# Professional English – II
# Reading

Chixiao Chen

cxchen@fudan.edu.cn

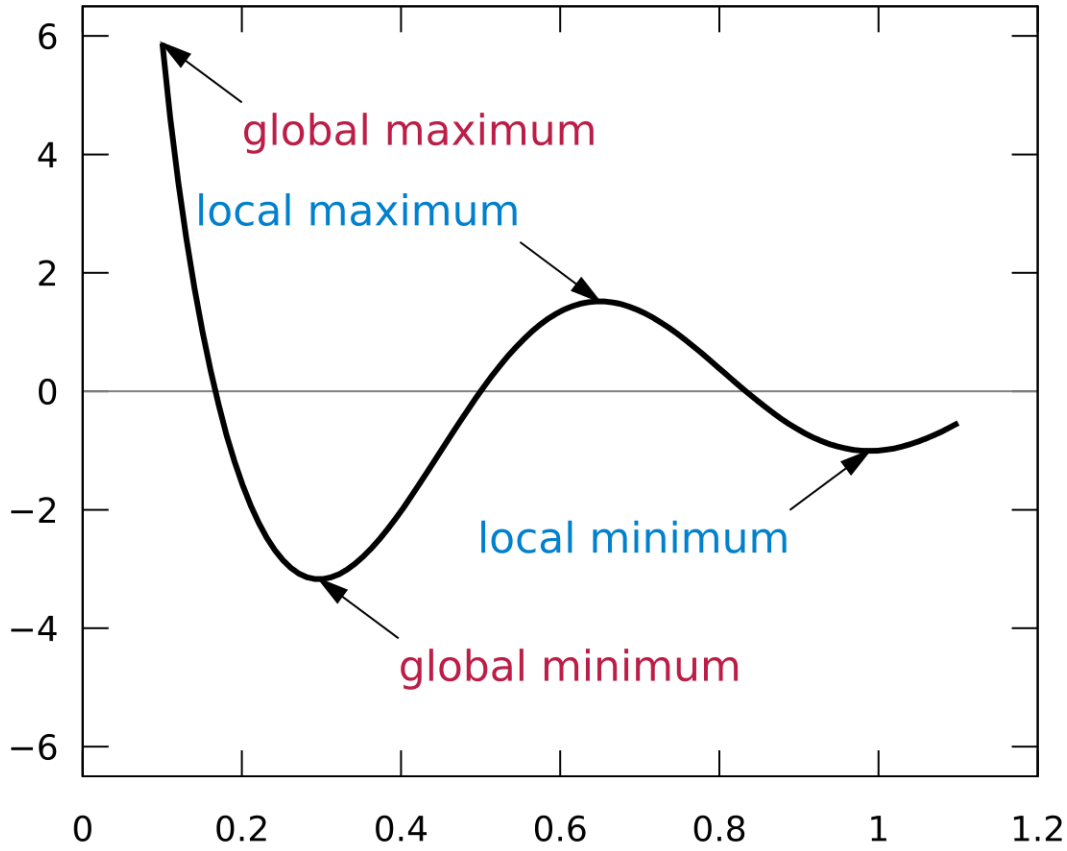
# Announcement

- We will have our first quiz next week !
- Our presentation vacancy is reserved until Nov. 22.
- I probably cannot attend the class on Nov. 29. If so, I will upload a video on the course website.

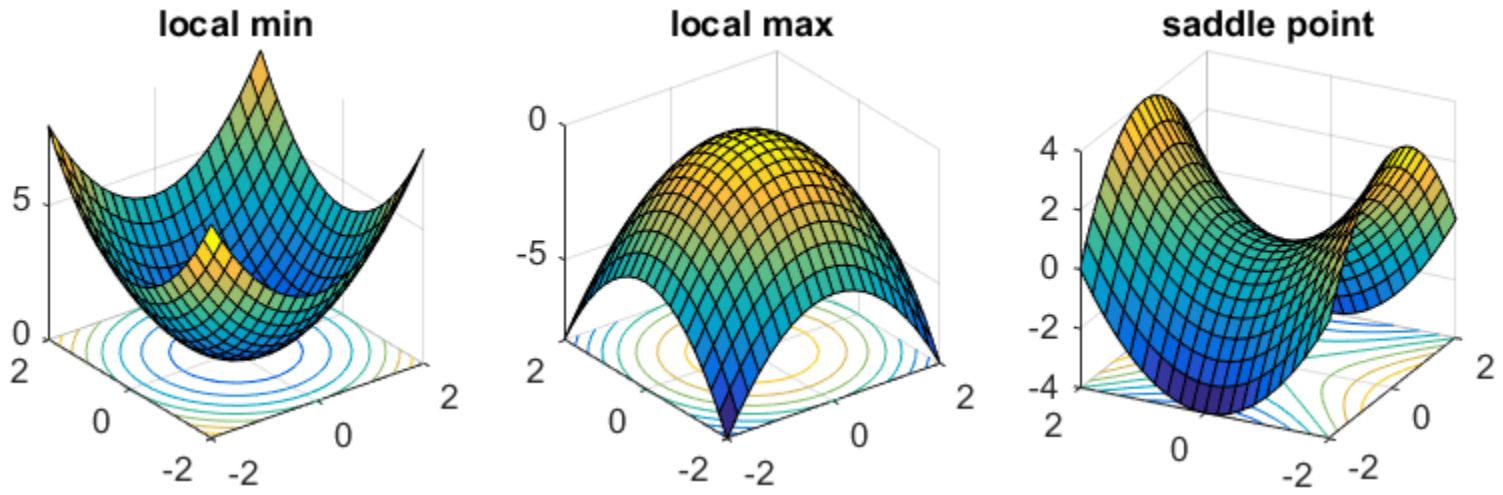
# Review on the last class

- Breakthrough
- non-linear
- Sequential
- General purpose
- Particle accelerator
- Knob
- Derivative
- Performance
- Procedure
- Molecule
- Promising
- Simultaneously
- Gradient
- Edge
- Regardless of
- vector

It was widely thought that learning useful, multistage, feature extractors with little prior knowledge was infeasible. In particular, it was commonly thought that simple gradient descent would get trapped in poor local minima — weight configurations for which no small change would reduce the average error.



Regardless of the initial conditions, the system nearly always reaches solutions of very similar quality. Recent theoretical and empirical results strongly suggest that local minima are not a serious issue in general. Instead, the landscape is packed with a combinatorially large number of saddle points where the gradient is zero, and the surface curves up in most dimensions and curves down in the remainders.



By ‘pre-training’ several layers of progressively more complex feature detectors using this reconstruction objective, the weights of a deep network could be initialized to sensible values.

Once deep learning had been rehabilitated, it turned out that the pre-training stage was only needed for small data sets.

In 2009, the approach was used to map short temporal windows of coefficients extracted from a sound wave to a set of probabilities for the various fragments of speech that might be represented by the frame in the center of the window. It achieved record-breaking results on a standard speech recognition benchmark that used a small vocabulary and was quickly developed to give record-breaking results on a large vocabulary task.

There was, however, one particular type of deep, feedforward network that was much easier to train and generalized much better than networks with full connectivity between adjacent layers. This was the convolutional neural network (ConvNet) .

Many data modalities are in the form of multiple arrays: 1D for signals and sequences, including language; 2D for images or audio spectrograms; and 3D for video or volumetric images.

The reason for this architecture is twofold. First, in array data such as images, local groups of values are often highly correlated, forming distinctive local motifs that are easily detected. Second, the local statistics of images and other signals are invariant to location. **In other words**, if a motif can appear in one part of the image, it could appear anywhere, hence the idea of units at different locations sharing the same weights and detecting the same pattern in different parts of the array.

Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher-level features are obtained by composing lower-level ones.

The convolutional and pooling layers in ConvNets are directly inspired by the classic notions of simple cells and complex cells in visual neuroscience, and the overall architecture is reminiscent of the LGN–V1–V2–V4–IT hierarchy in the visual cortex ventral pathway.

The document reading system used a ConvNet trained jointly with a probabilistic model that implemented language constraints. By the late 1990s this system was reading over 10% of all the cheques in the United States.

A number of ConvNet-based optical character recognition and handwriting recognition systems were later deployed by Microsoft.

These were all tasks in which labelled data was relatively abundant, such as traffic sign recognition, the segmentation of biological images particularly for connectomics, and the detection of faces, text, pedestrians and human bodies in natural images.

Despite these successes, ConvNets were largely forsaken by the mainstream computer-vision and machine-learning communities until the ImageNet competition in 2012.

ConvNets are easily amenable to efficient hardware implementations in chips or field-programmable gate arrays.

Learning word vectors turned out to also work very well when the word sequences come from a large corpus of real text and the individual micro-rules are unreliable.

The issue of representation lies at the heart of the debate between the logic-inspired and the neural-network-inspired paradigms for cognition.

The instance of a symbol has no internal structure that is relevant to its use; and to reason with symbols, they must be bound to the variables in judiciously chosen rules of inference.

Before the introduction of neural language models, the standard approach to statistical modelling of language did not exploit distributed representations: it was based on counting frequencies of occurrences of short symbol sequences of length up to N (called N-grams).

The number of possible N-grams is on the order of $V^N$, where V is the vocabulary size, so taking into account a context of more than a handful of words would require very large training corpora.

RNNs are very powerful dynamic systems, but training them has proved to be problematic because the backpropagated gradients either grow or shrink at each time step, so over many time steps they typically explode or vanish.

This rather naive way of performing machine translation has quickly become competitive with the state-of-the-art, and this raises serious doubts about whether understanding a sentence requires anything like the internal symbolic expressions that are manipulated by using inference rules.

It is more compatible with the view that everyday reasoning involves many simultaneous analogies that each contribute plausibility to a conclusion.

To correct for that, one idea is to augment the network with an explicit memory.

implicit

Unsupervised learning91–98 had a catalytic effect in reviving interest in deep learning, but has since been overshadowed by the successes of purely supervised learning.

Natural language understanding is another area in which deep learning is poised to make a large impact over the next few years.