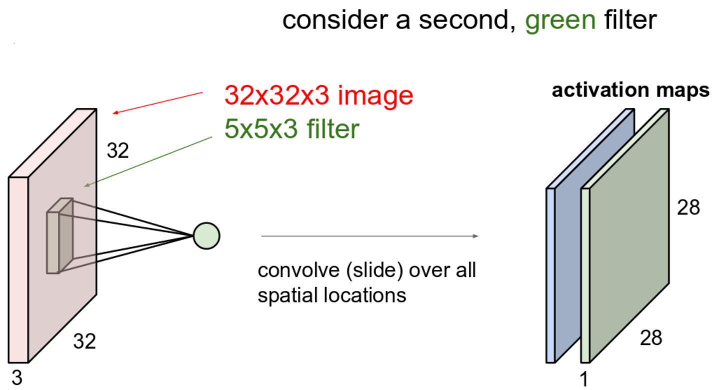
1 number:
the result of taking a dot product between the filter and a small 5x5x3 chunk of the image
(i.e. 5*5*3 = 75-dimensional dot product + bias)

$$w^T x + b$$

Convolution layer

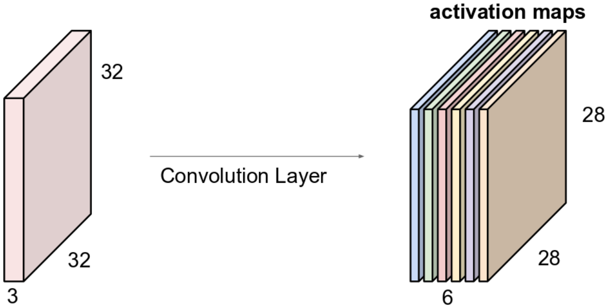


Convolution layer



Convolution layer

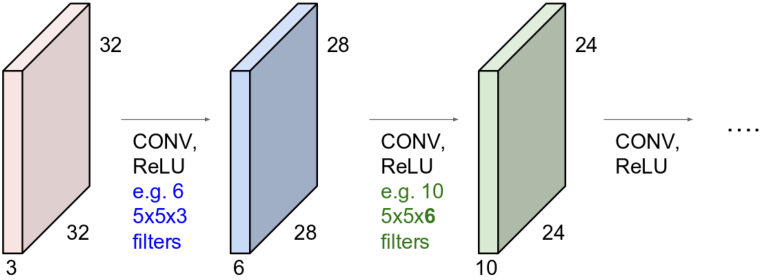
For example, if we had 6 5x5 filters, we'll get 6 separate activation maps:



We stack these up to get a "new image" of size 28x28x6!

Convolution layer

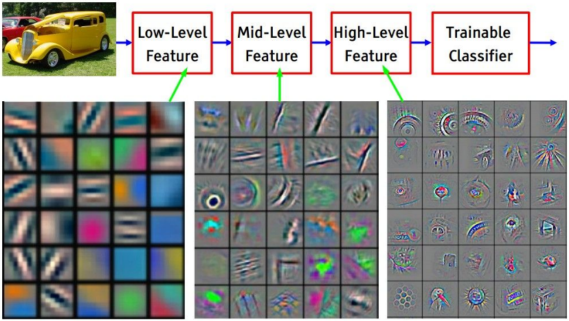
Preview: ConvNet is a sequence of Convolutional Layers, interspersed with activation functions



Convolution layer

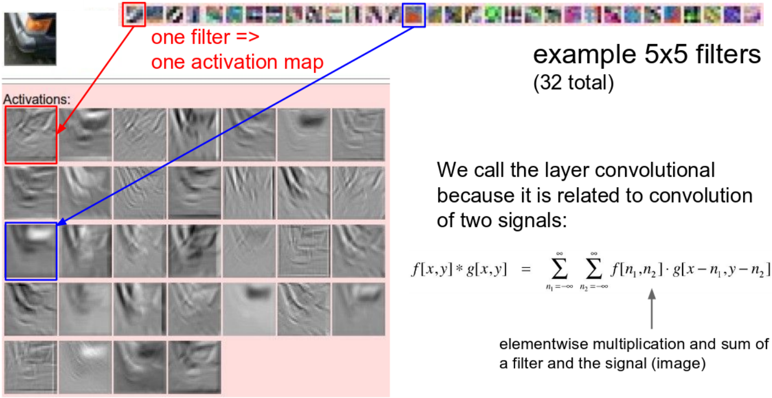
Preview

[From recent Yann LeCun slides]



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

Convolution layer

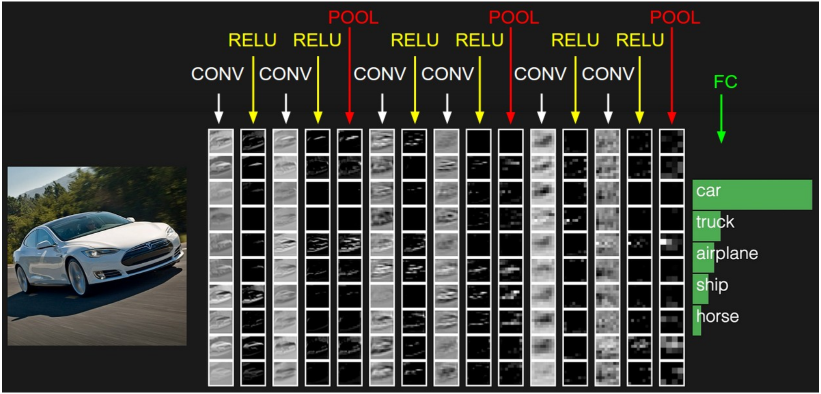


We call the layer convolutional because it is related to convolution of two signals:

$$f[x,y] * g[x,y] = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f[n_1,n_2] \cdot g[x-n_1, y-n_2]$$

elementwise multiplication and sum of a filter and the signal (image)

Convolutional Neural Network



References

- `cs231n.stanford.edu/slides/winter1516_lecture6.pdf`
- `cs231n.stanford.edu/slides/winter1516_lecture7.pdf`
- `cs231n.github.io/`
- IRC server *freenode* - channel *#naiveneuron*