Protecting Against Contrastive Learning Poisoning with SAS

Mohammad Akbarnezhad, Adithya Embar, Dylan Gunn, and Anthony Holmes

Agenda

- Background
- Related Work
- Problem Formulation
- Methods and Challenges
- Experiments
- Work to be Continued

Background

- Poisoned data, if it is caught up in training data, is disastrous for visual models
- Real world implications
    - Tesla lane markings
    - Crowdsourced malware detection classifiers
    - Google's image recognition
- Ease of permutation
- Difficulty of identification

What is Contrastive Learning?

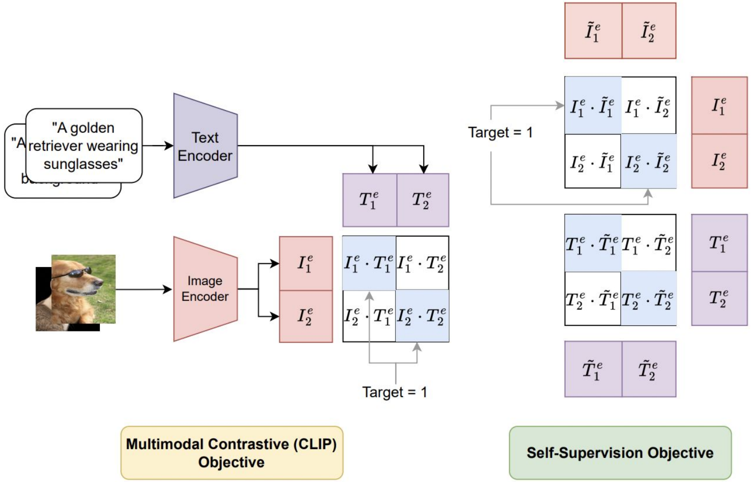
- Data Augmentation
- Representation Learning
- Loss Function

SAS and Protection from Poisoning

- Identifying Representative Examples
  - Low likelihood that central points are noise or poisoned
  - Poisoned examples usually atypical or extreme
- Submodular Optimization
- Training with Selected Subsets

Related Work

- CleanClip
  - Fine-tuning framework that weakens the effects of backdoor attacks
  - Weakening Spurious Associations
  - Independent Re-alignment
- Data Augmentation based on Matrix Completion (2-Steps)
  - The augmentation first randomly drops pixels in the image
  - Then it reconstructs the missing pixels via matrix completion



Defense Methods	AP-CL	EMP-CL-S	EMP-CL-C	Average
No Defense	80.2	44.9	68.9	64.7
Random Noise ($\sigma = 8/255$)	83.2	54.1	**90.3**	75.9
Random Noise ($\sigma = 64/255$)	72.2	73.6	73.6	73.1
Gauss Smooth ($k = 3$)	83.6	47.8	87.9	73.1
Gauss Smooth ($k = 15$)	63.0	59.7	62.0	61.6
Cutout	82.5	47.7	75.0	68.4
Adversarial Training	78.5	79.3	82.3	80.0
Matrix Completion	**83.6**	**85.6**	88.2	**85.8**
Clean Data		91.8		

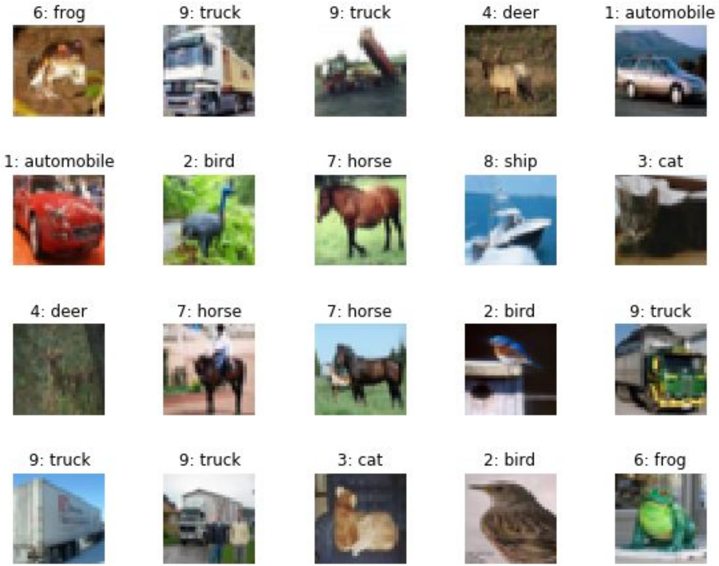
Problem Formulation

- Main Idea: Identify *how robust* different contrastive learning approaches are to poisoned data
- Select standardized clean and poisoned data
- Train and test each model for a certain number of epochs on each dataset
- Compare the approaches and their final test accuracies, identify what might make some approaches more robust than others

Methods and Challenges

- Dataset
  - CIFAR-10, original and poisoned
- Models
  - MoCo, MoCo v2, SimCLR, CMC

- Challenges
  - Compute
  - Model availability



Experiments

	CIFAR-10	CIFAR-10C
MoCo (25 epochs)	62.37%	40.13%
MoCo (50 epochs)	71.89%	48.26%

Work to be Completed

- Determine feasibility of continued work with CIFAR, pick other dataset? Shrink dataset? Get compute?
- Continue experiments on other models, get an idea of which models perform better against adversarial attacks

Thank You!

References

- https://arxiv.org/pdf/2202.11202v1.pdf
- https://openaccess.thecvf.com/content/ICCV2023/papers/Bansal_CleanCLIP_Mitigating_Data_Poisoning_Attacks_in_Multimodal_Contrastive_Learning_ICCV_2023_paper.pdf