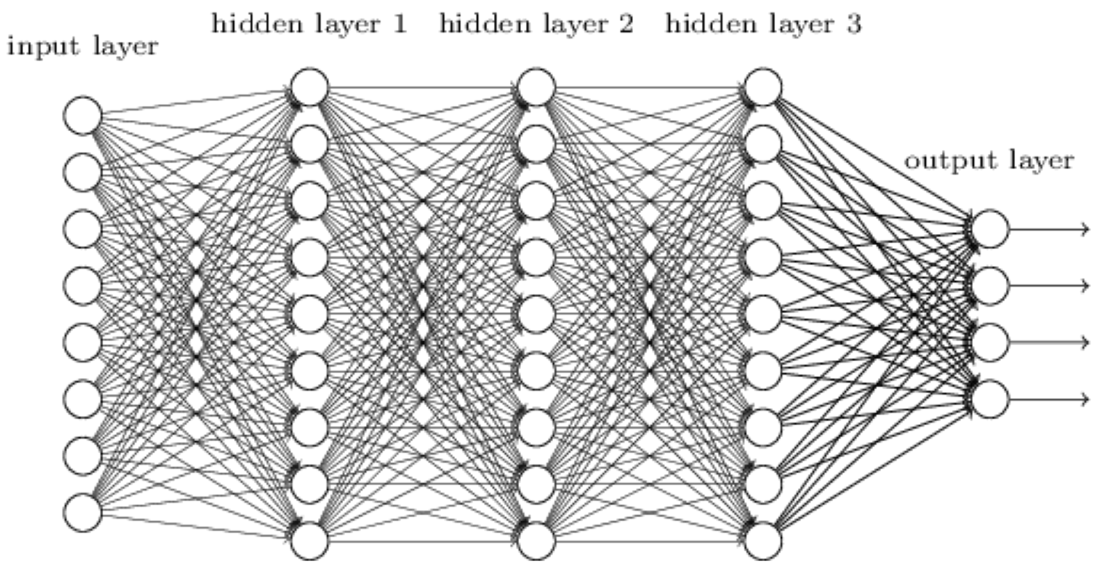
An introduction to deep learning in CASP

Prof. David Jones
UCL Dept. of Computer Science &
The Francis Crick Institute

What is a Deep Neural Network?

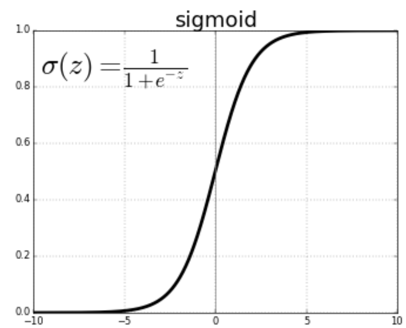


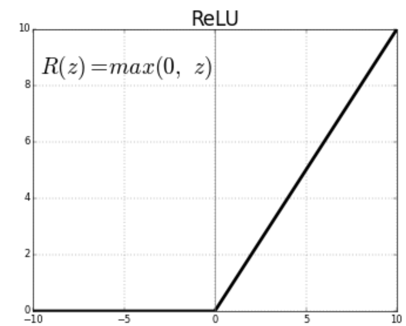
Deep nets are defined as having at least 3 layers (2 hidden layers), but typically many more.

Training of very deep networks is difficult because of the VANISHING GRADIENT PROBLEM.

Key Developments

- 2006, Hinton’s Layer by Layer training of Deep Belief Nets
- 2011, Rectified Linear Units
- 2015, Batch Normalization
- 2016, Residual nets





Turned out that this is all we needed to change to train deep nets!

- 2010, Acceleration by GPUs (CUDA/theano)

THE FRANCIS CRICK INSTITUTE

Open Source Resources

theano

Lasagne

PYTORCH

Caffe

cuDNN

TensorFlow

K

Deep Learning Tutorials:
deeplearning.net/tutorial/
deeplearning4j.org

torch

Deep Learning

THE FRANCIS CRICK INSTITUTE

- Many connected nonlinear processing units (e.g. layers in a DNN)
- These units learn a hierarchy of representations that correspond to different levels of abstraction

"Deep learning has proven to be good at machine vision or text processing problems, but not a lot else." Anon. reviewer

This is obviously not true, but there may be some wisdom in it.

Deep learning is effective at any problem where there is a strong element of <u>hierarchical parsing</u> to be done. This means that deep learning does not work particularly well on <u>unstructured data</u>.

Convolutional Neural Networks

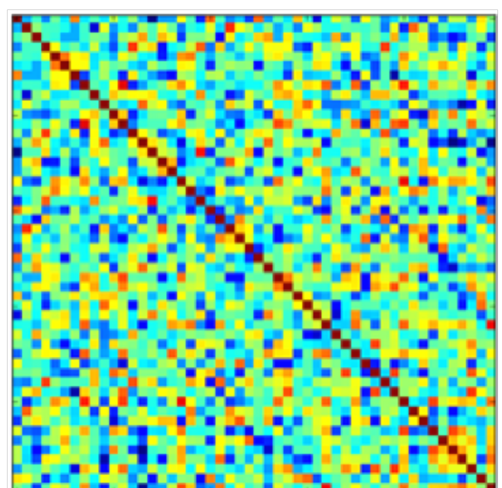
Convolutional networks are popular because they are good for machine vision problems

The basic idea is to take an input image and apply "filters" to produce new images that highlight specific features of the input image, e.g. edges

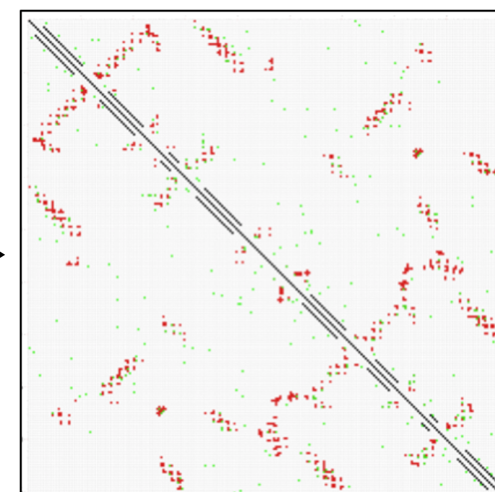
In this second case, residue covariance matrices are the inputs, and want to model the probability of each residue pair being in contact

Input image

Edges highlighted



Residue covariance matrix



Contact probabilities

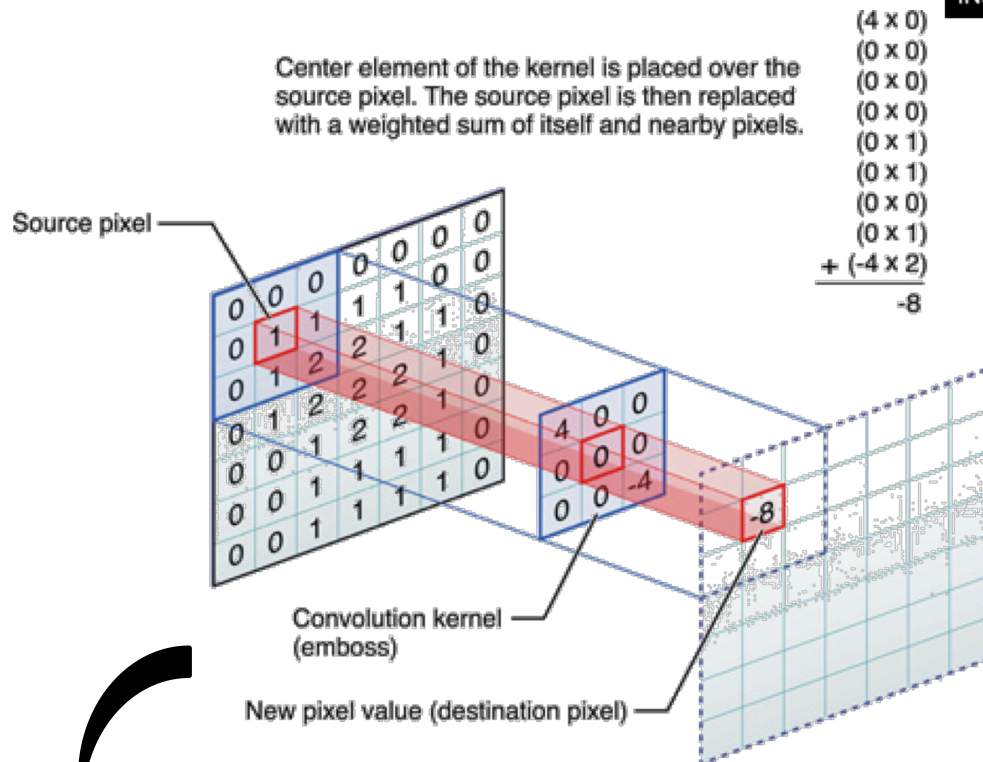
Convolutional layers

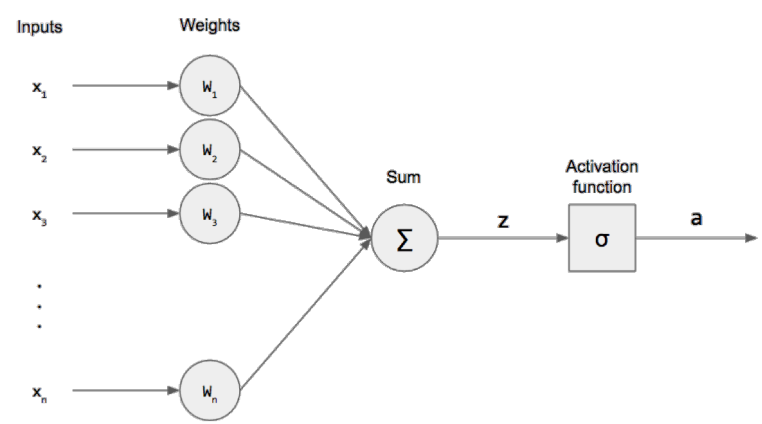
Convolutional nets act on image-like inputs by applying small filters to co-located groups of pixels in the image

The numbers in the filter are trainable weights

After training, the outputs are also image-like, but now convey extra information e.g. presence of an edge

Actually a “filter” is just a small neural network that is applied at every set of pixels (with the same set of weights)





Growing the receptive field by stacking convolutional layers

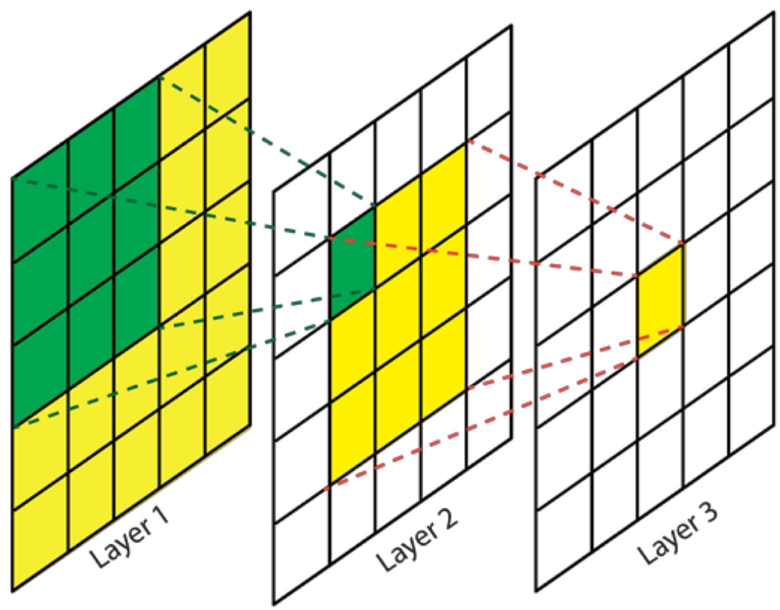


Figure from Lin et al., (2017). Maritime Semantic Labeling of Optical Remote Sensing Images with Multi-Scale Fully Convolutional Network. *Remote Sensing* **9**(5): 480.

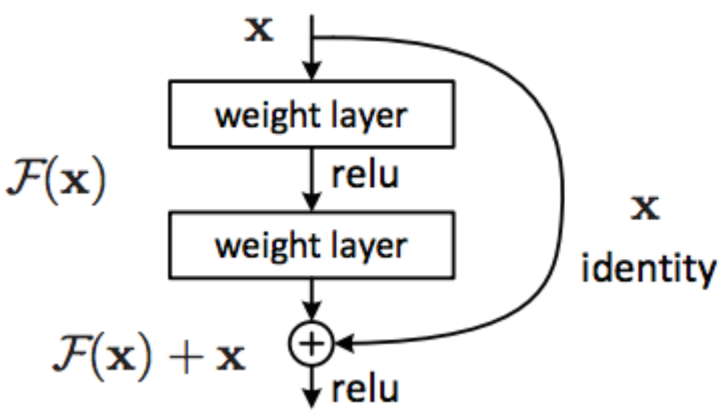
Deep Residual Networks

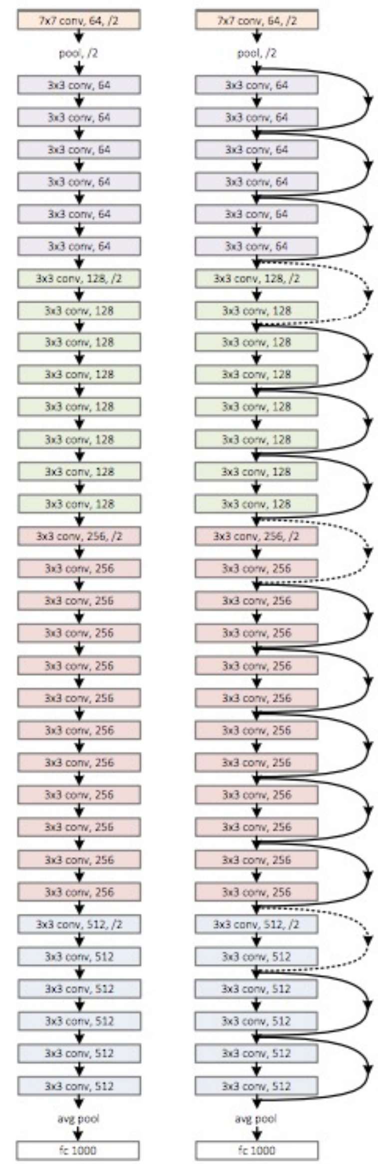
THE FRANCIS CRICK INSTITUTE

Residual neural networks avoid vanishing gradient effects in very deep networks by utilizing *skip connections* or *short-cuts* to jump over some layers.

Further improvements may also come from effective ensembling of different architectures during training.

Deep ResNets are now more or less the standard deep convolutional network models.

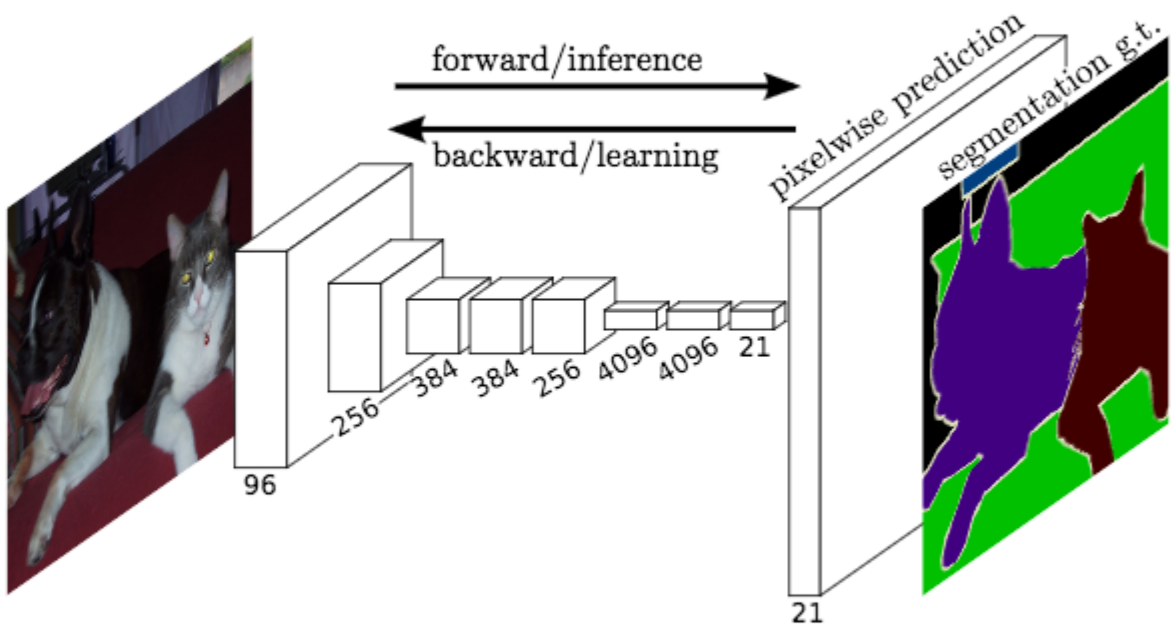




Fully Convolutional Deep Neural Nets

- FCNs are a powerful AI approach to locating objects in images.

- Successful examples are recognizing tumours in X-rays, pedestrians crossing the road, faces in photographs or even cats and dogs…



Fully Convolutional Networks for Semantic Segmentation.
Long et al. CVPR2015

Covariance between MSA columns

THE FRANCIS CRICK INSTITUTE

n rows

```
                                 i                          j
VENLVVYNDGADQRAAEYLADRLACPTINNARKFDYSNVKNVYAVGGNKEQYTSYLTTLIAGSTRYTTMQAVLDYIKNLK
VKNLVVYNNICDQRAAEYLADKLNCPTIWGARPFDYSCVQNVIGVGGKKEQYTSYLKTLVSGNNRYDTMQAVLNYNK---
---LVVYNNICDQRAAEYLADKLNCPTIWGSRPFDYSCVENVIGVGGKKEQYTSYLKTLLTGNNRYDTMQVVLNYSK---
VENLVVYNDGADQRAAEYLADRLACPTINNARKFDYSNVKNVYAVGGNKEQYTSYLTTLIAGSTRYTTMQAVLDYIKNLK
---LVVYNNIADARAAEYLADKLNCPTIWGARSFDYSVVENVIGVGGKKEQYTSYLKTLISGNNRYYTMQAVLNYNK---
VEHLILVSRGADERAAGYLADYLQAPILYLDRLSNLDSAKKIYVVGGNTKPVDRAI--LISGTDRYGTCQKVLDFIRTGK
LKNLILVGHGPDERAAGYLADFLKAPVAYLDQADDLNSAQNIYVIGGSVKPLERA--TLISGATRYDTCQKVIDFIHTGK
MKNLIITNRGADERAAGYLADYLKAPVVLLDQLTDVQGAENIYVVGGSDKPVSSAI--LISGSNRYETCRKVLDICGN--
VDYIIQYSNSTDQAIAEIMADRLNCPTINCLRPYAYGQYKTVIAVGEAKNK-SGYTNVEIKGANRKETLDKAIEYCEKL-
MENIVVYYYPLDQRSAEYVAGELNCTTIYVARTSNYSCVKNIIAVGGRIGKYKEYNITIIAGNGRYDTLKAVVDYIK---
```

We need to use an input feature set in which each element only has information from given pairs of columns in an MSA (no information from any other columns). We consider a simple covariance measure S:

$$S_{ij}^{ab} = \frac{1}{n} \sum_{k=1}^{n} \left(x_i^{ak} - \bar{x}_i^{a}\right)\left(x_j^{bk} - \bar{x}_j^{b}\right),$$

where:

n: number of rows in the MSA

k: row index in MSA

i, j: column indices in MSA

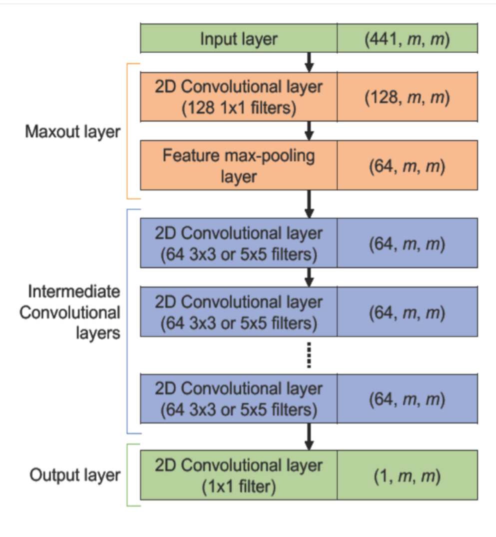
a, b: residue types (Ala, Gly etc., with gaps considered as a 21st type)

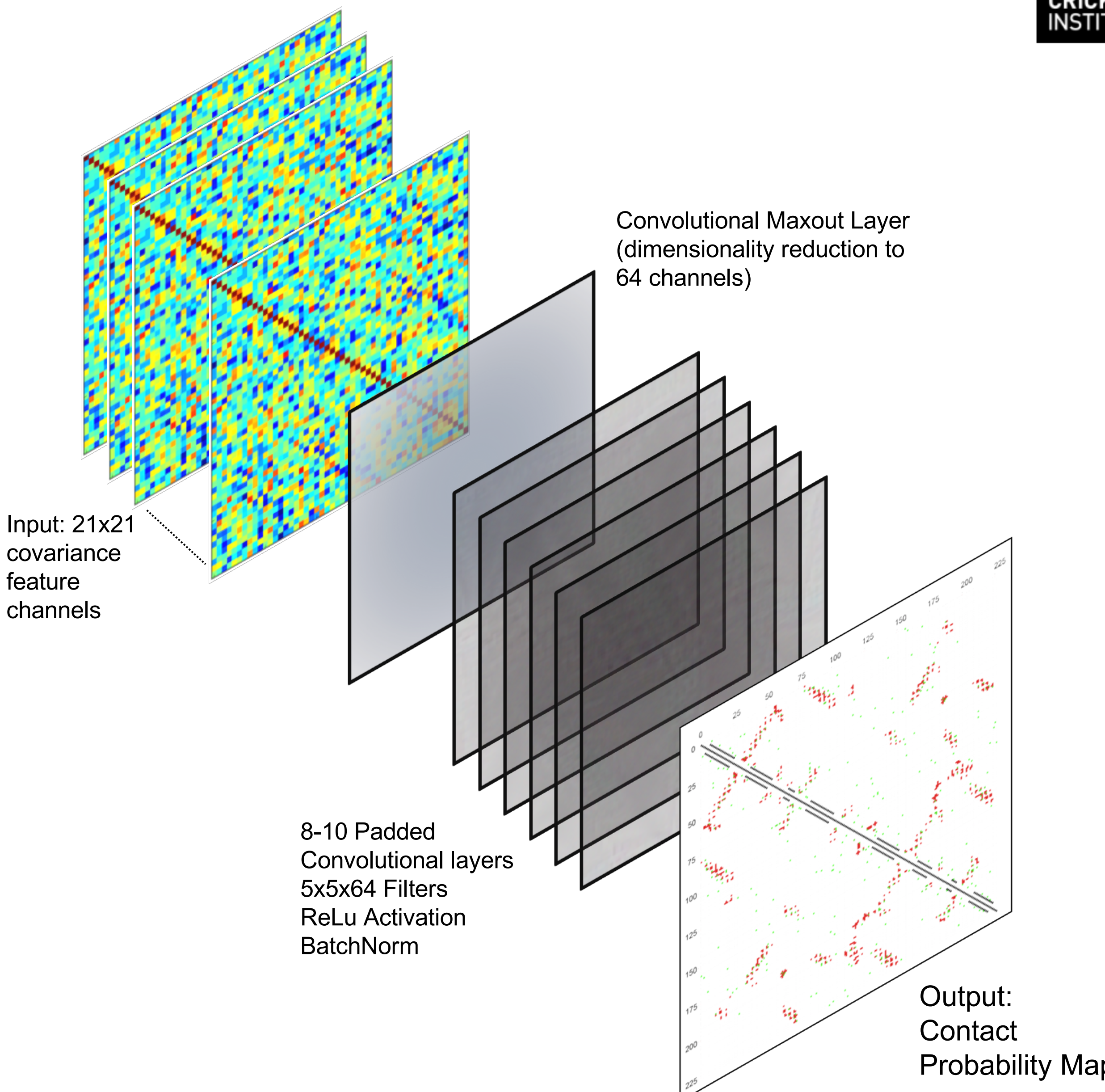
x: residue presence or absence (1 or 0) in specified MSA column and row

$\bar{x}$: relative frequency of residue in specified MSA column

DeepCOV: Analysing Residue Covariation using FCNs

THE FRANCIS CRICK INSTITUTE





Determining sequence locality of chaining effects using varying sized receptive fields

THE FRANCIS CRICK INSTITUTE

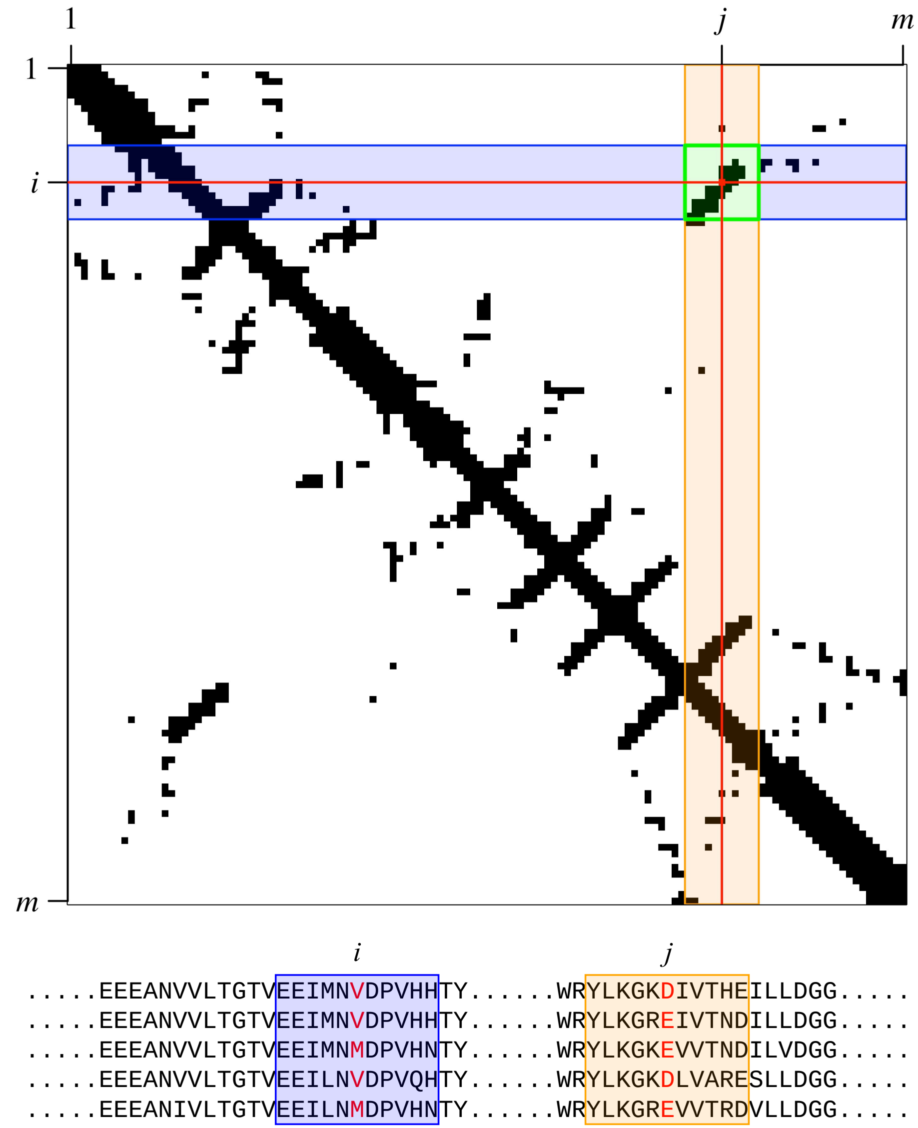
Alignment has *m* columns
Contact map is *m* x *m* matrix

Features calculated in pairwise fashion on alignment columns will also be *m* x *m*

The receptive field is a square set of the covariances that are used to get the contact prediction for residue pair (*i*, *j*)

This corresponds to 2 windows in the MSA around columns *i* and *j* (blue and orange)

Fully convolutional networks (FCNs) allow us to easily control the receptive field size

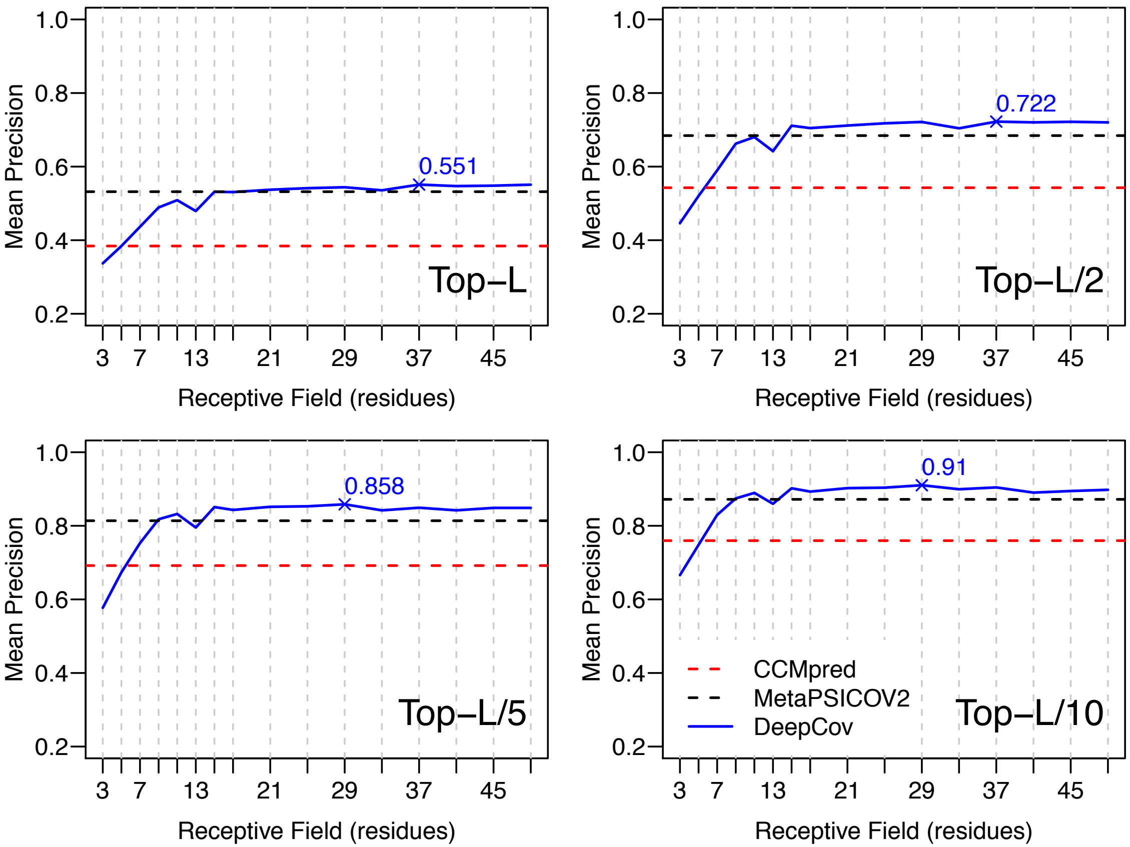


Performance on PSICOV150 test set using only covariance as input

Precision increases with receptive field size

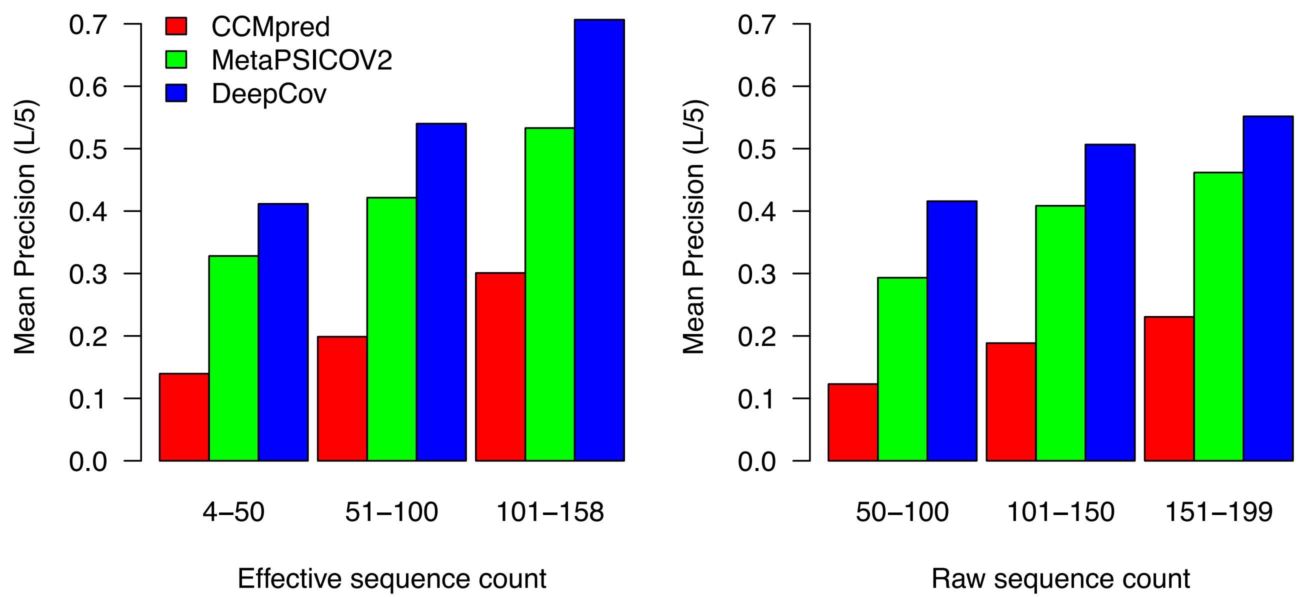
Mean precision plateaus at receptive field of only around 15 residues

DeepCov is more precise than CCMpred and MetaPSICOV2



Performance on shallow alignments

THE FRANCIS CRICK INSTITUTE



232 alignments with fewer than 200 raw sequences (effective count < 159)

Retrained DeepCov model using different training set with no overlap to this test set (by ECOD classification)

Fewer sequences needed to get reasonably accurate predictions! **Why?**

Do we still need deep alignments?

- It appears that DL methods are able to derive contacts from relatively shallow alignments
- Classical global covariation methods e.g. PSICOV, plmDCA worked well with deep alignments but fail on shallow alignments

- This is not necessarily a new aspect of DL methods but more a failure mode of global methods e.g. SICE model in PSICOV:

$$\sum_{ij=1}^{d} S_{ij}\Theta_{ij} - \log\mathrm{d}et\Theta + \rho\sum_{ij=1}^{d} |\Theta_{ij}|$$

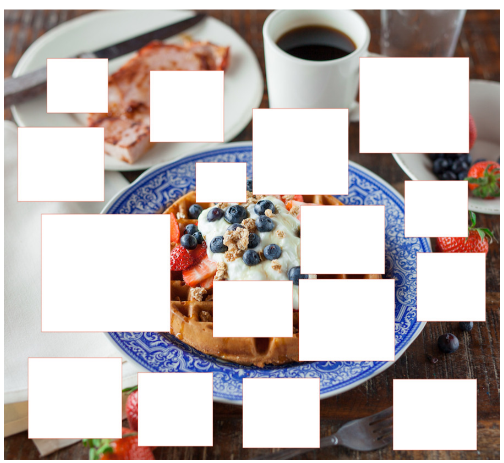
- Note that in global models we are forced to fill-in missing data e.g. by using shrinking and pseudocounts

- DL models just learn to ignore the missing data!

Deep learning models are relatively insensitive to missing data…

THE FRANCIS CRICK INSTITUTE

no person	0.995
Breakfast	0.988
Food	0.984
Milk	0.960
Delicious	0.957
Fruit	0.950
Dawn	0.941
Refreshment	0.938
dairy product	0.932
Homemade	0.931
Cream	0.924
Table	0.922

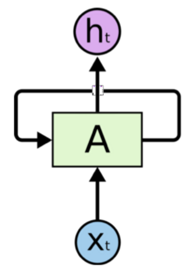


no person	0.977
Table	0.957
Business	0.937
Coffee	0.929
Indoors	0.924
Paper	0.909
Horizontal	0.908
Illustration	0.896
Furniture	0.867
Design	0.851
Family	0.848
Food	0.843

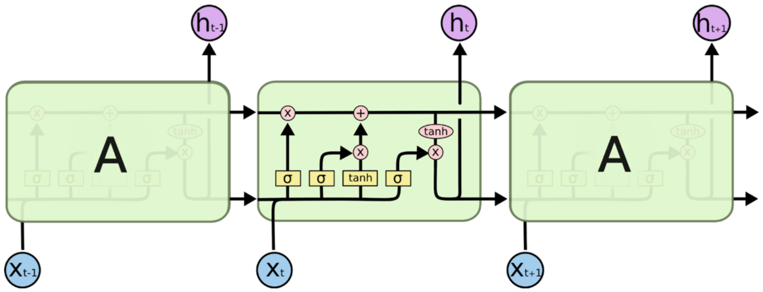
https://clarifai.com/demo

Sequence-to-sequence models

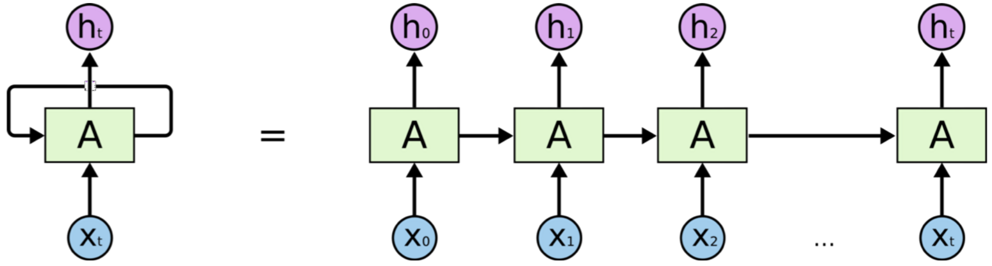
Simple Recurrent Network (RNN)



LSTM RNN



RNN = CNN?



Sequence-to-sequence models

Useful for…

- **Secondary structure prediction**
  - Sequence profile → DSSP codes

- **Torsion angle prediction**
  - Sequence profile → main chain torsion angles

- **Estimation of model accuracy**
  - Sequence of 3-D coordinates → GDT score

- **End-to-end differentiable folding models**
  - Amino acid sequence → torsion angles
    - Torsion angles → 3-D coordinates

Summary

THE FRANCIS CRICK INSTITUTE

- **Deep learning (deep residual networks) has improved covariation-based contact prediction**
  - Problem fits the deep learning model of hierarchical feature extraction well, and deep learning is robust to missing data (shallow alignments)
  - By varying the network depth we can gain some insight into the most significant contributions to indirect correlation chaining
  - Can easily be adapted to predict distance distribution outputs, or to produce an EMA method
- **Sequence-to-sequence models have proven effective in predicting 1-D protein features e.g. secondary structure & torsion angles**
  - End-to-end differentiable models may be interesting to explore

UCL

THE FRANCIS CRICK INSTITUTE

Acknowledgements

BBSRC
bioscience for the future

erc