EECS 4422/5323 Computer Vision

Image Understanding 2

Calden Wloka

7 October, 2019

Announcments

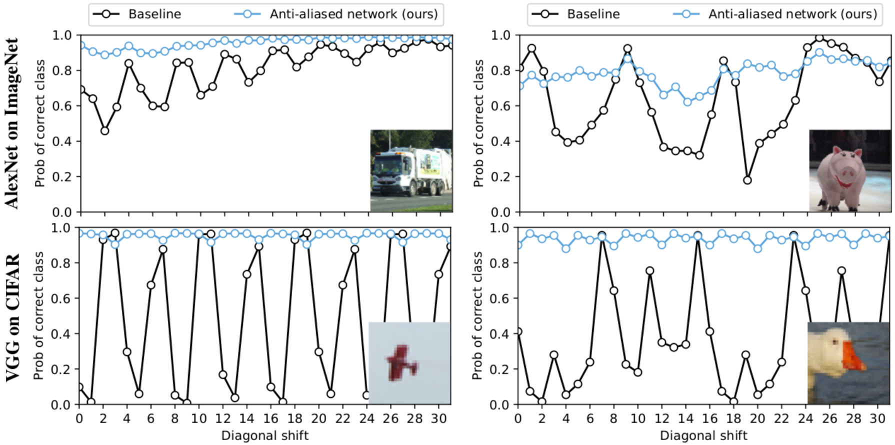
- Lab today TBD
- Solutions to Assignment 1 to be posted Wednesday
- If you did not collect your white paper or did not turn one in, please talk to me ASAP

Outline

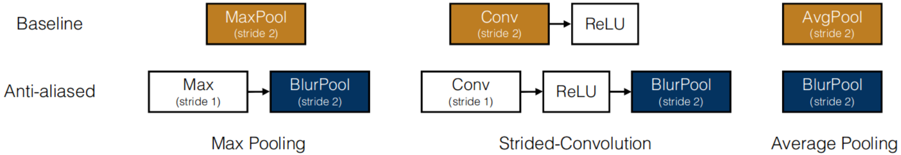
- Shift Invariance in CNNs
- Hierarchical Processing
- Ramifications
- Attention

Shift Invariance in CNNs

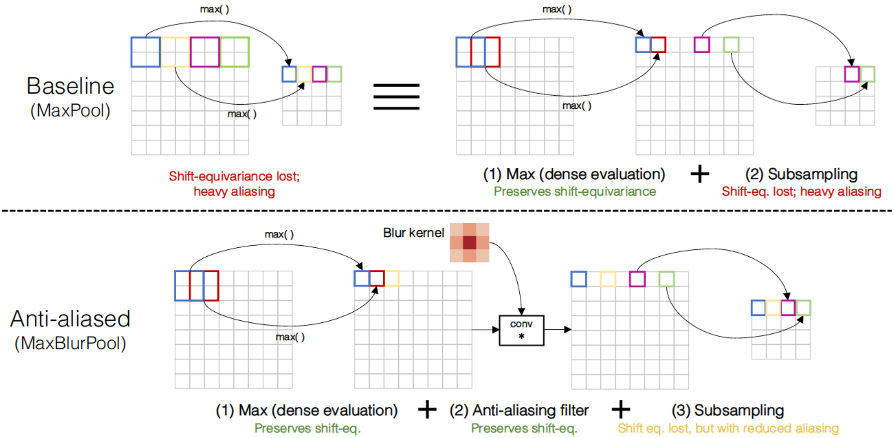
The paper I mentioned last class:
Zhang, "Making Convolution Networks Shift-Invariant Again", 2019



The issue seems to be max-pooling and downsampling - mitigate with an extra filter step



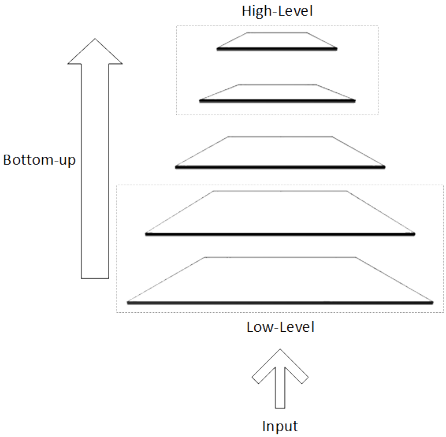
Example for Max-Pooling



Feed-Forward Processing

The style of information flow in a standard CNN architecture consists of *feed-forward processing*.

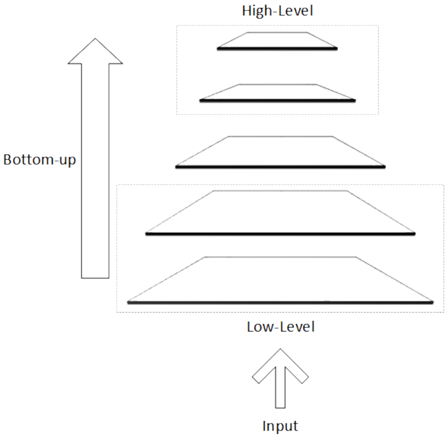
- Information flows in a *bottom-up* fashion



Feed-Forward Processing

The style of information flow in a standard CNN architecture consists of *feed-forward processing*.

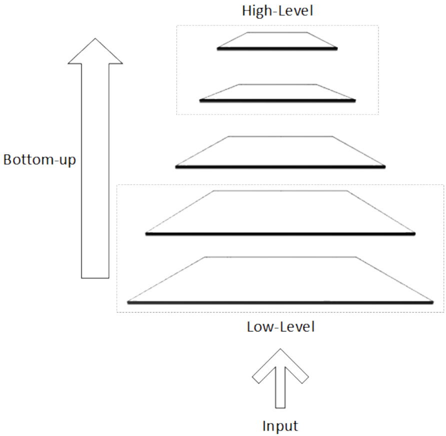
- Information flows in a *bottom-up* fashion
- Low-level features are closer to the input



Feed-Forward Processing

The style of information flow in a standard CNN architecture consists of *feed-forward processing*.

- Information flows in a *bottom-up* fashion
- Low-level features are closer to the input
- High-level features occur near the top of the hierarchy



Challenges with Hierarchies

Much like the shift variance issue introduced by max-pooling downsampling, feedforward hierarchies face a number of general computational challenges.

- Boundary Problem

Note: the majority of the discussion in this section follows Chapter 2 of Tsotsos, 2011, *A Computational Perspective on Visual Attention*.

Challenges with Hierarchies

Much like the shift variance issue introduced by max-pooling downsampling, feedforward hierarchies face a number of general computational challenges.

- Boundary Problem
- Blurring

Note: the majority of the discussion in this section follows Chapter 2 of Tsotsos, 2011, *A Computational Perspective on Visual Attention*.

Challenges with Hierarchies

Much like the shift variance issue introduced by max-pooling downsampling, feedforward hierarchies face a number of general computational challenges.

- Boundary Problem
- Blurring
- Cross-talk Problem

Note: the majority of the discussion in this section follows Chapter 2 of Tsotsos, 2011, *A Computational Perspective on Visual Attention*.

Challenges with Hierarchies

Much like the shift variance issue introduced by max-pooling downsampling, feedforward hierarchies face a number of general computational challenges.

- Boundary Problem
- Blurring
- Cross-talk Problem
- Sampling

Note: the majority of the discussion in this section follows Chapter 2 of Tsotsos, 2011, *A Computational Perspective on Visual Attention*.

Challenges with Hierarchies

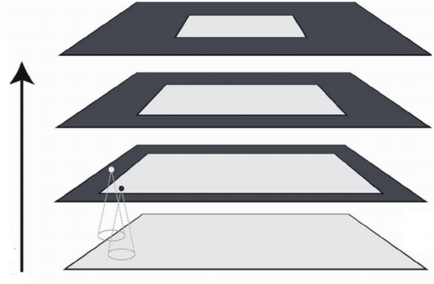
Much like the shift variance issue introduced by max-pooling downsampling, feedforward hierarchies face a number of general computational challenges.

- Boundary Problem
- Blurring
- Cross-talk Problem
- Sampling
- Context Problem

Note: the majority of the discussion in this section follows Chapter 2 of Tsotsos, 2011, *A Computational Perspective on Visual Attention*.

Each convolution layer loses information

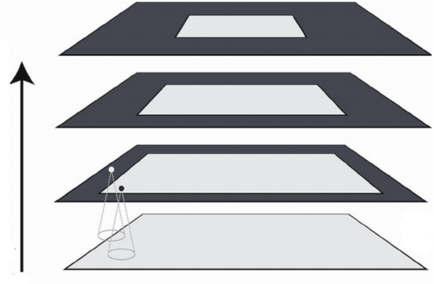
- Each increasing layer of the network loses the half-width of its precedent kernel in defined output



Source: Image modified from Tsotsos, 2011

Each convolution layer loses information

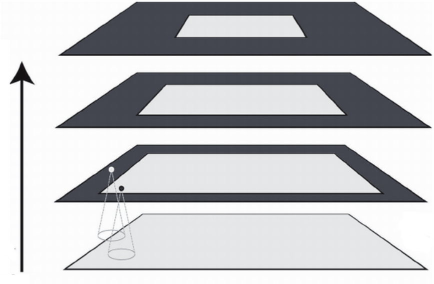
- Each increasing layer of the network loses the half-width of its precedent kernel in defined output
- Padding techniques each introduce their own potential artifacts which still propagates, even if the output stays the same size (demo)



Source: Image modified from Tsotsos, 2011

Each convolution layer loses information

- Each increasing layer of the network loses the half-width of its precedent kernel in defined output
- Padding techniques each introduce their own potential artifacts which still propagates, even if the output stays the same size (demo)
- The only real solution is to change viewpoints



Source: Image modified from Tsotsos, 2011

Active Vision

When an entity exerts control over the future perceptual input it acquires, that is termed *active perception*.

Active Vision

When an entity exerts control over the future perceptual input it acquires, that is termed *active perception*.

Many computer vision problem domains do not allow for active vision (*e.g*. any analysis of previously recorded image or video), but for any platform with motor capabilities (*e.g*. a security camera or a robot), active vision techniques can be very powerful.

Active Vision

When an entity exerts control over the future perceptual input it acquires, that is termed *active perception*.

Many computer vision problem domains do not allow for active vision (*e.g.* any analysis of previously recorded image or video), but for any platform with motor capabilities (*e.g.* a security camera or a robot), active vision techniques can be very powerful.

Choosing where to look next in an active vision system is highly non-trivial, and will vary greatly with task or the nature of the visual platform.

The Boundary Problem and CNNs

The boundary problem often does not seem like a big deal for CNNs due to a combination of:

The Boundary Problem and CNNs

The boundary problem often does not seem like a big deal for CNNs due to a combination of:

- CNNs typically use very small filters for computational reasons

The Boundary Problem and CNNs

The boundary problem often does not seem like a big deal for CNNs due to a combination of:

- CNNs typically use very small filters for computational reasons
- Most datasets have a compositional bias, with most images containing objects of interest near the image centre

The Boundary Problem and CNNs

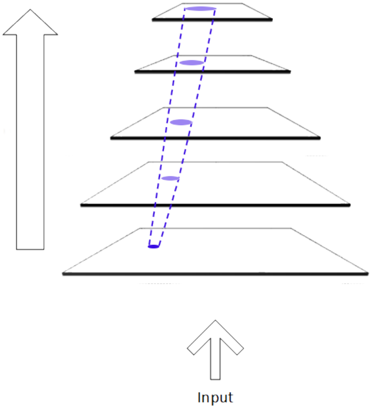
The boundary problem often does not seem like a big deal for CNNs due to a combination of:

- CNNs typically use very small filters for computational reasons
- Most datasets have a compositional bias, with most images containing objects of interest near the image centre

Nevertheless, the boundary problem does mean that for a given input size, there is a point at which increasing network depth will begin to suffer meaningfully from the boundary problem.

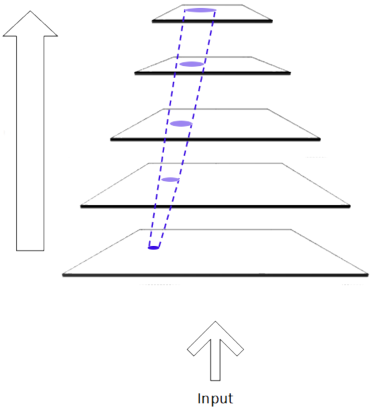
Information Gets "Smeared" Across the Hierarchy

- The influence of any given input element ends up affecting a diverging tree of activations as one moves up through the hierarchy



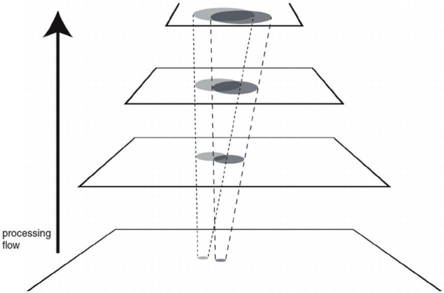
Information Gets “Smeared” Across the Hierarchy

- The influence of any given input element ends up affecting a diverging tree of activations as one moves up through the hierarchy
- Information which is well-localized at input is not well localized by the top of the hierarchy



Nearby Elements Mutually Interfere

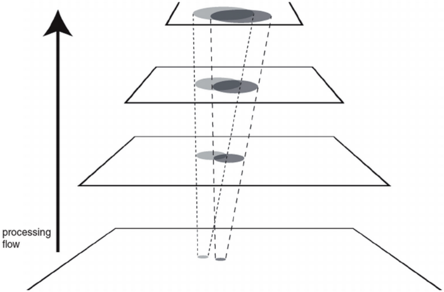
- The issue of blurring becomes much more acute when other visual elements appear in the scene



Source: Image modified from Tsotsos, 2011

Nearby Elements Mutually Interfere

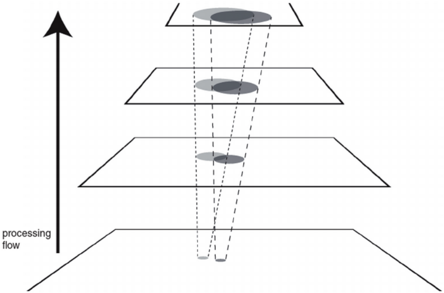
- The issue of blurring becomes much more acute when other visual elements appear in the scene
- At each stage of processing, the degree of activity entanglement tends to increase



Source: Image modified from Tsotsos, 2011

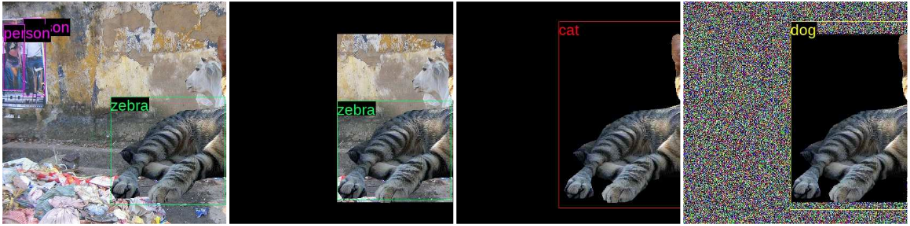
Nearby Elements Mutually Interfere

- The issue of blurring becomes much more acute when other visual elements appear in the scene
- At each stage of processing, the degree of activity entanglement tends to increase
- It can be hard to predict what the behaviour of interacting elements will be



Source: Image modified from Tsotsos, 2011

Cross-talk Example

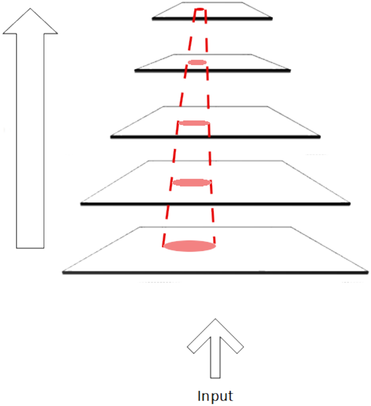


Source: Rosenfeld *et al.*, 2018

An example of the unpredictable interaction of scene elements can be seen with a partially in-frame cat which is initially detected as a zebra, becomes accurately detected once all surrounding pixels are removed, but switches to a dog when noise is added back in to the surround.

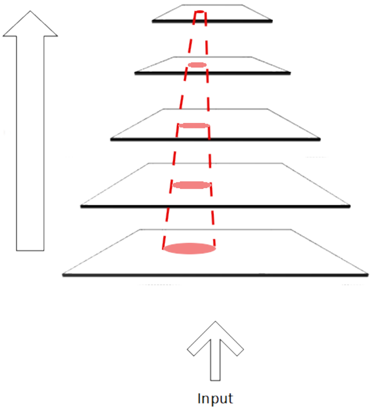
Feature Complexity vs. Spatial Acuity

- Essentially the inverse of the issue with blurring, high-level elements pull information from a wide range of inputs



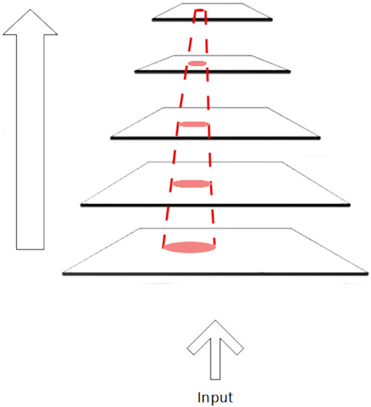
Feature Complexity vs. Spatial Acuity

- Essentially the inverse of the issue with blurring, high-level elements pull information from a wide range of inputs
- As feature complexity and abstraction increases, there is a decrease in spatial acuity



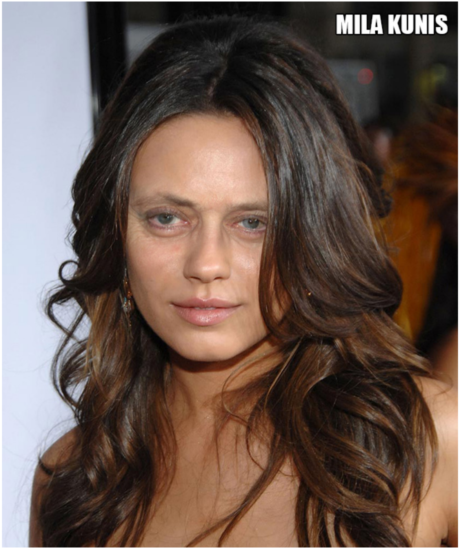
Feature Complexity vs. Spatial Acuity

- Essentially the inverse of the issue with blurring, high-level elements pull information from a wide range of inputs
- As feature complexity and abstraction increases, there is a decrease in spatial acuity
- What happens if we need to pair a high-level representation with high acuity information?



Steve “Buscemeyes”

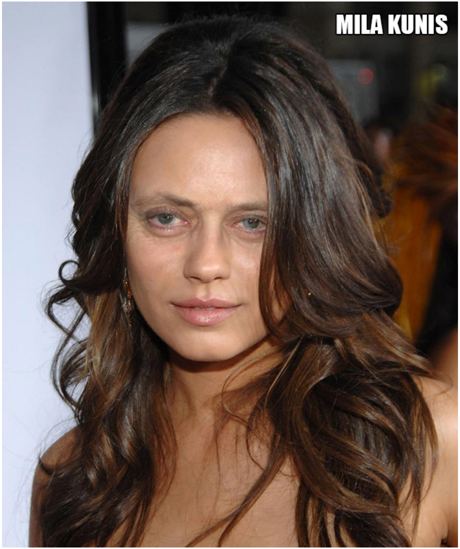
- All the parts of a face are there, and the global average of features still matches Mila Kunis



Source: Original image source unknown

Steve "Buscemeyes"

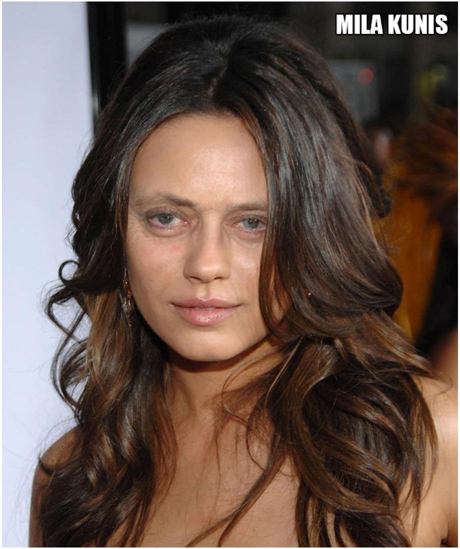
- All the parts of a face are there, and the global average of features still matches Mila Kunis
- We need to be able to see the individual elements in high detail to know that something has gone terribly wrong



Source: Original image source unknown

Steve "Buscemeyes"

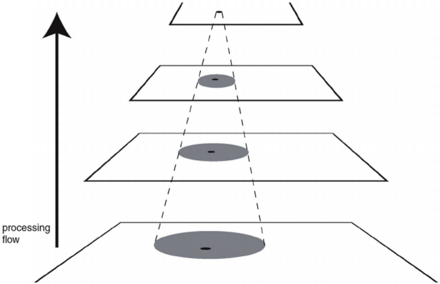
- All the parts of a face are there, and the global average of features still matches Mila Kunis
- We need to be able to see the individual elements in high detail to know that something has gone terribly wrong
- How does the Google network handle this?



Source: Original image source unknown

It is Difficult to Isolate a Specific Element

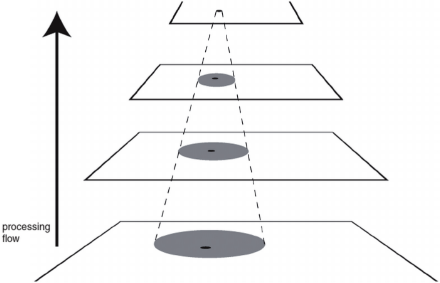
- The challenge of sampling is a problem if we need to isolate a given visual element from its surroundings



Source: Image modified from Tsotsos, 2011

It is Difficult to Isolate a Specific Element

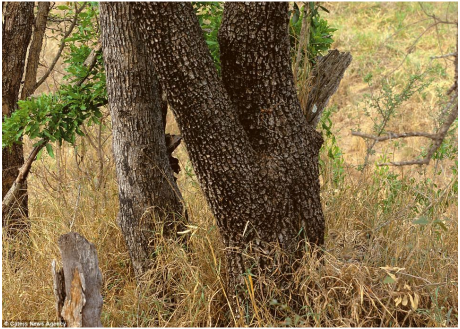
- The challenge of sampling is a problem if we need to isolate a given visual element from its surroundings
- Given the size of a high-level neuron's receptive field, smaller visual elements are always entangled with their surrounding context



Source: Image modified from Tsotsos, 2011

Isolating Scene Elements

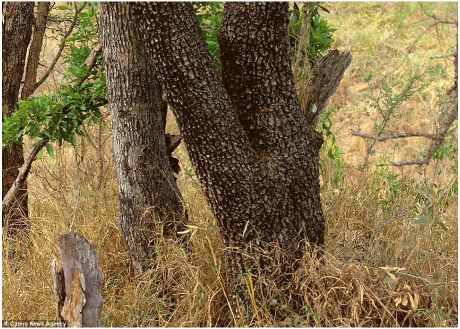
- Isolating a specific scene element is particularly important when the surroundings are particularly disruptive to our percept (*e.g.* camouflage)



Source: Original source unknown

Isolating Scene Elements

- Isolating a specific scene element is particularly important when the surroundings are particularly disruptive to our percept (*e.g.* camouflage)
- Can you spot the leopard on the right?



Source: Original source unknown

So... What?

As we mentioned at the end of last class, CNNs are still an amazing tool, and are frequently at the forefront of computer vision approaches today. So what are we to make of these issues we've just gone over? Do they matter?

So... What?

As we mentioned at the end of last class, CNNs are still an amazing tool, and are frequently at the forefront of computer vision approaches today. So what are we to make of these issues we've just gone over? Do they matter?

It really depends on what you are trying to accomplish. For certain tasks, like recognition or even detection, feedforward computations are sufficient to achieve very good results.

So... What?

As we mentioned at the end of last class, CNNs are still an amazing tool, and are frequently at the forefront of computer vision approaches today. So what are we to make of these issues we've just gone over? Do they matter?

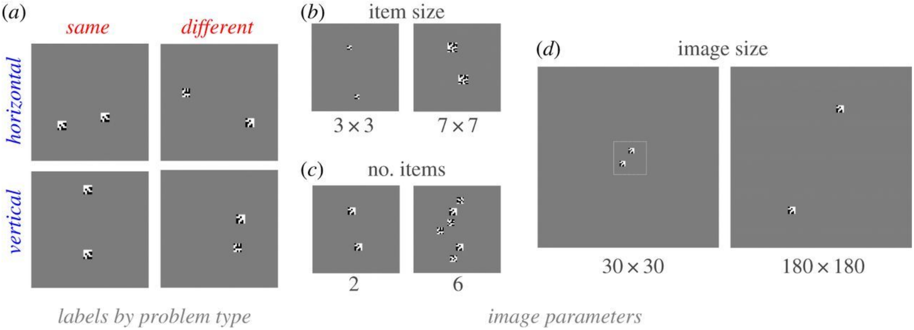
It really depends on what you are trying to accomplish. For certain tasks, like recognition or even detection, feedforward computations are sufficient to achieve very good results.

For other tasks, you will likely struggle to achieve your goals.

Example Limitation: Same-Different Task

An example of a problem class for which CNNs struggle is the comparison of two elements *within the same image*.

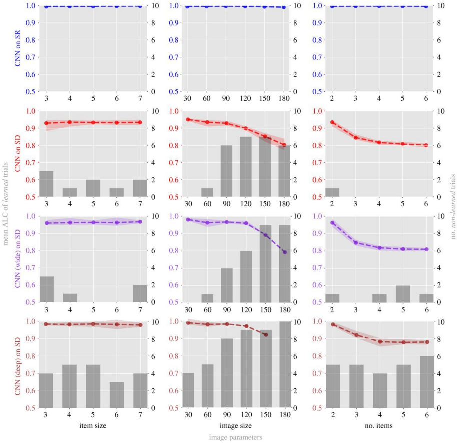
Kim *et al.* (2018), “Not-So-CLEVR: learning samedifferent relations strains feedforward neural networks”



Source: Kim *et al.*, 2018

Combinatorics of Comparison Cause Deficits

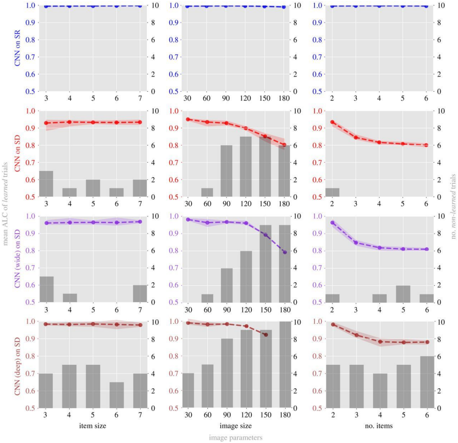
- Two tasks, SR = spatial relations, SD = same-different



Source: Original source unknown

Combinatorics of Comparison Cause Deficits

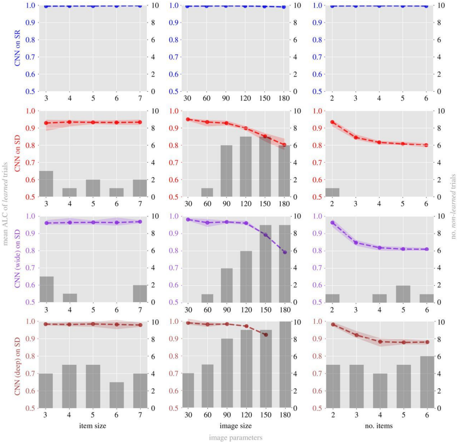
- Two tasks, SR = spatial relations, SD = same-different
- Coloured lines show mean Area-under-the-Learning-Curve (ALC) for trials which achieved more than 55% validation accuracy



Source: Original source unknown

Combinatorics of Comparison Cause Deficits

- Two tasks, SR = spatial relations, SD = same-different
- Coloured lines show mean Area-under-the-Learning-Curve (ALC) for trials which achieved more than 55% validation accuracy
- Grey bars indicate the number of trials which never beat 55% on validation



Source: Original source unknown

Recurrence

One solution to some of these problems is *recurrence*, also sometimes called *feedback*. The idea is to allow information to flow in more than one direction.

Recurrence

One solution to some of these problems is *recurrence*, also sometimes called *feedback*. The idea is to allow information to flow in more than one direction.

- Recurrent Neural Networks (RNNs)

Recurrence

One solution to some of these problems is *recurrence*, also sometimes called *feedback*. The idea is to allow information to flow in more than one direction.

- Recurrent Neural Networks (RNNs)
- Long Short-Term Memory (LSTM) units - small, repeated networks for storing information through time steps

Recurrence

One solution to some of these problems is *recurrence*, also sometimes called *feedback*. The idea is to allow information to flow in more than one direction.

- Recurrent Neural Networks (RNNs)
- Long Short-Term Memory (LSTM) units - small, repeated networks for storing information through time steps

Recurrent networks are often more difficult to train, as errors removed through time will frequently have vanishingly small gradients associated with them.

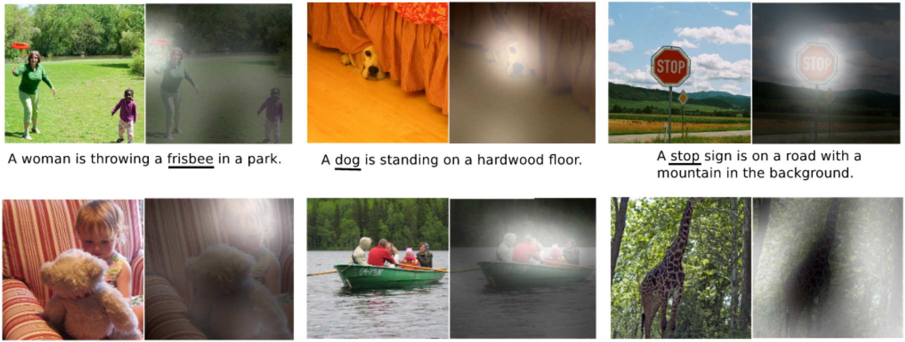
Controlling Recurrence

What information we want to propagate back through a network may change with time and task.

This sort of dynamic control over information flow falls under the umbrella of *attention*.

Spatial Attention

The most common use of an attention mechanism is applying a spatial mask or weighting at the input layer (like we've seen previously with saliency maps).



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.

Spatial attention applied in an image captioning network.

Source: Xu *et al.*, 2016

Attention Can Apply Throughout the Network

Zhang *et al.*, 2016 noted that CNNs already include a readily available signal for feedback: backpropagation.

Attention Can Apply Throughout the Network

Zhang *et al.*, 2016 noted that CNNs already include a readily available signal for feedback: backpropagation.

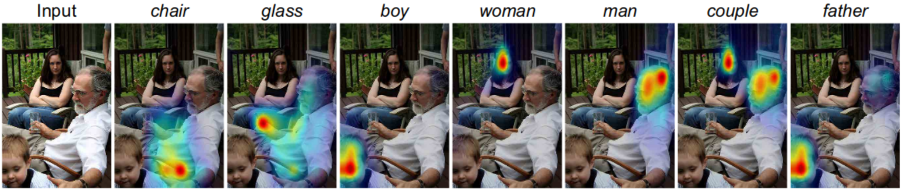
Zhang *et al.* reformulate the standard backpropagation signal as *Excitation Backprop*, whereby they select the set of neurons with the highest activations and positive weights as one moves back through the layers.

Attention Can Apply Throughout the Network

Zhang *et al*., 2016 noted that CNNs already include a readily available signal for feedback: backpropagation.

Zhang *et al*. reformulate the standard backpropagation signal as *Excitation Backprop*, whereby they select the set of neurons with the highest activations and positive weights as one moves back through the layers.

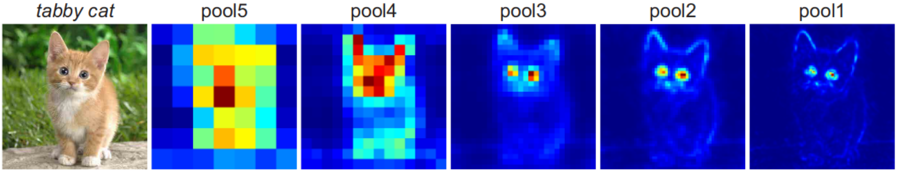
By selecting a different active output class, they can selectively create attention maps for different categories.



Source: Zhang *et al*., 2016

A Demonstration of Sampling Issues

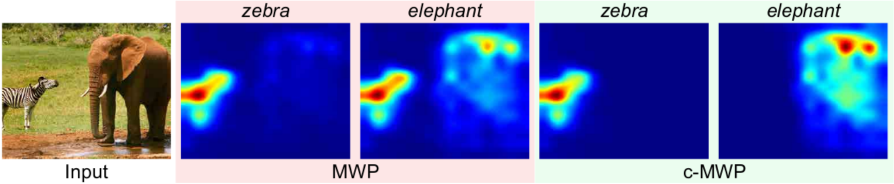
The method of Zhang *et al*. also provides a relatively nice visualization of the sampling issue with CNNs. Here we can see that the spatial acuity at which information is represented in the network is much coarser at higher layers (*pool5*) than at lower layers (*pool1*).



Source: Zhang *et al*., 2016

Not All Features Activate Equally...

A final note about the approach used by Zhang *et al.* is that not all features in a network tend to have the same average activation levels. Some neurons tend to be highly active for many types of input, whereas others are more selective. They therefore found it more effective to use a contrastive scheme whereby they would cancel out (suppress) the selection of neurons which were active in both the class (elephant) and not-class (non-elephant) selections.



Source: Zhang *et al.*, 2016