CS772: Research Project

Zero Shot Unlearning

Ashutosh Kumar - 210221
Krish Sharma - 210530
Labajyoti Das - 210552
Shubham Patel - 210709
Siddharth Kalra - 211032

April 24, 2024

Problem Statement

- Survey Paper
- Machine Unlearning
  - Model M, Data D
  - Request:
    - Forget Data $D_f \subset D$
    - Retain Data $D_r = D - D_f$

Problem Statement

- Survey Paper
- Machine Unlearning
  - Model M, Data D
  - Request:
    - Forget Data $D_f \subset D$
    - Retain Data $D_r = D - D_f$
  - Gold / Retrained Model: M^*
  - Unlearned Model: M_u
  - Aim: $M_u(x) \approx M^*(x)$

Problem Statement

- Survey Paper
- Machine Unlearning
  - Model M, Data D
  - Request:
    - Forget Data $D_f \subset D$
    - Retain Data $D_r = D - D_f$
  - Gold / Retrained Model: M^*
  - Unlearned Model: M_u
  - Aim: $M_u(x) \approx M^*(x)$
- Zero-Shot Machine Unlearning
  - No Access to D_f and D_r

Seed Paper

- Zero-Shot Machine Unlearning
- Introduces the novel problem
- Introduces a new metric - Anamnesis Index

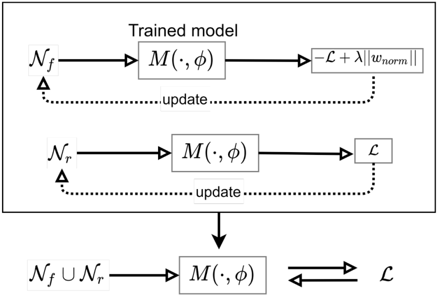
Seed Paper

- Zero-Shot Machine Unlearning
- Introduces the novel problem
- Introduces a new metric - Anamnesis Index
- Proposes two approaches - restricted setting of classification
- Setting
  - Set of Forget C_f and Retain Classes C_r

Seed Paper

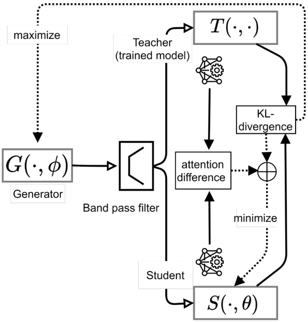
- Zero-Shot Machine Unlearning
- Introduces the novel problem
- Introduces a new metric - Anamnesis Index
- Proposes two approaches - restricted setting of classification
- Setting
  - Set of Forget C_f and Retain Classes C_r
- Approach
  1. Error Minimization-Maximization Noise
  2. Gated Knowledge Transfer

Error Minimization-Maximization Noise



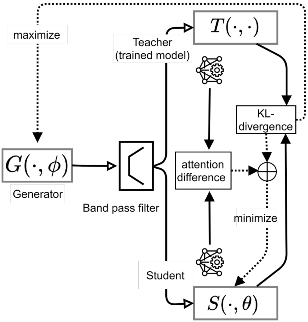
- Inspiration: Fast Yet Effective Machine Unlearning
- Anti-Samples N_f learnt by maximising loss
- Data representatives N_r learnt by minimising loss
- Updates the original model using noise

Gated Knowledge Transfer



- Inspiration: Zero-shot KT via Adversarial Belief Matching
- Knowledge Distillation to train the student from teacher

Gated Knowledge Transfer



- Inspiration: Zero-shot KT via Adversarial Belief Matching
- Knowledge Distillation to train the student from teacher
- Generator: Max $D_{KL}(T(x_g)||S(x_g)) = \sum_i^{|C|} t_p^{(i)} \log(t_p^{(i)}/s_p^{(i)})$
- Filter images belonging to C_f
- Student - Immitate Teacher's predictions
- Attention - Mimic Inner Layers - Minimise KL

Improvements and Extensions

- Ideas
  - Extending to Regression
  - Positive Samples instead of Anti-Samples
  - Entropy criteria on the filter
  - Better Alternatives for generators

Improvements and Extensions

- Ideas
  - Extending to Regression
  - Positive Samples instead of Anti-Samples
  - Entropy criteria on the filter
  - Better Alternatives for generators
- Entropy of predictions
  - Reject if $S(t_p) > \epsilon$
  - Poorer Retain Accuracy
  - Faster Retain Accuracy Restoration
  - Carrying out experimentations

Deep Inversion

- Difference in M^* and M_u
  - Non-zero probability for C_f
  - Due to Attention implicitly learn for C_f
  - Removing Attention: Impacts Performance
  - Reason: Poor Generated Images

Deep Inversion

- Difference in M^* and M_u
  - Non-zero probability for C_f
  - Due to Attention implicitly learn for C_f
  - Removing Attention: Impacts Performance
  - Reason: Poor Generated Images
- Inspiration: Deep Inversion
  - Idea: To minimise loss on x along with regularisation and matching lower-layer features
  - To determine when to halt in GKT
  - Much Better Images

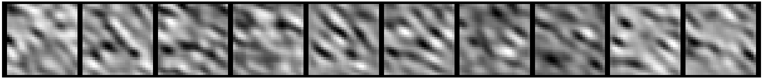
Generated Images

MNIST Numbers Dataset
Images of Digits from 0-9

Generated Images

MNIST Numbers Dataset
Images of Digits from 0-9

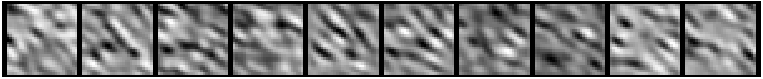
- GKT*



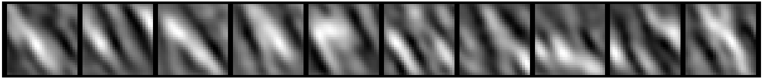
Generated Images

MNIST Numbers Dataset
Images of Digits from 0-9

- GKT*



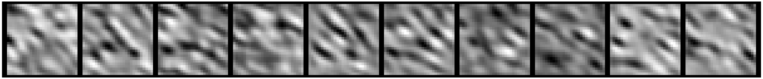
- GKT (with entropy criterion)*



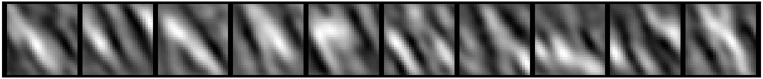
Generated Images

MNIST Numbers Dataset
Images of Digits from 0-9

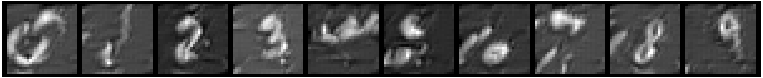
- GKT*



- GKT (with entropy criterion)*



- Deep Inversion



* Images not in order from 0-9. Images generated by the generator before forget accuracy begin to rise

Experimental Results

MNIST Numbers Dataset - `AllCNN` Model
Train: 60,000, Test: 10,000
Retain Accuracy on Test Set:

- Retrain Model: 99.25 %
- GKT: 97.12 %
- M-M: 10.57 %

Experimental Results

MNIST Numbers Dataset - `AllCNN` Model
Train: 60,000, Test: 10,000
Retain Accuracy on Test Set:

- Retrain Model: 99.25 %
- GKT: 97.12 %
- M-M: 10.57 %
- GKT (no attention): <50 %
- Deep Inversion (100 sample/class): 40 - 50 %
- Deep Inversion (6000 sample/class): 80 - 85 %

Experimental Results

MNIST Numbers Dataset - `AllCNN` Model
Train: 60,000, Test: 10,000
Retain Accuracy on Test Set:

- Retrain Model: 99.25 %
- GKT: 97.12 %
- M-M: 10.57 %
- GKT (no attention): <50 %
- Deep Inversion (100 sample/class): 40 - 50 %
- Deep Inversion (6000 sample/class): 80 - 85 %

If time permits will try to outperform the base paper on some dataset.

Conclusion

- Understand the internals of tackling zero-shot setting
- Determine the source of non-zero forget class accuracy and address it up to certain extent
- Improve upon the quality of images generated
- Decent Results

Learnings

- One of the first research experience
- Ability to read papers
- Exposure to tweaking complex machine learning code
- Repeatedly improving on strategies

Thank You