Computer vision

Giving computers the ability to understand visual information
Computer vision

Giving computers the ability to understand visual information

Examples:

- A robot that can move around obstacles by analysing the input of its camera(s)
Computer vision

Giving computers the ability to understand visual information

Examples:

- A robot that can move around obstacles by analysing the input of its camera(s)
- A computer system finding images of cats among millions of images (e.g., on the Internet).
How hard can it be?

When a user takes a photo, the app should check whether they're in a national park...

Sure, easy GIS lookup.
Gimme a few hours.

...and check whether the photo is of a bird.

I'll need a research team and five years.

In CS, it can be hard to explain the difference between the easy and the virtually impossible.

https://xkcd.com/1425/
From picture to pixels

- The camera image needs to be digitised for computer processing
- Turning it into millions of discrete picture elements, or pixels

There's a cat among some flowers in the grass

How do we get from pixels to understanding... or even some kind of useful/actionable interpretation.
From picture to pixels

- The camera image needs to be digitised for computer processing
- Turning it into millions of discrete picture elements, or pixels

“There’s a cat among some flowers in the grass”

- How do we get from pixels to understanding?
From picture to pixels

- The camera image needs to be digitised for computer processing
- Turning it into millions of discrete picture elements, or pixels

“There's a cat among some flowers in the grass”

- How do we get from pixels to understanding?
- ...or even some kind of useful/actionable interpretation.
Deep learning

Before

- Hand-crafted features, e.g., colour distributions, edge histograms
- Complicated feature selection mechanisms
- “Classical” machine learning, e.g., kernel methods (SVM)
Deep learning

Before

- Hand-crafted features, e.g., colour distributions, edge histograms
- Complicated feature selection mechanisms
- “Classical” machine learning, e.g., kernel methods (SVM)

About 5 years ago: deep learning

- End-to-end learning, i.e., the network itself learns the features
- Each layer typically learns a higher level of representation
- However: entirely data-driven, features can be hard to interpret
Deep learning

Before

- Hand-crafted features, e.g., colour distributions, edge histograms
- Complicated feature selection mechanisms
- “Classical” machine learning, e.g., kernel methods (SVM)

About 5 years ago: deep learning

- End-to-end learning, i.e., the network itself learns the features
- Each layer typically learns a higher level of representation
- However: entirely data-driven, features can be hard to interpret

Computer vision was one of the first breakthroughs of deep learning.
Deep learning for vision

While we don’t hand-craft features anymore . . .

In practice we still apply some “expert knowledge” to make learning feasible . . .
Deep learning for vision

While we don’t hand-craft features anymore . . .

In practice we still apply some “expert knowledge” to make learning feasible . . .

- Neighbouring pixels are probably related (convolutions)
Deep learning for vision

While we don’t hand-craft features anymore . . .

In practice we still apply some “expert knowledge” to make learning feasible . . .

- Neighbouring pixels are probably related (convolutions)
- There are common image features which can appear anywhere such as edges, corners, etc (weight sharing)
Deep learning for vision

While we don’t hand-craft features anymore . . .

In practice we still apply some “expert knowledge” to make learning feasible . . .

- Neighbouring pixels are probably related (convolutions)
- There are common image features which can appear anywhere such as edges, corners, etc (weight sharing)
- Often the exact location of a feature isn’t important (max pooling)
Deep learning for vision

While we don’t hand-craft features anymore . . .

In practice we still apply some “expert knowledge” to make learning feasible . . .

▶ Neighbouring pixels are probably related (convolutions)
▶ There are common image features which can appear anywhere such as edges, corners, etc (weight sharing)
▶ Often the exact location of a feature isn’t important (max pooling)

⇒ Convolutional neural networks (CNN, ConvNet).
Convolution in 2D

We arrange the input and output neurons in 2D

\[ S(i, j) = \sum_{m} \sum_{n} I(i+m, j+n) K(m, n) \]

The weights \( K(m, n) \) is what is learned.
Convolution in 2D

- We arrange the input and output neurons in 2D
- The output is the result of a weighted sum of a small local area in the previous layer – convolution

\[
S(i, j) = \sum_{m} \sum_{n} l(i + m, j + n) K(m, n)
\]
Convolution in 2D

- We arrange the input and output neurons in 2D
- The output is the result of a weighted sum of a small local area in the previous layer – convolution

\[ S(i, j) = \sum_{m} \sum_{n} l(i + m, j + n)K(m, n) \]
Convolution in 2D

▷ We arrange the input and output neurons in 2D
▷ The output is the result of a weighted sum of a small local area in the previous layer – convolution

\[
S(i, j) = \sum_{m} \sum_{n} I(i + m, j + n)K(m, n)
\]
Convolution in 2D

- We arrange the input and output neurons in 2D
- The output is the result of a weighted sum of a small local area in the previous layer – convolution

\[ S(i, j) = \sum_{m} \sum_{n} I(i + m, j + n)K(m, n) \]

- The weights \( K(m, n) \) is what is \textit{learned}.
Learning in layers

- The convolutional layer learns several sets of weights, each a kind of feature detector
Learning in layers

- The convolutional layer learns several sets of weights, each a kind of feature detector.
- These are built up in layers.
Learning in layers

- The convolutional layer learns several sets of weights, each a kind of feature detector
- These are built up in layers
- Until we get our end result, e.g., an object detector
Visualising convolutional layers

Figure 3: 96 convolutional kernels of size $11 \times 11 \times 3$ learned by the first convolutional layer on the $224 \times 224 \times 3$ input images.

Krizhevsky et al 2012
Map activations back to the image space
Real convolutional neural nets

- What we call CNNs, actually also contain other types of operations/layers: fully connected layers, non-linearities

LeNet5 (LeCun et al 1998)
Real convolutional neural nets

- What we call CNNs, actually also contain other types of operations/layers: fully connected layers, non-linearities
- Modern CNNs have a huge bag of tricks: pooling, various training shortcuts, 1x1 convolutions, inception modules, residual connections, etc.

LeNet5 (LeCun et al 1998)
Examples of real CNNs

AlexNet (Krizhevsky et al 2012)
Examples of real CNNs

GoogLeNet (Szegedy et al 2014)
Examples of real CNNs

Inception v3 (Szegedy et al 2015)
Object recognition challenge

ImageNet benchmark

- ImageNet Large Scale Visual Recognition Challenge (ILSVRC)
- More than 1 million images
- Task: classify into 1000 object categories
Object recognition challenge

- First time won by a CNN in 2012 (Krizhevsky et al)
- Wide margin: top-5 error rate from 26% to 16%
Object recognition challenge

- First time won by a CNN in 2012 (Krizhevsky et al)
- Wide margin: top-5 error rate from 26% to 16%
- CNNs have ruled ever since.
Accuracy vs model complexity

- Accuracy vs number of inference operations
- Circle size represents number of parameters

Accuracy vs model complexity

- Accuracy vs number of inference operations
- Circle size represents number of parameters
- Newer nets are both better, faster and have fewer parameters.

Computer vision applications
Object detection and localisation

R-CNN: Regions with CNN features

1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions


Object detection and localisation

https://github.com/facebookresearch/Detectron
Describing an image

Describing an image

Image and Video Captioning with Augmented Neural Architectures, Rakshith Shetty, Hamed R. Tavakoli and Jorma Laaksonen.
Describing an image

DenseCap: Fully Convolutional Localization Networks for Dense Captioning, Justin Johnson, Andrej Karpathy, Li Fei-Fei, CVPR 2016.
Visual question answering

Generative Adversarial Networks (GANs)

“The coolest idea in machine learning in the last twenty years”
– Yann LeCun
Generative Adversarial Networks (GANs)

“The coolest idea in machine learning in the last twenty years”
– Yann LeCun

- We have two networks: generator and discriminator
Generative Adversarial Networks (GANs)

“The coolest idea in machine learning in the last twenty years”
– Yann LeCun

- We have two networks: generator and discriminator
- The generator produces samples, while the discriminator tries to distinguish between real data items and the generated samples
Generative Adversarial Networks (GANs)

“The coolest idea in machine learning in the last twenty years”
– Yann LeCun

- We have two networks: \textit{generator} and \textit{discriminator}
- The generator produces samples, while the discriminator tries to distinguish between real data items and the generated samples
- The discriminator tries to learn to classify correctly, while the generator in turn tries to learn to fool the discriminator.
GAN examples
Generated bedrooms

https://arxiv.org/abs/1511.06434v2
GAN examples

CycleGAN

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks
https://junyanz.github.io/CycleGAN/
GAN examples


this small bird has a pink breast and crown, and black primaries and secondaries. this magnificent fellow is almost all black with a red crest, and white cheek patch.

the flower has petals that are bright pinkish purple with white stigma this white and yellow flower have thin white petals and a round yellow stamen

Figure 1. Examples of generated images from text descriptions. Left: captions are from zero-shot (held out) categories, unseen text. Right: captions are from the training set.
AI vs humans?
AI vs humans?

Recall our ImageNet benchmark . . . where do humans stand?

![ImageNet classification top-5 error rate chart](http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/)
AI vs humans?

Recall our ImageNet benchmark ... where do humans stand?

http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/
AI better than humans?

Don't confuse classification accuracy with understanding!

Neural nets learn to optimize for a particular problem really well.

But in the end it's just pixel statistics.

Humans can generalize and understand the context.
AI better than humans?

▶ Don’t confuse classification accuracy with understanding!
▶ Neural nets learn to optimize for a particular problem really well
AI better than humans?

- Don’t confuse classification accuracy with understanding!
- Neural nets learn to optimize for a particular problem really well
- But in the end it’s just pixel statistics
- Humans can generalize and understand the context.
AI better than humans?

https://karpathy.github.io/2012/10/22/state-of-computer-vision/
AI better than humans?

Microsoft CaptionBot: “I think it’s a group of people standing next to a man in a suit and tie.”

https://karpathy.github.io/2012/10/22/state-of-computer-vision/
Google’s AI thinks this turtle looks like a gun, which is a problem

New research shows how machine vision systems of all kinds can be tricked into misidentifying 3D objects

by James Vincent | @jVincent | Nov 2, 2017, 8:19am EDT
Adversarial examples

- Deep nets fooled by deliberately crafted inputs

https://blog.openai.com/adversarial-example-research/
Adversarial examples

- Deep nets fooled by deliberately crafted inputs
- Revealing: what deep nets learn is quite different from what humans learn

https://blog.openai.com/adversarial-example-research/
Project at Aalto: reasoning with scene graphs

Despite great strides, cognitive tasks which require reasoning about the world, still remain elusive.

Figure 1: Actual results using a popular image search engine (top row) and ideal results (bottom row) for the query *a boy wearing a t-shirt with a plane on it*.
Despite great strides, cognitive tasks which require reasoning about the world, still remain elusive.

Recent idea: train on densely annotated image data: objects, attributes and their relationships.

Can we learn to really understand a visual scene?
Project at Aalto: reasoning with scene graphs

Such a database has recently been created: Visual Genome

- 108,077 Images
- 5.4 Million Region Descriptions
- 1.7 Million Visual Question Answers
- 3.8 Million Object Instances
- 2.8 Million Attributes
- 2.3 Million Relationships
- Everything Mapped to Wordnet Synsets

Project at Aalto: reasoning with scene graphs

DeepGraph project

- The CBIR group at Dept. of Computer Science, Aalto University, https://research.cs.aalto.fi/cbir/
- Funded by Academy of Finland, 1.1.2018 - 31.12.2019 (1. period), ICT 2023: Computation, Machine Learning and Artificial Intelligence

In conjunction with FCAI, Finnish Center for Artificial Intelligence, http://www.fcai.fi/

Project at Aalto: reasoning with scene graphs

DeepGraph project

- The CBIR group at Dept. of Computer Science, Aalto University, https://research.cs.aalto.fi/cbir/
- Funded by Academy of Finland, 1.1.2018 - 31.12.2019 (1. period), ICT 2023: Computation, Machine Learning and Artificial Intelligence
- In conjunction with FCAI, Finnish Center for Artificial Intelligence, http://www.fcai.fi/
Conclusion

- Deep learning has been a big leap for computer vision
- We can solve some specific problems really well

Thank you for listening!
Conclusion

- Deep learning has been a big leap for computer vision
- We can solve some specific problems really well
- Still far away from true understanding of visual information
Conclusion

- Deep learning has been a big leap for computer vision
- We can solve some specific problems really well
- Still far away from true understanding of visual information
- Scene graphs are one promising future direction
Conclusion

- Deep learning has been a big leap for computer vision
- We can solve some specific problems really well
- Still far away from true understanding of visual information
- Scene graphs are one promising future direction

Thank you for listening!