Be Careful What You Smooth For

Label Smoothing Can Be a Privacy Shield but also a Catalyst for Model Inversion Attacks



Lukas Struppek [1 4] **Dominik Hintersdorf** [1 4] **Kristian Kersting** [1 2 3 4]

[1] Technical University of Darmstadt [2] Centre for Cognitive Science [3] Hessian Center for AI (hessian.AI) [4] German Research Center for AI (DFKI)

At a Glance

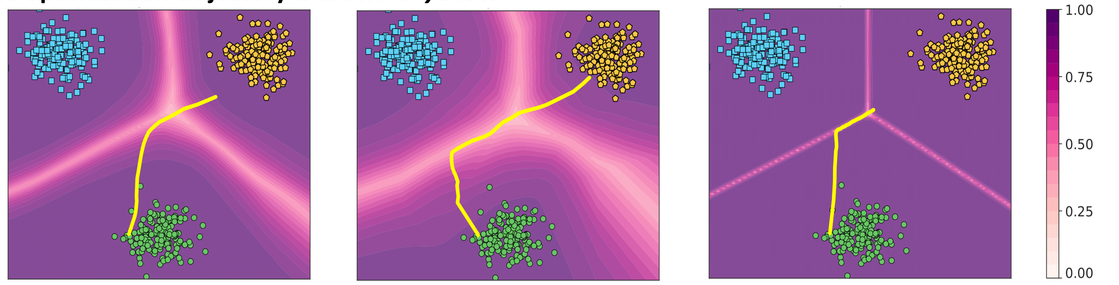
Label smoothing – using softened labels instead of hard ones – shows diverse benefits, such as enhanced generalization and calibration. However, its implications for preserving model privacy have remained unexplored.

Traditional label smoothing with positive smoothing factors fosters model inversion attacks and increases a model's privacy leakage.

Negative label smoothing counteracts this trend and emerges as a practical and viable defense mechanism.

Motivational Example

Model inversion attack on a toy dataset. The optimization starts from a sample from the *green circle* class and tries to reconstruct a sample from the *orange pentagons* class. The optimization trajectory is drawn in *yellow*.



Hard Labels $(\alpha = 0.0)$ **Positive LS** $(\alpha = 0.05)$ **Negative LS** $(\alpha = -0.05)$

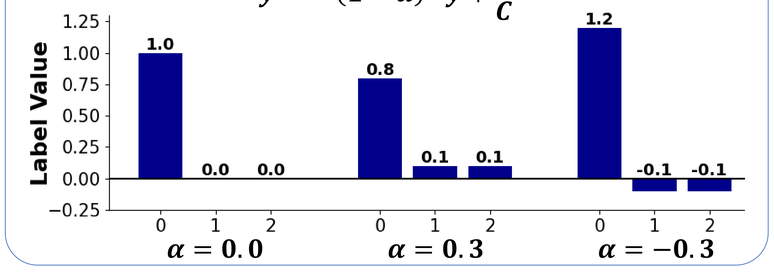
Code & Paper

Github.com/LukasStruppek/Plug-and-Play-Attacks

Label Smoothing

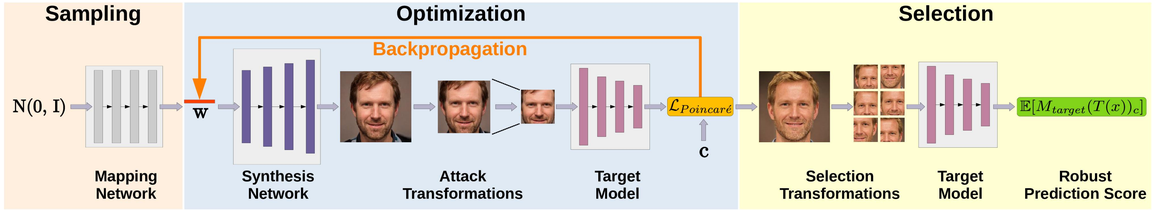
Label Smoothing replaces the hard-coded label y with a mixture of the hard label and a uniformly distributed vector.

$$y^{LS} = (1 - \alpha) \cdot y + \frac{\alpha}{C}$$



Model Inversion Attacks

Model inversion attacks aim to create synthetic images that reflect the class-wise characteristics from a target classifier's private training data by exploiting the model's learned knowledge.



[*Struppek et al. Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks, ICML '22*]

Contact

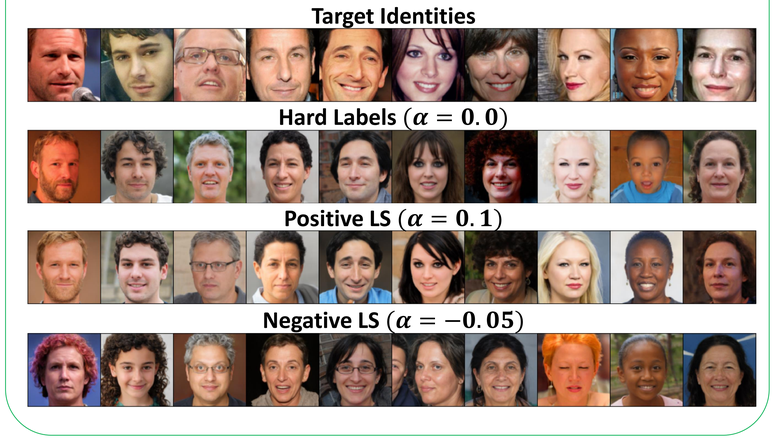
Please feel free to reach out to us!

Lukas Struppek
Technical University of Darmstadt
struppek@cs.tu-darmstadt.de
𝕏 @LukasStruppek

Dominik Hintersdorf
Technical University of Darmstadt
hintersdorf@cs.tu-darmstadt.de
𝕏 @D0miH

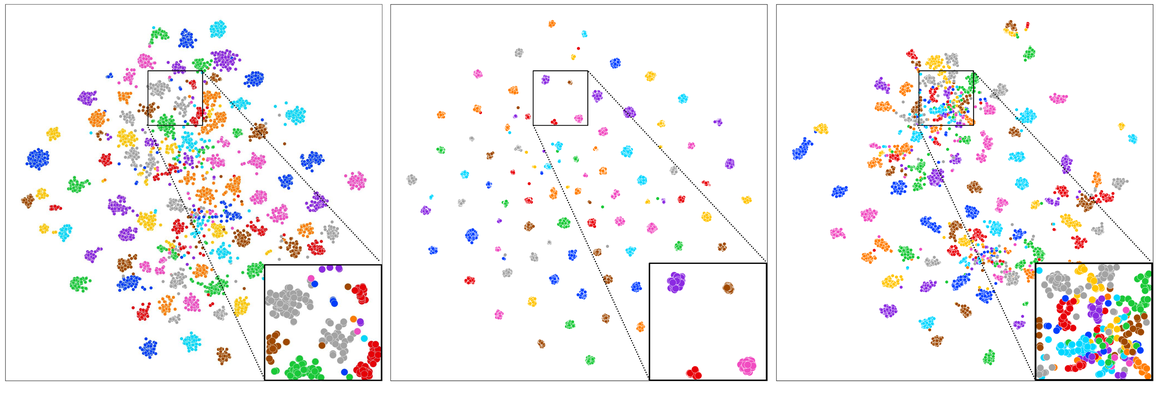
Qualitative Attack Results

The model trained with positive smoothing clearly reveals more visual characteristics of the identities, whereas attacks on the negative smoothing model generate misleading results.



Embedding Space Visualization

Positive label smoothing clusters samples from the same class together. Smoothing the labels with a negative factor reverses this effect and builds a less clearly separated space.



Hard Labels $(\alpha = 0.0)$ **Positive LS** $(\alpha = 0.1)$ **Negative LS** $(\alpha = -0.05)$

Gradient Stability

Mean cosine similarity between consecutive image gradients for the individual attack steps. The optimization path on the negative smoothing model is characterized by many changes of direction.