Robustness in AI

Nicolas THOME – Prof at Sorbonne University
ISIR Lab, MLIA Team
M. Careil’s PhD Defense
Context: AI/ML summer

- AI in the last decade: huge performance boost
  - Vision, NLP, multi-modal prediction, robotics
Context: robustness in deep learning

Several brittleness aspects in deep learning models
• Explainability, biases & shortcuts, fairness, etc

i) Stability: adversarial examples, mistake severity
ii) Uncertainty Quantification (UQ)

“Know when you do not know”

Abstain to make a prediction

UQ: a challenge in DL

- Which uncertainty score?
- Calibration, ranking correct/incorrect prediction
Context: robustness in deep learning

iii) Training: direct optimization of target metrics, image retrieval

- Non-differentiable losses

  “Good” (with guarantees) surrogates for classification, what about other metrics, rank losses?

- Non-decomposable losses

  - Rank losses, e.g. Average Precision

  \[ \frac{1}{3} \sum_{k=1}^{3} AP(b_k) = 0.78 \]

- Many other losses (IoU, Dice in segmentation), including global constraints (fairness)
Improving robustness in deep learning

1. Uncertainty quantification
2. Direct optimization of rank losses
3. Controlling mistake severity
Sources of uncertainty

• Aleatoric uncertainty: data
  • Class confusion, ambiguous data, sensor noise

• Epistemic uncertainty: model
  • Distribution shift in $p(x,y)$, e.g. $x$ (snow, image->cartoon), or $y$ (open set, new classes)
Uncertainty quantification in deep learning

• **Uncertainty for failure prediction [CBT+19]:** correct vs incorrect predictions

• **Our proposal:** True Class probability (TCP) vs Maximum Class Probability (MCP)

• TCP better than MCP for failure prediction

Uncertainty quantification in deep learning

TCP unknown at test time: learning it! => ConfidNet

- Pre-trained prediction model (blue)
- Learning to regress TCP with an auxiliary model (orange)

\[ L_{\text{conf}}(\theta; D) = \frac{1}{N} \sum_{i=1}^{N} (\hat{c}(x_i, \theta) - c^*(x_i, y_i^*))^2 \]

## Results

<table>
<thead>
<tr>
<th></th>
<th>MNIST</th>
<th>SVHN</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>CamVis</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP [Hendrycks &amp; Gimpel, 2017]</td>
<td>47.3 ± 1.7</td>
<td>36.1 ± 3.6</td>
<td>46.2 ± 0.5</td>
<td>48.4 ± 0.7</td>
<td>71.3 ± 0.4</td>
</tr>
<tr>
<td>MC Dropout [Gal et Graharnani, 2015]</td>
<td>41.0 ± 1.2</td>
<td>42.1 ± 5.5</td>
<td>45.2 ± 1.3</td>
<td>48.1 ± 1.0</td>
<td>71.9 ± 0.7</td>
</tr>
<tr>
<td>TrustScore [Jiang et al, 2019]</td>
<td>52.1 ± 1.8</td>
<td>33.5 ± 3.8</td>
<td>44.8 ± 1.3</td>
<td>41.8 ± 2.0</td>
<td>66.8 ± 0.5</td>
</tr>
<tr>
<td>ConfidNet</td>
<td><strong>59.7 ± 1.9</strong></td>
<td><strong>45.5 ± 3.8</strong></td>
<td><strong>48.6 ± 1.0</strong></td>
<td><strong>53.7 ± 0.6</strong></td>
<td><strong>73.6 ± 0.6</strong></td>
</tr>
</tbody>
</table>

AP errors (%)
Learning confidence for self-labelling

- Extension for domain adaptation [CTS+21]

Learning confidence for self-labelling

• Extension for Medical image segmentation [PTS21]

Out-Of-Distribution (OOD) detection

• Post-hoc OOD detection: leveraging any state-of-the-art prediction model
• Accurate OOD detection $\Leftrightarrow$ accurate in-distribution (ID) density estimation
  • State-of-the-art ID density estimation: prior densities, e.g., GMM, Energy Logits (EL)

- Prior density: not accurate $\Rightarrow$ **Energy correction**
  - GMM good for far-OOD, EL for near-OOD $\Rightarrow$ **Energy composition**
• **HEAT [LRR+23]:** Hybrid Energy Based Model (EBM) in the feature space for OOD detection

- **Energy-based correction** of prior energy terms, e.g. Gaussians
- **Energy composition** of several terms (Gaussian, Energy Logits, std for style)

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>Imagenet</th>
</tr>
</thead>
<tbody>
<tr>
<td>EL (NeurIPS, 2020)</td>
<td>32.6</td>
<td>74.3</td>
<td>57.0</td>
</tr>
<tr>
<td>SSD+ (ICLR, 2021)</td>
<td>35.1</td>
<td>66.2</td>
<td>49.7</td>
</tr>
<tr>
<td>DICE (ECCV, 2022)</td>
<td>35.1</td>
<td>69.2</td>
<td>35.9</td>
</tr>
<tr>
<td>KNN (ICML, 2022)</td>
<td>32.78</td>
<td>68.8</td>
<td>55.0</td>
</tr>
<tr>
<td>ViM (CVPR, 2022)</td>
<td>33.0</td>
<td>63.9</td>
<td>48.4</td>
</tr>
<tr>
<td>HEAT (Ours, ICML 2023)</td>
<td>23.5</td>
<td></td>
<td>34.4</td>
</tr>
</tbody>
</table>
Robustness: recent contributions

1. Uncertainty quantification
2. Direct optimization of rank losses
3. Robustness
Direct optimization of rank losses for image retrieval

1. Theoretically sound surrogates for non-differentiable rank losses, e.g., Average Precision (AP)
2. Reducing the decomposability gap

\[ \mathcal{L}_\text{SupAP}(\theta) = \sum_{j \in \Omega^-} H^{-}(s_j - s_k) \]

\[ \mathcal{L}_\text{DG}(\theta) = \sum_{j \in \Omega^+} H^+(s_j - s_k) \]

\[ \mathcal{L}_\text{ROADMAP}(\theta) = (1 - \lambda) \cdot \mathcal{L}_\text{SupAP}(\theta) + \lambda \cdot \mathcal{L}_\text{DG}(\theta) \]
Image retrieval: non-smooth metrics & losses

- Standard losses, e.g., triplet loss, NSM [A]
  - Coarse upper-bounds, not well-aligned with metrics: supports bottom vs. top of the ranking

- Upper bounds: structural SVMs, Blackbox optim [B]
  - General methods, theoretical guarantees
  - Coarse upper bounds

- Rank approximation: binning approaches, smoothAP [C,D]
  - Tighter approximations
  - No theoretical guarantees

---

[A] A. Zhai, and H.Y Wu. Classification is a strong baseline for deep metric learning. BMVC 2018
[C] A. Brown, W. Xie, V. Kalogeiton, A. Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. ECCV 2020
[D] Y. Patel, G. Tolias, and J. Matas, “Recall@ k surrogate loss with large batches and similarity mixup,” in CVPR, 2022
Image retrieval: addressing non-decomposability

Fewer works, brute-force approaches

• Sampling informative batches or constraints in batch
• Storing the datasets, e.g., x-batch memory [D]: increased in memory
• Large batches + 2-step approach for AP and back-prop [E]: increase in training time

Robust and decomposable AP (ROADMAP)

\[
\text{AP} = \frac{1}{|\Omega^+|} \sum_{k \in \Omega^+} \frac{\text{rank}^+(k)}{\text{rank}(k)} = \frac{1}{|\Omega^+|} \sum_{k \in \Omega^+} \frac{\text{rank}^+(k)}{\text{rank}^+(k) + \text{rank}^-(k)}
\]

\[
\text{rank}(k) = 1 + \sum_{j \in \Omega} H(s_j - s_k)
\]

- Optimizing rank \(\rightarrow\) smooth approximation & upper bound of AP
- Not optimizing rank \(^+\) \(\rightarrow\) well-behaved gradients.

\[
\mathcal{L}_{\text{SupAP}} = 1 - \frac{1}{|\Omega^+|} \sum_{k \in \Omega^+} \frac{\text{rank}^+(k)}{\text{rank}^+(k) + \text{rank}^-(k)}
\]

Better than
Improving decomposability

Decomposability gap:
\[ DG(\Omega) = \frac{1}{|B|} \sum_{b \in B} M(b) - M(\Omega) \]

- Calibrates scores across batches
  - Positive scores \( \geq \alpha \)
  - Negative scores \(< \beta \)

- Proof: Theore an upper-bound on the decomposability gap

\[ L_{DG}(\theta) = \frac{1}{|\Omega^+|} \sum_{x_j \in \Omega^+} [\alpha - s_j]^+ + \frac{1}{|\Omega^-|} \sum_{x_j \in \Omega^-} [s_j - \beta]^+ \]

\[ L_{ROADMAP}(\theta) = (1 - \lambda) \cdot L_{SupAP}(\theta) + \lambda \cdot L_{DG}(\theta) \]
Robustness: recent contributions

1. Uncertainty quantification
2. Direct optimization of rank losses
3. Controlling mistake severity
Hierarchical Image Retrieval for Robust Ranking

- Binary image retrieval → do not take into account mistake severity
- HAPPIER: Hierarchical Average Precision training for Pertinent Image Retrieval
  Extending AP to graded setting to take importance of errors into account

![Diagram showing HAPPIER results with query images and ranked retrieved images with H-AP and AP values]
Relevance function: graded similarities

- Relations between categories → proxy for mistake severity.
- Relevance function: \textit{graded similarity} between categories
- Decreasing function of the distance in the hierarchical tree.

\[
\text{rel}(k) = \frac{l/L}{|\Omega(l)|}
\]

- $l$: level of the closest ancestor in the tree.
- $L$: total number of levels.
Hierarchical average precision ($\mathcal{H}$-AP)

$$\mathcal{H}\text{-rank}(k) = \text{rel}(k) + \sum_{j \in \Omega^+} \min(\text{rel}(k), \text{rel}(j)) \cdot H(s_j - s_k)$$

- Errors in ranking → weighted by relevance
- Correct $\mathcal{H}$-rank → decreasing order of relevance

$$\mathcal{H}$-AP = \frac{1}{\sum_{k \in \Omega^+} \text{rel}(k)} \sum_{k \in \Omega^+} \frac{\mathcal{H}\text{-rank}(k)}{\text{rank}(k)}$$

- Consistent generalization of AP.
- Flexible wrt. the relevance
\[ \mathcal{L}_{\text{Sup-H-AP}}(\theta) = 1 - \sum_{k \in \Omega^+} \frac{1}{\text{rel}(k)} \sum_{k \in \Omega^+} \frac{\mathcal{H}\text{-rank}(k)}{\text{rank}^+(k) + \text{rank}_s^-(k)} \]
Results

• On par for fine-grained retrieval (“Species”)
• **Large gains** on other hierarchical levels from “Family”
ℋ-GLDv2: a hierarchical landmark dataset

GLDv2 → large scale landmarks retrieval dataset [F]

No hierarchical annotations → how difficult is it to create hierarchical annotations?

ℋ-GLDv2

1. Scraping Wikimedia Commons
2. Post-processing

Results

![Bar chart showing results for mAP@100 and H-AP.]
Perspectives

• Uncertainty quantification:
  • Global measure of uncertainty (aleatoric, epistemic)
  • For foundation models, e.g., CLIP
  • Test-time adaptation

• Non-smooth & non-decomposable metrics beyond ranking

• Mistake severity robustness
  • Adaptation to multi-modal models
Thank you for you attention!


HEAT: Energy correction

- Hybrid energy: $E^h_{\theta_k}(z) = E_{q_k}(z) + E_{\theta_k}(z)$
  $$p^h_{\theta_k}(z) = \frac{1}{Z(\theta_k)} \exp(-E^h_{\theta_k}(z))$$

- Controlling the residual $\mathcal{L}_C(\theta_k) = \mathbb{E}_{p_{in} \sim p_{\theta_k}} \left[(E^h_{\theta_k} - E_{q_k})^2\right]$
  - Correction with mimical norm

$$\mathcal{L}_{tot}(\theta_k) = \mathcal{L}_{MLE}(\theta_k) + \lambda \mathcal{L}_C(\theta_k) \quad (6)$$
HEAT: Energy composition

- Prior energy function, e.g. Gaussians or Energy Logit (EL)
  - Avg or Std (style) features as inputs for the energy model

- Energy composition:
  - $\beta \rightarrow +\infty$, max/
  - $\beta = -1$, logsumexp

$$E_{HEAT}^\beta = \frac{1}{\beta} \log \sum_{k=1}^{K} e^{\beta E_{\theta_k}^h}$$
1. Scraping Wikimedia Commons

Wikimedia Commons → largest open database of landmarks.

- GLDv2 sourced from wikimedia commons
- « Instance of » => super-category

Ex of scraped labels:
- Church building.
- Church building (1172–1954)
- Cathedral
- Castle
- Corsican nature reserve
- New Zealand great walks
- Waterfall
- Arch-gravity dam
- Canal
- Association football venue
- Astronomical observatory
- Village
2. Post-processing (manual + automatic)

- K-Means clustering from CLIP’s textual encoder
- Manual verification + adding natural/man made in hierarchy
- 78 super categories