Iris Groen – Day 1 Lecture 1

07.07

Part 1

→ Hubel & Wiesel (cats :))

→ orientation tuning

→ retinotopic organisation

* Kobatake & Tanaka (1994)
* Desimone et al (1984)
* Grill-Spector & Weiner (2014)

fMRI ← where

MEG, EEG ← when

! Alexia part appears when you learn to read and it's always in the same place of the brain.

→ hierarchical theory of object recognition

Part 2

How human vision influenced computer vision?

→ Gabor filters

→ computer vision meets deep learning

→ solve invariant object recognition

→ ImageNet & AlexNet

* Krizhevsky, Suts... (2012)
* Fukushima (1980) Neocognition
* LeNet, LeCun
→ poloclub github cnn explainer
* Lindsay (2020) computational neural networks
* Zeiler & Fergus (2014)
* Gucly & van Gerven (2015)

LeCun et al, Nature (2015)

O'reilly (2020) (?)

diCarlo & Cox (2007)

* The Woreds I See Fei-Fei Li

Part 3

→ Do CNN networks predict (features) brain activity?

→ HCNN layers (?)

→ Grouping similarity

→ Networks seem to mimic human brain

- Lindsay (2020)
- Yamins et al. (2014)
- Kriegeskorte (2014)
- Kriegeskorte & Mur (2012)
- King & Groen et al (2018)
- Doerig et al. (2022)
- Groen, Silson & Baker (2017)

→ Brain-Score Algonauts

→ Neuroconnectionist research cycle

→ object recognition & scene processing

Part 4 How computer vision & human vision (still) different?

→ retinal image vs cortical map.

→ transformers are better than neural networks

→ Neuroscience & computer vision work together, they feed eachother

→ global statistics extraction, feedback/recurrent connections, peripheral sampling, ecological training, etc. ? ! ?

- * Da Costa et. al. (2021)
- * Cheung, Wells ...
- * Muller, Scholte & Groen (2024)
- * Geirhos et. al. (2018)
- * Geirhos et. al. (2019)
- * Dapello, Marques et. al Simulating a Primary ...
- * An image worth 16x16 words
- * Tuli, Dasgupta, ... Are convolutional NNs more like com. vis.?
- * Conwell et. al (2024)

Andrew Tracks - Day 1 Lecture 2

"no learning"?

* Classical tracking
  → motion detection & tracking
  → individual vs. multiple targets

→ pixel-level motion
→ optic flow
→ background models

} detect movement not tracking ∴

* if we can model the motion, we can predict objects' future locations.
  → HCI, robotics, surveillance, medicine...

flow-field ↔ quiver plot

→ motion difference & background subtraction

two basic approaches in motion detection

motion detection

- multiple object tracking -

* Markov Chain Monte Carlo tracking
* Metropolis-Hastings' algorithm
  → chain of predictions

→ Khan Z. et.al. (2004) An MCMC...

* social extensions
  → behaviours
  → sharing motion information

- tracking -

* uncertainty
* we want to be able to PREDICT where our target will be, and UPDATE our guess with a measurement.
* Kalman Filter
* Particle filters
  → real tracking is often multimodal

→ Isard & Blake (1998) Condensation

* appearance model(?)
* contour tracking
* mixed-state condensation
  → transition probability
* behavioural recognition !
* the curse of dimensionality

} still relevant in the deep-learning era

Michael Wray - Day 2 Lecture 1

08.07

Egocentric Vision - Making Sense of the First-Person Perspective

* Steve Mann, WearCam, WearComp
  applications in Personal Imaging
* Wearable devices
* Not limited to humans
  ↳ dogs, robots, shopping cards
* Blue Sky Outcomes
  ↳ robotics to understand surroundings
  ↳ personal assistants
    - remind you what you forget
    - help you fix things
    - remind you a recipe's next step etc.
  ↳ healthcare for assisting people
  ↳ ethic & privacy is an issue.!
  ↳ Hololens, Google Glass, etc.?

Ego4D dataset
GTEA, BEOID, GTEA Gaze, UT Ego,
ADL, EGTEA Gaze +, EPIC-Kitchens-55
Charades-Ego, EPIC-Kitchens-100
↳ combining every dataset
Ego4D

Past — **Present** — **Future**

- Episodic memory
- Hand & object
- Audio visual diarization
- Social interaction
- Forecasting

HD-EPIC

Digital Twin of kitchen videos
- narration, fine grained annot
  audio annotation, object detection

- object-fixture assignment

- Blind-Language: Llama 3.2, Gemini Pro
- Video-Language: VideoLlama 2, LongVA, Gemini Pro

Ego-Centric Tasks

- video understanding tasks
  ↳ action recognition, VQA
- captioning, retrieval, grounding
  audio localisation, activity & object recognition

→ EgoVis workshop

HierVL: learning Hierarchical Video-Language Embeddings

- EgoVLP - previous study, not good (?)
- Assembly101 dataset (2022)
  ↳ assembly Hands
  ↳ SVEgoNet : single view ego net

LEGO: Learning Egocentric Action Frame Generation via Visual Instruction Tuning

---

Privacy & Ethics

- Egocentric data is highly personal
- capturing should be done carefully
- biases can be added easily.
- indoors only, other people are not involved & recorders review their videos.
- anonymisation / de-identification

Prof. Victor Sanchez - Warwick

Day 2 Lecture 2

CLIP: Contrastive Language-Image Pretraining for Video Analytics

* Deepfakes, IAPR head of.
* SIPlab,
- multimedia forensics & biometrics
  - ↳ platform provenance

Project: multicamera Trajectory Forecasting tracking via person re-identification

Styles, Guha, Sanchez (2021) Pattern Analysis...
Nooker, Sanchez (2025) WACV

---

* CV + DL

which? → when? → where?

trajectory tensor

→ Enc Dec → which, when, where? CNN-GRU

mapping features to outputs

feature extraction

historical data as input

→ used tensors instead of vectors to convey the uncertainty

* FC, CNN, GRU, tCNN
* binary cross-entropy loss

Tiny Faces project

MTCNN + IFS

Deyra, Sanchez, Li (2018)

Deyra, Shen, Bahadur ... (2025)

(Automatic Face & Gesture Recogn.)

IEEE

Tiny Faces IFS detector

↳ Fisher Vector

* IFS to FVs → reduces the # of operations

Krishnashitha & Guha (2018)

(ICIP)

MEVA dataset

multiview extended video with activities

Qi, Tan, Yao, ... (2023) – Yolo5face

Yu, Huang ... (2024) – YoloFacev2

tinyfaces is still a challage to be solved.!

———— o ————

Recall: CLIP is proposed to be used in downstream tasks

→ WebImageText (WIT)

CLIP for segmentation

Lüddecke, Ecker (2022) → CVF

DACE: Distance-Aware Cross-Entropy Loss

HuggingFace: spaces/yiming-M/CLIP-EBC (?)

Crowd Counting

→ rely on annotated 2D coordinates of individuals' head centres in images.

Ma, Sonche2, Guha (2022) → ICIP

Ma, Sonche2, Guha (2025) → ICME (CLIP-EBC)

↳ Reframe crowd counting as a classification task

⇓

they introduced this idea to CLIP.

* zero-shot learning: can we facilitate?

↳ natural language supervision

* contrastive learning

Radford, Kim, Hallacy ... learning transferable visual models from natual language supervision

* symmetric crossentropy loss

Attention is all you need

Encoders of CLIP => ResNet50: attention pooling

Vision Transformer

Dr Andrew Gilbert - UV Surrey

Day 2 Lecture 3

Impact & Application

Surveillance
Healthcare
Autonomous Vehicles
Entertainment

From Paragraphs to Pixels: LLMs in Video Understanding

@andrewjohngilbert.github.io

* Video understanding
  - ↳ long range temporal dependencies
  - ↳ ambiguous / overlapping events
  - ⋮
  - lots of challenges in this area.

YouTube: WhoDunnIt? It is easy to miss stuff you're not looking for.

* Video Understanding with Large Language Models: A Survey, Yunlong Tang et. al.

* Unsupervised learning of Visual Representations Using Videos, ICCV 2015
* Shuffle Learn ECCV (2016)
* Self-Supervised Video Rep CVPR 2017
* Arrow of Time, CVPR 2018 (?)
* ECCV 2018 - Tracking Emerges by Colorizing Videos
* VideoMAE, NeurIPS 2022
* MOFO:, NeurIPS 2022 workshop
* CVPR 2023 - learning Video ..
* Vid2Seq CVPR 2023
* VAST NeurIPS 2023, Sihan Chen
* FILS: Self-Supervised Video Feature Prediction...
* Fine-Grained audible video description
* SWINBert (?)
* AutoAD: CUF 2023
- → ? Donte AD ?

gap between pixels & nl reasoning

AI Storytelling

Art in Storytelling.

* Visual Agents in Software
  - ↳ unreal engine
  - or Game Engines nowtey

} Visual Interface Agent.

* MAGMA: Multimodal Agentic Foundation

  Yang et al. 2025

* UI-TARS, 2025

Audio-Visual Fusion (?)

* Ramazonaw... "OWL..." CUF 2023
* Egocentric audio-visual object localization, CUF 2023
* Self-supervised moving vehicle...
* Epic-Fusion: Audio visual temporal binding for egocentric action recognition

→ Dataset: Youtube8M THUMOS:

Local (?)
Learnable Gating
Cross Attention

* Multi-Resolution Audio-Visual Feature Fusion...
* DEL: Dense Event Localisation...? 2025

→ Action Quality Assessment ↔ related to this. ECCV 2022 with temporal parsing transformer

* learning to Score Olympic Events

Oisin Mac Aodha - U. of Edinburgh

Day 2 lecture 6

Representation learning -

@ sthalles.github.io/a-few-words-on-representation-learning

* bad vs good features

* Once trained, the representation can be transferred to other tasks.

- Wei Koh et. al. Concept Bottleneck Models, ICML 2020

* success depends on data representation

↳ what is good representation?

- compact (minimal)
- explanatory (sufficient)
- disentangled (independent factors)
- interpretable
- make subsequent problem easier

Eli Cole, Caltech (slide credits)

* Transfer learning (ImageNet)

- Bengio et al. Representation learning: A Review and Perspectives

credit: Justin Johnson eecs498

Self-Supervised learning

"Most of human and animal learning ..." ~Yann lecun (On true AI)

- How to learn features from the unlabelled data?

* Papers With Code: Self Supervised Image Classification

- Effective methods for learning representations.
- Gidaris et al. Unsupervised Repr. learning... ICLR 201
- Weng & Kim, Self-Supervised learning, NeurIPS 2021

Self-prediction

SimCLR (2018)

- Chen et. al. A Simple framework for contrastive learning ...
- Github: google-research/simclr

Multi-modal Contrastive learning

- Radford et. al. learning Transferable Visual models from NL Supervision ICML 2021

* zero-shot classification

* there are limitations

↳ can require large batch size

↳ relies on "good" negative selection

↳ space of plausible augmentations can be dataset specific and must be defined in advance

* masked language models

↳ BERT masked language models

* vision transformers (ViT)

- Dosovitskiy et. al. An Image is worth 16x16 Words. ICLR
- He et al. Masked Autoencoders Are Scalable Vision learners CVPR 2022

- Zhang et al. 2016, 2017
- Doersch et al. 2015
- Noroozi & Favaro, 2016?
- Noroozi et. al. 2017

Neill DF Campbell - UCL

Day 3 lecture 1

09.07

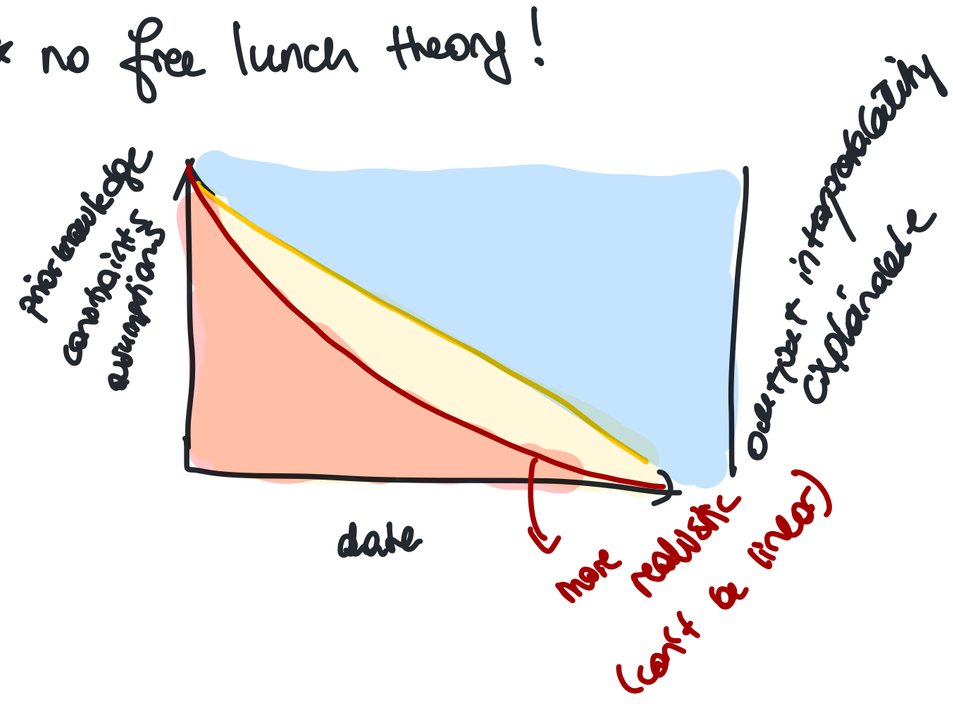
Uncertainty & Evaluation in Computer Vision

- Poggi ... On the uncertainty of self-supervised...

* why do we care about uncertainty?=
  - ↳ ambiguity in task, in our models
  - ↳ downstream decision making
    - safe
    - robust
    - transparent
  - ↳ improved performance
    - data efficiency
    - self supervision
  - ↳ evaluation & model selection!

* there are danger if we avoid probabilistic approch

* no free lunch theory!



Laplace: "the theory probability is the common sense reduced to calculus." (?)

* statistical significance

→ Youtube: Crimes against data.

* Human36M dataset

* ablation studies

- Simple and Scalable predictive uncertainty estimation using deep ensembles
- on uncertainty of self-supervised monocular depth estimation
-

Recent advances

* non-contrastive Siamese networks
* BYOL, SimSiam
* DINO

- Emerging Properties in self-Supervised Vision Transformers, ICCV 2021

* DINOv2 builds on DINO & iBOT

- DinoV2 : learning Robust Visual Features...

Limits of SSL

* most SSL pretrained with ImageNet

- Cole et al. When Does Contrastive Visual Representation learning work ? CVPR'22

* dataset size matters
  - ↳ amount of unlabelled data for pretraining
  - ↳ amount of labelled data for supervised learning

* iNAT

- Van Horn et al Benchmarking... (?)
- Zong, Aodha... Self Supervised ...

Vision @ UoE

---

* Monty Hall problem.
* Bayes' rule

→ How can we make it better if we don't know where it fails...

→ Understanding Deep Learning - Simon Pierce

* Bias-variance trade-off