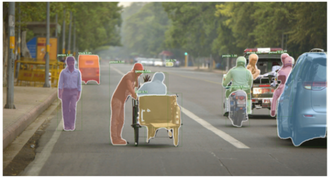
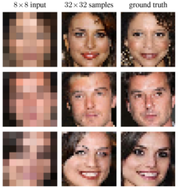
Possibilities and Challenges for Artificial Intelligence in Military Applications

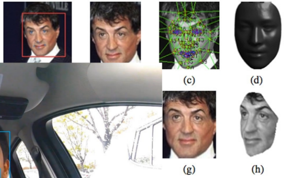
Peter Svenmarck, Linus Luotsinen,
Mattias Nilsson and Johan Schubert

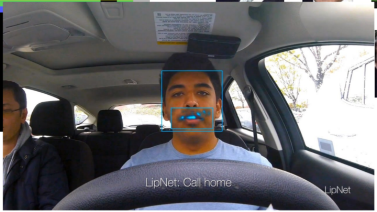
May 31, 2018

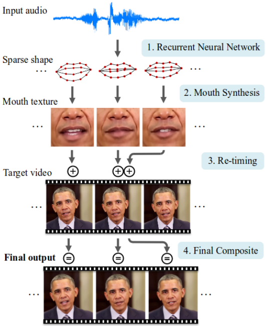
DL Boosts Performance in a Large Number of Applications











Potential Advantages of DL

- Efficiency:
    - Reduced development costs and development time
- Availability:
    - No programming skills required (software 2.0)
- Complexity:
    - Computer generated programs perform better than any human implementation
- Creativity:
    - Computers provide creative solutions to problems that humans can study and learn from
- Objective:
    - Computers are unbiased and fair whereas humans can be corrupt, unfair, racist and so on

Examples of Military AI-Applications

- Maritime surveillance
  - Unsupervised machine learning
  - Low probability events are anomalies
- Underwater mine warfare
  - Supervised machine learning
  - Image classification
- Intrusion detection
  - Supervised machine learning
  - Signature classification
- Penetration testing
  - Deep reinforcement learning
  - Planning of mitigation strategies

Challenges

- Optimization:
  - Local vs. global
- Generalization:
  - Under-fitting vs. over-fitting
- Hyper-parameter tuning:
  - Meta-learning
- Production grade AI:
  - Reproducibility
  - Version control for data
  - Power efficiency
  - Real-time processing
  - Up to date after deployment
- AI-compute and data centers

- **Black-box:**
  - **Transparency, interpretability, explainability**
- **Vulnerabilities:**
  - **Adversarial examples, transfer learning and data poisoning**
- **Data:**
  - **Learning with limited data**

Transparency, Interpretability, and Explainability

- Types of need
  - Trust
  - Causal relationships
  - Generalizability
  - Inform decision making
  - Fairness

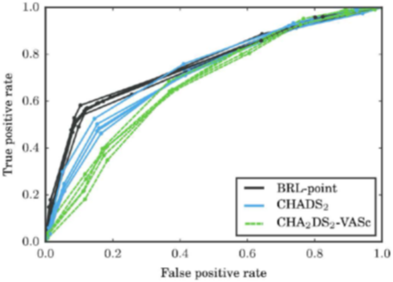
Approaches for Transparency

- Interpretable models
  - Linear models, Rule-based systems, Decision trees
  - Predictability, Decomposability, Training method
- Explanations
  - Textual or visual
  - Perceived beliefs, desires, and intentions
  - Abnormality, Preferences, Norms, Recency, Controllability
  - Contrast relative other recommendation
  - Selective
  - Conversations for transfer of knowledge

Examples of Interpretable Models

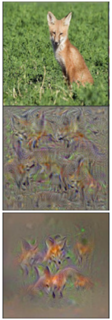
- Bayesian Rule List
- Stroke Prediction

if hemiplegia **and** age > 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age > 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age ≤ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)



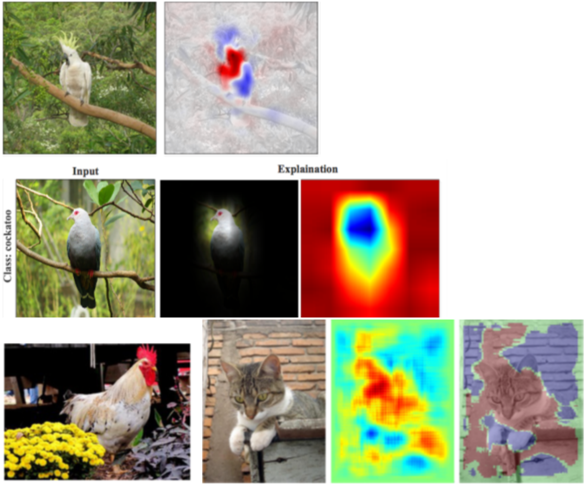
Examples of Feature Visualization: Activation maximation

- Semantic information in images is spread out
- Multifaceted features
- Synthesize images with GAN for
  - Coherent global structure
  - Realistic looking colors
  - Sharpness



Examples of Feature Visualization: DNN explanation

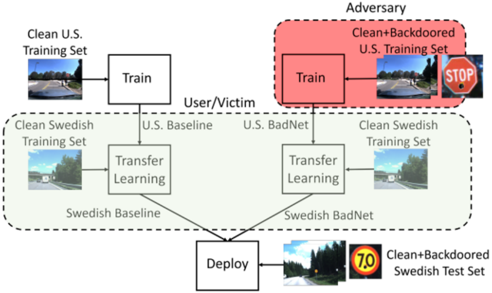
- Highlight discriminative features or regions
- Sensitivity methods are vulnerable to occlusion
- Relevance propagation considers both presence and reaction



Vulnerabilities

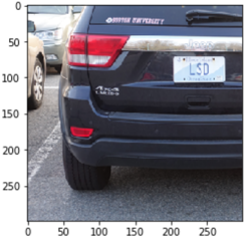
- Adversarial examples:
  - It is easy to adjust the input so that the classification system fails completely
  - The main idea is to use SGD and back-prop as usual, but instead of updating weights the input signal is updated
  - When input dimensionality is large then the changes are often imperceptible
  - Black-box attacks are also possible

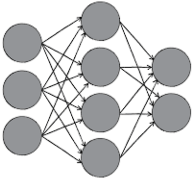
- Transfer learning:
  - The idea is to exploit hidden backdoors in pre-trained DNNs

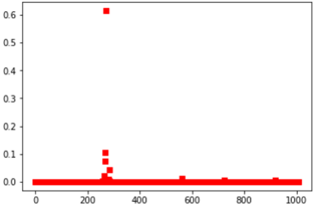


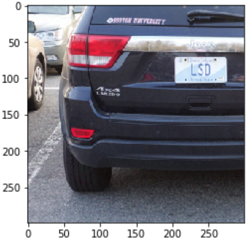
Example 1: Manipulation of Input Signal

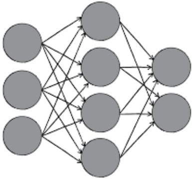
Input | Model | Output

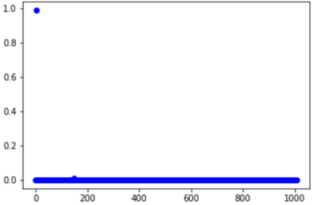






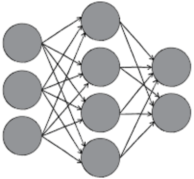


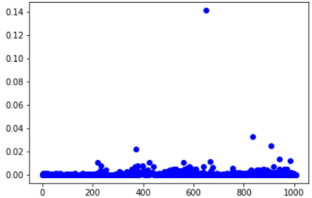


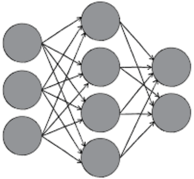


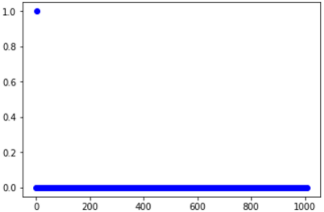
Example 2: Manipulation of Input Signal

Input | Model | Output

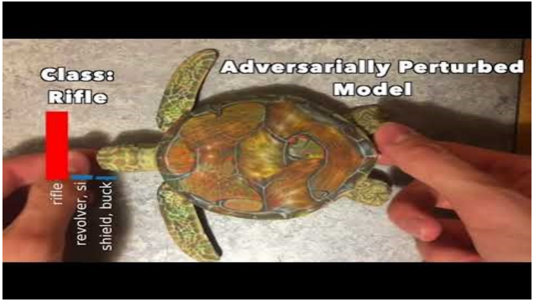








Example 3: Manipulation of Input Signal



Vulnerabilities

- Even though this is a hot research area, there are no solutions to these problems
- Defence mechanisms exists but they do not always work
- Recommendation:
  - Always protect the model, its architecture and weights
  - Minimize the possibility for outsiders to interact with the model
  - Be careful when using transfer learning
  - When reusing training data, always check for poisoning

Learning with Limited Data

- Data for military ML-applications is limited:
    - Data is collected but typically not for ML-purposes
    - Data is not easily shared
- Techniques that can be used to learn with limited data:
    - Transfer learning
    - Generative Adversarial Networks (GANs)
    - Modeling and simulation

Conclusions

- There is currently no silver bullet for the challenges highlighted in this talk
- But, the AI-field is moving fast:
  - Partial solutions continues to emerge
  - Keeping up-to-date is a challenge
- More AI-applications are reaching human or even superhuman performance
- Many AI-services are now available as products on the cloud (transcribing, sentiment analysis, face recognition, etc.)
- Deep learning solves domain specific tasks only:
  - Other breakthroughs are needed for AGI

Questions?

Thanks for listening

Acknowledgment

This work was supported by the FOI research project “AI for decision support and cognitive systems”, which is funded by the R&D programme of the Swedish Armed Forces.