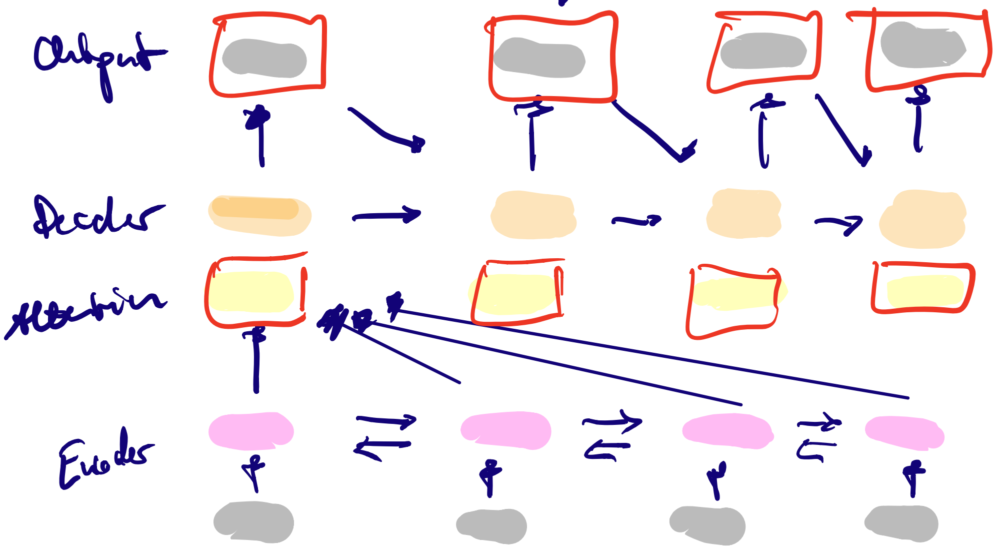
RANLP Conference – Bucharest – Day 1 — 2/10/19

Diana Inkpen – Detecting Signs of Mental Illness and Behavioral Issues from Online Communication (University of Ottawa)

- NLP-based user modeling → Big 5 model: (vs. message level)
  ① extraversion, ② emotional stability, ③ agreeableness, ④ conscientiousness, ⑤ openness to experiences
- Social media provides unique info source! → 'natural setting'
- Mental disorders: insomnia, distress, postpartum depression, PTSD → detectables!
  → a lot of old school NLP → but more recently deep learning end-to-end
- Datasets: 'Bell Let's Talk' (Twitter), CLPsych 2015 shared dataset (Twitter), George Lewis dataset (Reddit forum)
  DEPRESSION
  * CLPsych 2017 shared dataset ReachOut.com → labelled distress
  * Cyberbullying datasets
- Problem of deep learning: little data on user-level
  → share features ⇒ Multi-Task Deep Learning (different mental disorders)
  → train models independently and expose layer sharing!
- FLAIR → Embedding → Conv → Predict Out // Bidirectional LSTM
  (Word level, Character level, Multi-Lingual)
  ⇒ often combining tasks can improve performance for all tasks!
  → Form of data augmentation as in FLAIR case! More labels per point!
- Compared different pre-trained embedding techniques! → GloVe ⊕ end-to-end optimized with
- Message level models are hard due to noise/little data!
  → but: Ensures more privacy security! // Population-based: Easiest to analyze/extract insights for policy dept!

Vlad Niculae - Sparse Sequence-to-Sequence Models

- S-to-S w. Attention → Bahdanau et al 2015
- Architecture classically:



SOFTMAX used at times:
- Attention weights
- Output prob.

Bi-Directional LSTM

- Usually beamsearch used to artificially induce sparsity as form of post-processing.

- HERE → Peters and Way 2019 ⇒ Sparse Attention Weights/Alignments
  ↳ Softmax: $p_i = \exp z_i / \sum \exp z_j$ ⇒ Simplex: discrete probability over classes

softmax maps to set

  ⇒ maximizes expected score + entropy
  → Generalize different entropy measure!

$$\pi_H(z) = \arg\max_{p \in \Delta} p^T z + H(p)$$

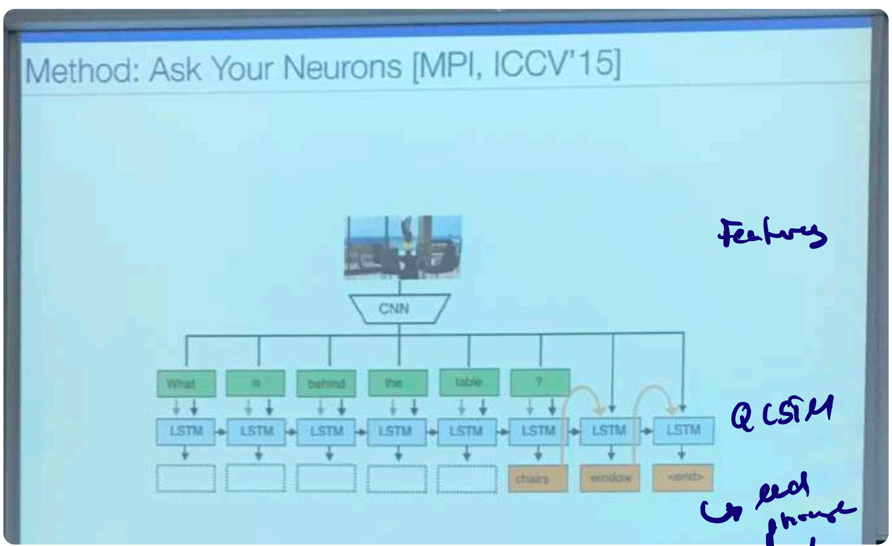
  ↳ $H^0(p) = 0$ → argmax
  ↳ $H^S(p) = -\sum_j p_j \log p_j$ → softmax
  ↳ $H^G(p) = 1/2 \sum_j p_j (1 - p_j)$ → gini entropy → P & W 2019
  ↳ Here: $H^T_\alpha(p) = \frac{1}{\alpha(\alpha-1)} \sum_j (p_j - p_j^\alpha)$ → $\alpha = 2$: sparsemax; $\alpha = 1$: softmax
  ↳ if $\alpha \neq 1$ else $H^S(p)$

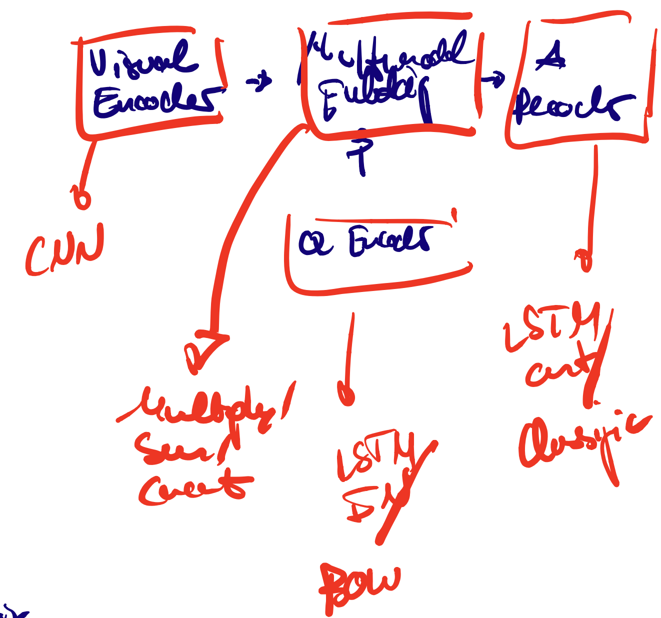
- Grid search: α for one fixed → learn the other
  ↳ try to learn end-to-end! → might be better to learn more same shape?

RAAI Conference – Bucharest – Day 2

Mateusz Malinowski – Grounded Language: Answering Qs about Images + Navigation (DeepMind)

- Visual Turing Test – 1 image ↔ many Qs → Multi-modal Task
- Visual Illusions! → Downstream behavior vs. Learned representations
  - ↳ Have to test algo on both in order to achieve human understanding
- Grounding language into percepts → association! (Motivation)
- S. Gershman → "Human Similarity in Humans and Machines"
- Asking different Qs → What is in image? → Object Recognition
  - ↳ How many are there? → ⊕ Counting
- Architecture Design:



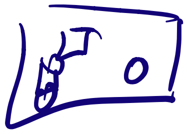


- Multimodal Embedding: Mapping of the arch. to vector → univariate signal
  - ↳ Concatenation: Form of Interaction!
  - ↳ Hadamard: $f = p^T (U^{(l)T} x \odot V^{(l)} y)$ — x: 1st Modality, y: 2nd Modality → Qs
  - ⊕ Sum: FiLM
- ↔ Relation to Bilinear Pooling: $x^T U^{(l)} V^{(l)} y = 1^T (U^{(l)T} x \odot V^{(l)} y)$, with $U^{(l)} V^{(l)} = W^{(l)}$ — Can lead to dim. red.
  - ↳ Quadratic ↔ Hadamard — Compressed via low-rank [?]

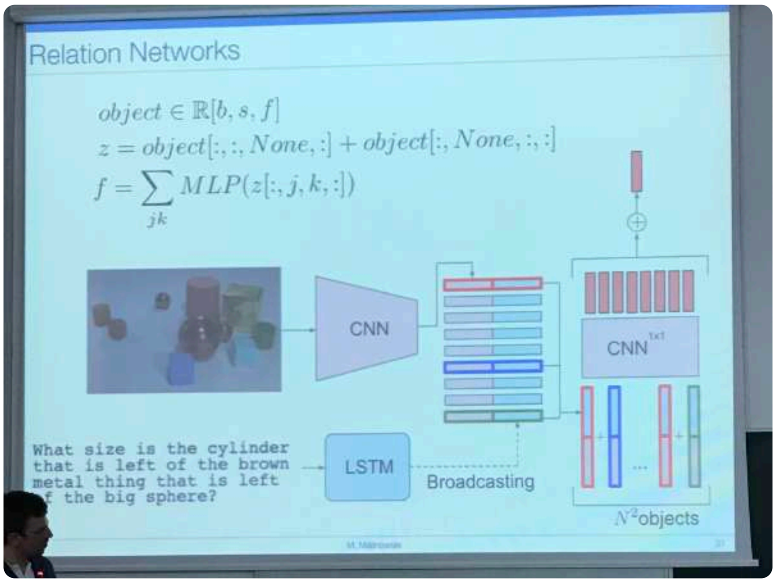
→ Could theoretically add auditory percepts to x LSTM streams

→ Usually only pretrain the vision component but not the embedding

↳ otherwise vision component tends to overfit!

• Data: CLEVR → shapes  → Facebook ⇒ Somewhat solved!

• Relation Networks:



→ Problem of scalability ⇒ N^2 iterations

↳ Reduce to k^2 subset

⇓

('Hard Attention')

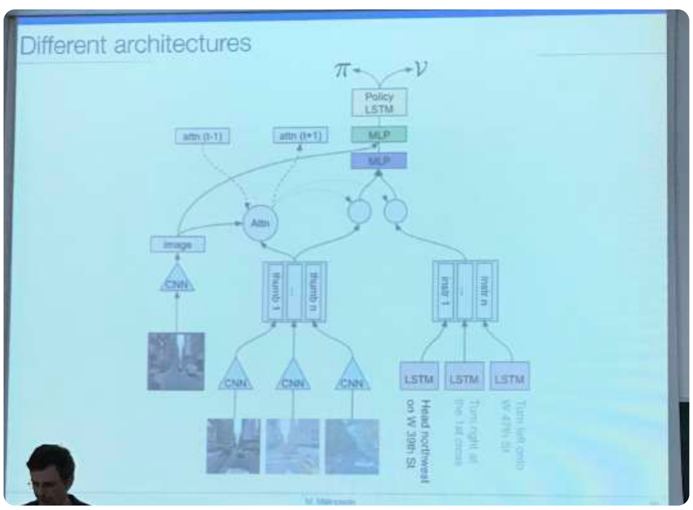
↳ Santoro et al (2018) – Relational Reasoning

↳ Don't need to attend to full object

• want to work with ambiguous naturalistic images and not abstract graphic renderings!

↳ Map navigation based on instructions → Street View + Google Maps

↳ Easy testing on different city environments!



• Different reward signals

↳ step-by-step, Incremental, Only Goal!

Agent has to discover relevant heuristics!

Still uses reward fct.

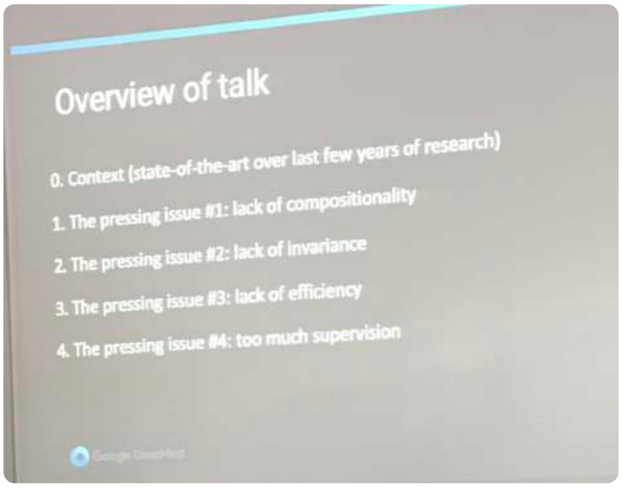
↳ Reward Shaping w. intermediate steps

• Supervised vs RL agents

• Vocabulary ↔ Options? → Hierarchy in language!

João Carreira — 'Pressing problems in video understanding'

- Primitive video understanding → rapidly evolved! Not learned!
- Social-IQ dataset CVPR 2019 → advanced video understanding



- Pascal VOC / COCO → segmentation task! datasets

↳ State of the art by transfer from Imagenet

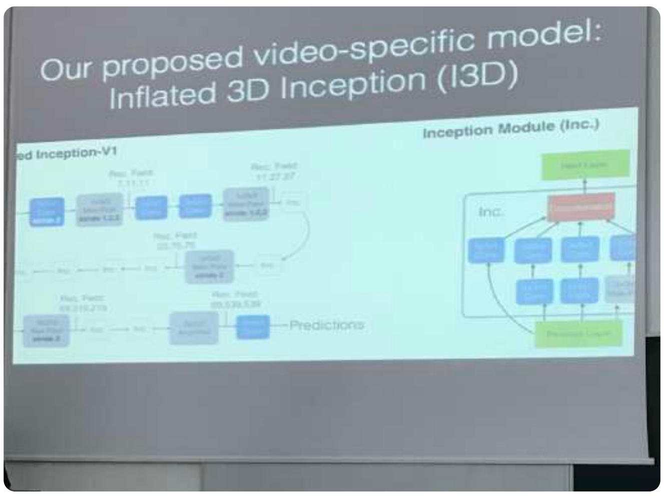
→ Did not really work for Videos!

↳ Have to learn directly from video → motion / time correlation?

→ Collect via Mech. Turk

- Kinetics dataset (300k videos) → aim to be 'Imagenet for video'

↳ 700 classes → 650k videos 2019!



- Model: Take Google Inception & add temporal dimension 3d convolutions

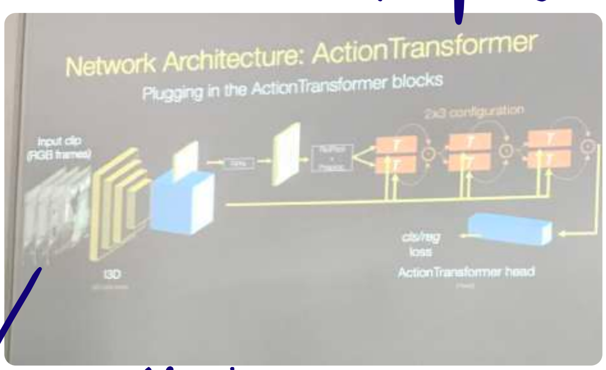
(+) Global pooling

→ No frame skipping! ⇒ Use all frames!

(1) Compositionality: Problem → Multiple actors + multiple actions

↳ requires attention to extract!

⇒ Action Transformer Blocks



↳ what do transformers do?

↳ Map Layer-specific heads to 3d

⇒ Look at RGB images

↳ Detect people // Track people

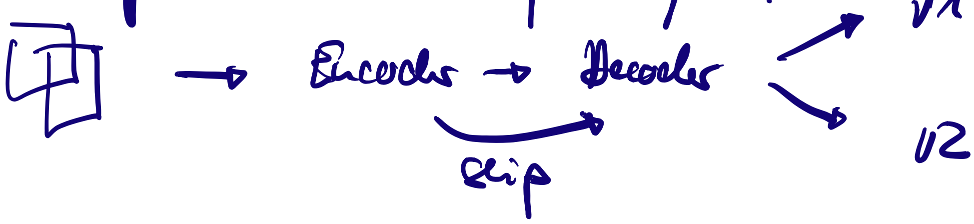
Processing full video!

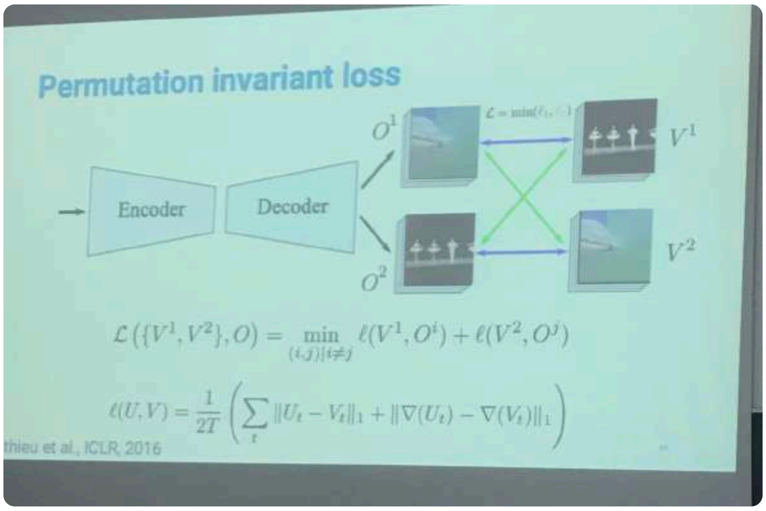
② Invariance → Get rid of nuisance factors (shadows, etc.)

→ Layer decomposition of images! → Reflection / Occlusion = Lin. Comb. Eqs

→ Usually optimization → What is good way to learn? → Self-Supervised

↳ Average two videos in pixel space!





- Works on synthetic data
  ↳ How to transfer to the wild?
- Get real invariance to features of naturalistic images!
  ↳ Go to occlusion!
- Test different data gen. methods
  ↳ Not only linear but perhaps or GAN generated

③ Efficiency → Fast + Online (no batch!) processing

→ Brain = Massively parallel → Not sequential

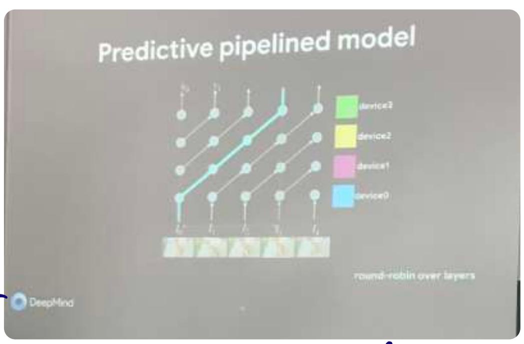
↳ Depth prediction processing → Improve throughput by processing on different devices

Spread over devices! ←

↳ Go predictive to improve latency issues

→ Trade-off between latency & accuracy!

↳ Fully sequential → Offline & but more accurate!



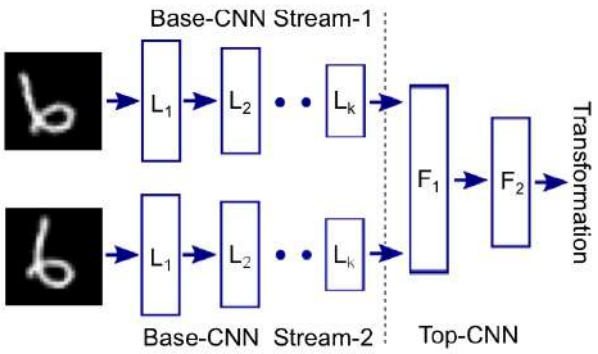
↳ Parallelize layers!

④ Human Supervision ⇒ Self-Supervised Learning

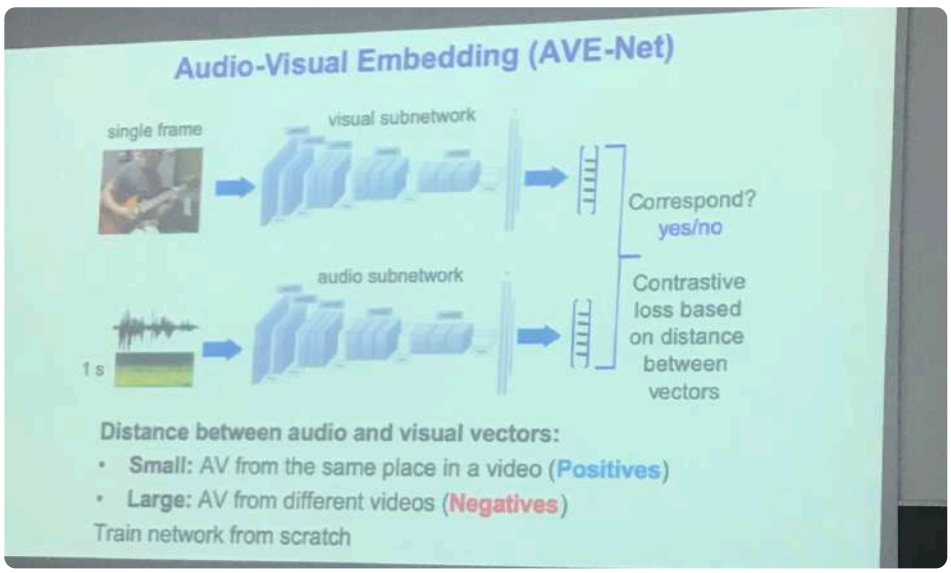
→ Video passive ⇒ Active inferring perception → Not always good!

J.J. Gibson – "We move in order to see and we see in order to move."

→ Visual features should predict motion!



→ Fusion of two streams predictions that feed into FC layers



→ With audio

⇒ Objects that sound

↓

Disconnect URL

↓

also by transfer tasks to segment!

↓

Meta-Learning ⇒ few shot learning!

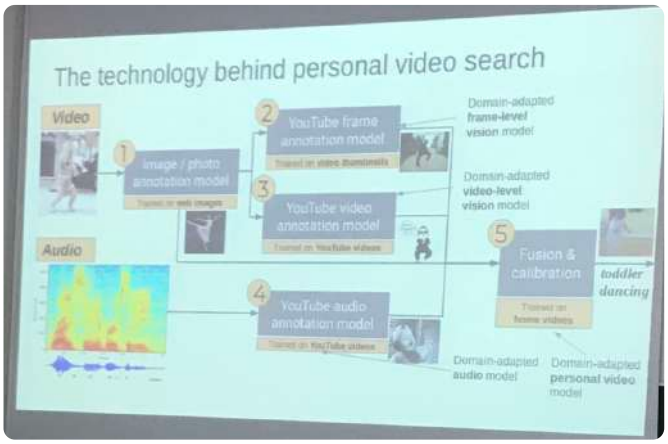
→ For NLP self-supervised is already better than supervised → BERT

- General Idea: Train in self-supervised fashion on massive data
  - ↳ perform few shot learning on supervised datasets
  - ↳ Is this the future? Bayesian NN → Active Inference?!
- Putting all 4 developments together?!

RAH Conference – Rehovot – Day 3

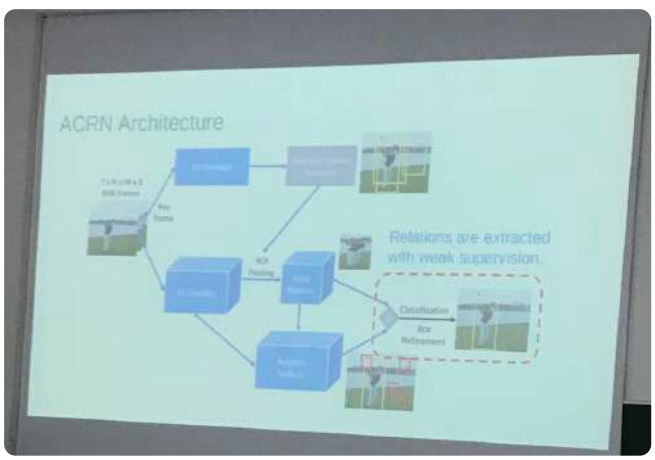
Rahul Sukthankar – Robot Research in Machine Perception

- NN Architectures + Labeled data + Infrastructure
- Comp. photography → High dynamic range imaging
- Privacy tooling and blurring on YouTube
- Connectome work @ Google #1



// Large-scale Video Annotation for YouTube //

- Importance of Meta-Data
  → Transfer from ET descriptions to private
  → Domain adaptation



// Actor-Centric Relation Network for Action Detection //

- Faster RCNN for object detection sequentially
  to need more spatio-temporal context
  to Here: Focus on actors + other objects over time + space
  to Graph Convolutional Nets → when are weakest/strongest edges associated with specific actions

// Pixel 3 Sharpness //

// Open Source datasets

- TF Object Detection API
- Open Images V5
- Audio Set

// Weakly / Self-Supervised Learning on YouTube DB //

- Fusion of View Synthesis / Learn depth + pose of Camera
- LDI – layered depth image
- NN = differentiable Renderer
- Mannequin Challenge → Static + Different views

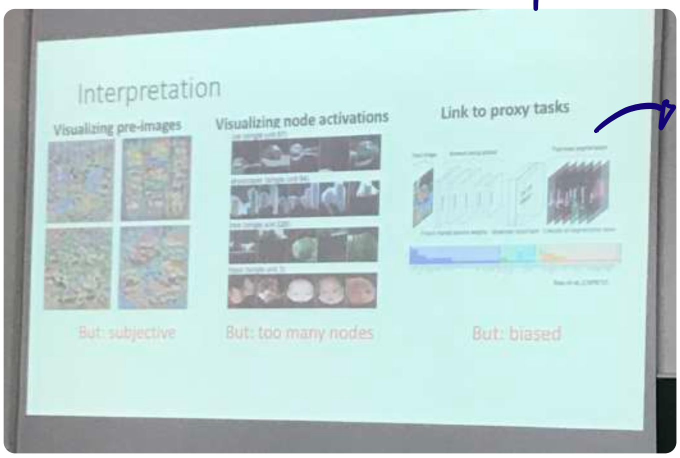
Tinne Tuytelaars – Explanation by Interpretation

- 'Interpretability is the degree to which a human can understand the cause of a decision' – Miller, 2017
  → subjective → Need to adapt to the recipient of explanation
  → explain vs. understand → Teacher might be able to understand a concept but not be good at explaining it → LANGUAGE
- EU guidelines → predictive systems must provide explanation of internal model

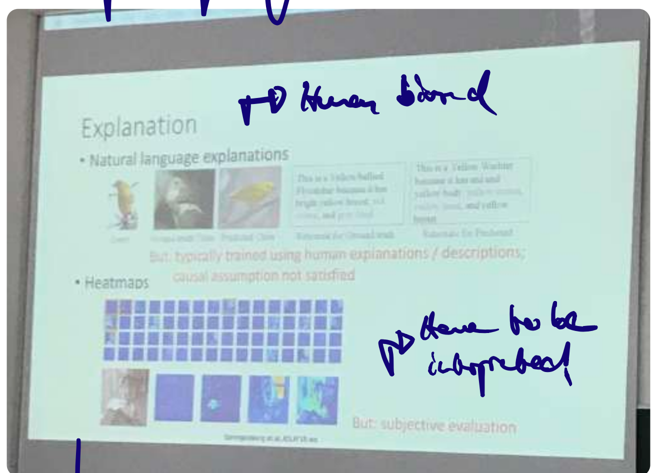
↳ Ease of interpretation → towards humans → images / natural language → SUBJECTIVE

INTERPRETATION: explain model → introspection

EXPLANATION: explain decision → input-specific



Form of semantic segmentation on top!



→ Human based

→ have to be interpreted

↳ Proxy task: Zhou et al CVPR 16

↳ User study: Zeiler & Fergus ECCV 14
→ Visualize only frame?

Problems: ① Too many neurons
↳ Which neurons matter?!
↳ Reduce # visualizations via guided backprop + Guided Backprop
↳ Select top k-images with highest response of relevant features!

② Visualizing heatmaps ↳ Guided BP not for class label but for relevant features

③ Evaluation
↳ strongly biased to what humans expect! → Go to synthetic data!
↳ Synthetic → Have access to ground truth masks! → Can compare heatmaps