2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

A Study Regarding Machine Unlearning on Facial Attribute Data

Emircan Gündoğdu[1], Altay Unal[1] and Gozde Unal[2]
[1] Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey
[2] Department of AI and Data Engineering, Istanbul Technical University, Istanbul, Turkey

***Abstract*— Machine learning (ML) models require large amounts of data and many of the stored data is used to train ML models. However, the ML models learn insights about the data during their training and this raises privacy concerns of the individuals regarding personal data. These concerns led to the introduction of legislation focusing on the ”right to be forgotten” and machine unlearning has emerged to address these concerns. Although machine unlearning studies focus on data privacy issues generally, machine unlearning is also used to fix the mistrained machine learning models as well. Mistraining may occur due to problems in the data such as mislabeling. Machine unlearning can solve this problem by discarding the information regarding the problematic data. In this study, the effects of machine unlearning on facial attribute classification are discovered. Experimental results on CelebA dataset show the effectiveness of machine unlearning methods. The code repository can be accessed at https://github.com/ituvisionlab/face-attribute-unlearning .**

I. INTRODUCTION

Despite their celebrated performances on classification and regression tasks, machine learning (ML) models are mainly data-hungry models that require large amounts of data to achieve their performances [23]. Some of those tasks require personal data such as medical records, financial records, and personal information. Training an ML model requires consent from the data owners since processing this data would allow a model to obtain insights about the people who are subject to the training data [1].

The people may not give their consent for their data to be processed or they may forego their consent. The legislations such as General Data Protection Regulation (GDPR) [27] and California Consumer Privacy Act (CCPA) [10] include such provisions that the people have the *right to be forgotten*. For such cases, the trained models cannot use that data and it must be deleted. However, the deletion of the data is not enough since many ML models may already have been trained by processing it. Since that data cannot be used by the ML models, the ML models must forget it in a way. However, this requires another training session without the deleted data, which is computationally expensive. Training a model from scratch is unfeasible and the deletion process may be continuous since more users decide not to provide consent for their data to be processed at various times. Machine unlearning is proposed so that this process can be avoided and the users can preserve their privacy. Machine unlearning methods have been proven to be efficient in nullifying the effect of personal data in several tasks such as face recognition [26]. In addition to the privacy concerns, machine unlearning can also be used to improve the learning of the ML models. The mistakes in the data such as misclassification may cause an ML model not to perform in an intended way. Discarding the mistaken data and finetuning the model with the fixed data will improve the performance of the model, enabling it to perform as intended [17]. However, this process may also be computationally expensive [12], and it does not guarantee that the model actually ”forgot” the mistaken data. So, the faulty data may still affect the ML model.

Fig. 1: Example images from CelebA dataset.

Machine unlearning can provide a solution for such issues as it aims to obtain an ML model that performs as if it had never observed the discarded data [17]. The requests from the people who want to be forgotten can be continuous or the mislabelling may occur more than once. Since it is not feasible or sustainable to train ML models from the start by discarding a certain amount of data, machine unlearning becomes an option for such cases. Handling data issues has been critical for ML models and machine unlearning allows handling these issues with a lower computational cost. Considering that many of the state-of-the-art models are computationally expensive, the importance of machine unlearning is increasing [29].

Machine unlearning has started to be applied in different areas such as face recognition and object detection [1], [26], [5]. In this study, we used this method on facial attribute classification, becoming the first study to our knowledge. In addition, our study becomes one of the first examples of machine unlearning applications on multi-label data.

II. RELATED WORK

A. Machine Unlearning

Machine unlearning is a paradigm that aims to make an ML model forget about a particular data. There are different challenges concerning this paradigm such as catastrophic unlearning [21] and the nature of the training process. Since machine unlearning aims for a model to forget a particular data, it will decrease the performance of the model as the model is not allowed to use that particular data anymore. However, this may cause a performance loss on the model exponentially. In addition to catastrophic unlearning, the nature of the training process [1] is also a factor in

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

unlearning. The training process for a model is inherently stochastic, characterized by unpredictability in the influence of individual data samples. The data batches and their order are provided randomly to the model, therefore the process becomes stochastic. In addition to the stochasticity of the training, this process is also incremental. A model's performance on a data sample is affected by the prior samples. Therefore, the model update on a given data sample affects the model's performance on future samples. The current research on machine unlearning can be categorized into two main approaches: exact and approximate methods [22], each addressing the challenges posed by catastrophic unlearning and stochasticity, thereby contributing to more effective unlearning techniques.

Exact unlearning aims to obtain an ML model such that the effect of the removed data on model weights is completely removed. To be more precise, the weight distributions of the unlearned model and the retrained model must be equal. Several studies are achieving exact data deletion with different conditions. SISA [1] proposes dividing the data and supporting partial retraining by creating shard structures. However, exact unlearning is still a computationally expensive operation despite its computation cost being lower than the cost of retraining an entire model [31].

On the other hand, approximate unlearning defines boundaries regarding the unlearning to deal with the constraint issues caused by exact unlearning so that the ML model can perform as closely as possible to the retrained model. The constraints are more relaxed and several different strategies are developed such as gradient-based [26], [9] and influence-based methods [32]. These strategies decrease the computational cost as the ML models discard the information regarding the "forgotten" data. Due to their constraints being more relaxed than exact unlearning, approximate unlearning is a more applicable method among unlearning methods.

B. Multi-label Classification

Multi-label classification [24] is a supervised problem where a sample may be associated with more than one label. The single-label learning methods may ignore the correlations among labels. However, the correlation among labels can be discovered through multi-label classification.

Multi-label classification is mainly used on text classification [30] since text data can be classified in more than one class. However, multi-label classification is also applied in different areas such as map labeling, bioinformatics [6], scene classification [2], and object detection [11]. Multi-label classification can also be applied to facial image data as well. There are multiple characteristics of human faces and recognition of these characteristics altogether is possible with multi-label classification.

C. Facial Attribute Classification

There are many studies associated with facial images such as face recognition [33], [13] and facial expression recognition [18], [7]. Facial attribute classification [34] is another application area where the face attributes are recognized from face images such as mustache and glasses. Facial attribute classification has different application areas such as face verification [16], and image retrieval [3].

Mislabelled data may cause problems in facial attribute classification since this approach is fused with different areas. For instance, an image retrieval model can be constructed according to the facial attributes. However, in case of mislabelled data, the image retrieval model cannot perform as planned. To handle mislabelled data samples, machine unlearning, and its effects are discovered in this study.

III. METHOD

A. Problem Definition

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a dataset consisting of N samples, where x_i is the i^{th} sample with label $y_i \subseteq \{1, 2, \ldots, K\}$, and K is the number of classes. Each x_i may have multiple labels among $\{1, 2, \ldots, K\}$. $\mathcal{D}_f$ and $\mathcal{D}_r$ denote data corresponding to the forgotten classes C_f and retained classes C_r, respectively. In our problem, $\mathcal{D}_f$ is defined as the set of data points (x_i, y_i) where $y_i[c_f] = 1$ for all $c_f \in C_f$, and $\mathcal{D}_r$ as $\mathcal{D} \backslash \mathcal{D}_f$. The objective is to discard the information regarding D_f so that an ML model can be obtained from an ML model that is previously trained on $\mathcal{D}$.

B. Implementation Details

1) Model Architectures: During the experiments, ResNet-18 and ResNet-50 are used as the main architectures for the unlearning methods. The last layer was adjusted according to the number of classes in the used dataset. In addition to ResNet-18 and ResNet-50, a pre-trained Vision Transformer (ViT) is also used. During the experiments, ResNets are fed with 64x64 sized image data while ViT is fed with 224x224 since ViT is not compatible with smaller sizes due to patch-based processing.

2) Approaches: The used unlearning approaches in this study are introduced as follows:

- Original: The original model trained on the complete dataset, including both D_r and D_f.
- Retrain: The model retrained on the retained data D_r.
- Finetune: The model finetuned on the retained data D_r.
- NegGrad [9]: The model fine tuned on D_f by moving in the direction of the increasing loss.
- NegGrad+ [4]: The model finetuned on both D_r and D_f by moving in the direction of the joint loss with increasing loss for D_f and decreasing loss for D_r.
- UNSIR [26]: A machine unlearning approach that introduces noise into the model so that the information regarding D_f can be discarded.
- Bad Teacher [5]: A machine unlearning approach using a teacher-student network with two teacher networks. The teacher network regarding D_r is competent while the teacher network regarding D_f is incompetent so that the student network can be manipulated to forget D_f.

TABLE I: Results of single class unlearning ($|C_f| = 1$) on retain test set ($\mathcal{D}_r^{test}$) and forget test set ($\mathcal{D}_f^{test}$) of CelebA dataset. JSD score is calculated between Retrain model and corresponding models.

Method	Retain Test Set				Forget Test Set			
	ResNet-18		ResNet-50		ResNet-18		ResNet-50	
	Hamming (↑)	JSD (↓)	Hamming (↑)	JSD (↓)	Hamming	JSD (↓)	Hamming	JSD (↓)
Original	0.99±0.00	-	0.98±0.00	-	0.99±0.00	-	0.98±0.00	-
Retrain	0.96±0.00	-	0.96±0.00	-	0.84±0.00	-	0.98±0.00	-
Finetune	0.98±0.00	0.07±0.00	0.98±0.00	0.06±0.00	0.98±0.00	0.01±0.00	0.84±0.00	0.08±0.00
NegGrad	0.76±0.00	0.46±0.00	0.72±0.00	0.43±0.00	0.69±0.00	0.64±0.00	0.68±0.00	0.58±0.00
NegGrad+	0.84±0.00	0.23±0.00	0.84±0.00	0.23±0.00	0.76±0.00	0.35±0.00	0.76±0.00	0.32±0.00
BadT	0.91±0.00	0.11±0.00	0.83±0.00	0.10±0.00	0.87±0.00	0.10±0.00	0.79±0.00	0.09±0.00
UNSIR	0.88±0.00	0.08±0.00	0.87±0.00	0.06±0.00	0.83±0.00	0.07±0.00	0.83±0.00	0.06±0.00

C. Metrics

During the experiments, the Hamming score [8] and Jensen-Shannon Divergence (JSD) [19] are used to evaluate the performance of the approaches. The Hamming score is a generally used metric for multi-label classification. It shows the ability of the model to determine multiple classes successfully. The Hamming score formula is given in Equation 1.

$$HammingScore = 1 - \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} \left[I\left(y_i^{(j)} \neq \hat{y}_i^{(j)} \right) \right]. \tag{1}$$

where n denotes the number of samples, L denotes the number of classes, $y_i^{(j)}$ represents the ground truth label of i^{th} sample's j^{th} class, and $\hat{y}_i^{(j)}$ represent the corresponding prediction. While the Hamming score is used to evaluate the model's accuracy in finding multiple classes, the unlearning performance is evaluated using JSD. JSD is used to determine the probability difference between the unlearned model and the retrained model [28]. JSD is given in Equation 2.

$$JSD(p(x), q(x)) = 0.5(KL(p(x)||q(x)) + KL(q(x)||p(x))) \tag{2}$$

where $p(x)$ and $q(x)$ represent the weight distribution of the compared two models. JSD is calculated by using the KL-Divergence [15] between the unlearned model and the retrained model on D_r. Since the aim of unlearning is to become as close as possible to the retrained model, the JSD should ideally approach 0.

IV. EXPERIMENTS

In this section, the details of the experiments and the results are provided.

A. Experimental Settings

Our experiments were performed in one NVIDIA Titan RTX GPU. Adam optimizer [14] is mainly used for training and finetuning with a learning rate of 0.0001. However, the optimizer and learning rate varied for different unlearning approaches. Specifically, for the NegGrad and NegGrad+ methods, SGD [25] with a learning rate of 0.001 is applied. UNSIR algorithm used Adam optimizer with learning rates set at 0.1 for the noise learning, 0.02 for the impair step, and 0.01 for the repair step. Lastly, the Bad Teacher approach used the Adam optimizer with a learning rate of 0.0001.

B. Dataset

The CelebA dataset [20] is used for the experiments. This dataset consists of 40 facial attributes (e.g., eyeglasses, mustache, and smiling). Since it has been an established dataset for studies focusing on facial attributes and has a large variety of facial attributes, this dataset is chosen. Some examples from the dataset can be observed in Figure 1.

C. Results

The results for the one class removal can be observed in Table I. The results show that the unlearning process is successful in facial attribute classification. Its accuracy is close to retraining an ML model with retained data, showing that the unlearning methods can perform in a similar way with a lower computational cost. In addition, the JSD score shows that the difference between the distribution of the model weights is low. Machine unlearning aims to obtain an ML model as close as possible to the retrained model without retraining it, therefore the unlearning methods can perform in the intended way. The results for the multiple class removal can be observed in Tables II and III. Two and three classes are removed for the experiments. It is observed once again that the unlearning model can manage the removal of multiple classes without a huge performance loss. Catastrophic unlearning is an important issue for machine unlearning since it may disrupt the previously learned information of the model. In addition, since some of the classes are correlated in facial attribute data, the damage caused by catastrophic unlearning could be larger. However, the results show that the unlearning methods can handle the multiple class removal. Figure 2 shows the changes in both the Hamming score and the JSD for the unlearning approaches when the forgotten class number is increased. It is observed that unlearning models can maintain their performances and discard the information regarding D_f. Their weight distributions tend to be close to the retrained model which is the main aim of unlearning.

It can also be observed that the Hamming scores on D_f are high. The reason for that is although D_f is discarded

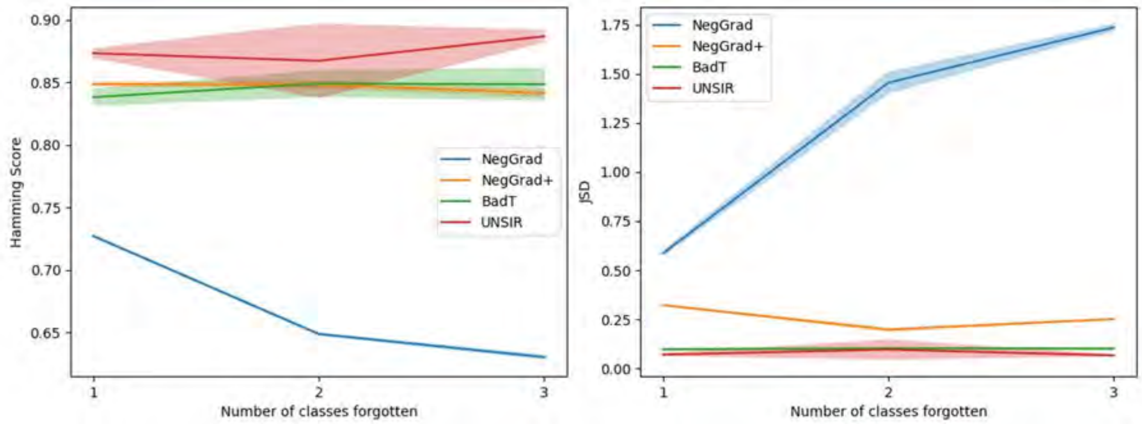

Fig. 2: The plot for the Hamming score on retain test set D_r^{test} and JSD on forget test set D_f^{test} with different numbers of class removal ($|C_f| = \{1, 2, 3\}$).

TABLE II: Results of 2-classes unlearning ($|C_f| = 2$) with ResNet-50 architecture.

Method	Retain Test Set		Forget Test Set	
	Hamming (↑)	JSD (↓)	Hamming	JSD (↓)
Original	0.98±0.00	-	0.98±0.00	-
Retrain	0.96±0.00	-	0.84±0.00	-
Finetune	0.98±0.00	0.06±0.00	0.98±0.00	0.08±0.00
NegGrad	0.64±0.00	1.39±0.05	0.60±0.00	1.45±0.05
NegGrad+	0.84±0.00	0.18±0.01	0.78±0.00	0.19±0.01
BadT	0.84±0.00	0.09±0.00	0.80±0.00	0.10±0.00
UNSIR	0.86±0.03	0.09±0.00	0.81±0.03	0.09±0.05

TABLE III: Results of 3-classes unlearning ($|C_f| = 3$) with ResNet-50 architecture.

Method	Retain Test Set		Forget Test Set	
	Hamming (↑)	JSD (↓)	Hamming	JSD (↓)
Original	0.98±0.00	-	0.98±0.00	-
Retrain	0.97±0.00	-	0.84±0.00	-
Finetune	0.98±0.00	0.06±0.00	0.98±0.00	0.09±0.00
NegGrad	0.63±0.00	1.66±0.02	0.58±0.00	1.73±0.02
NegGrad+	0.84±0.00	0.22±0.00	0.77±0.00	0.25±0.00
BadT	0.84±0.01	0.09±0.00	0.79±0.01	0.10±0.00
UNSIR	0.88±0.00	0.06±0.00	0.83±0.00	0.06±0.00

TABLE IV: Results of single class unlearning ($|C_f| = 1$) with ResNet-50 and ViT architectures.

Arch.	Method	Retain Test Set		Forget Test Set	
		Hamming (↑)	JSD (↓)	Hamming	JSD (↓)
ResNet-50	Original	0.98±0.00	-	0.98±0.00	-
	Retrain	0.96±0.00	-	0.98±0.00	-
	UNSIR	0.87±0.00	0.06±0.00	0.83±0.00	0.06±0.00
ViT	Original	0.96±0.00	-	0.96±0.00	-
	Retrain	0.95±0.00	-	0.87±0.00	-
	UNSIR	0.80±0.00	0.10±0.00	0.76±0.00	0.13±0.00

and its class information is forgotten, the class information from D_r remains, since D_f has multiple labels, including C_f. Therefore, a dramatic drop in Hamming scores is not observed. However, when the unlearning approaches are compared to the retrained model by the JSD, the divergence between the weight distributions is close to 0, meaning that the class information regarding C_f is discarded properly. In addition, the Hamming scores for the unlearning approaches are closer to the retrained model and this also shows the information regarding D_f is discarded properly. The results from Tables I, II and III also show that the finetuned model is effective in terms of the Hamming score and the JSD for D_r. However, machine unlearning does not only focus on the model performance on the retained data D_r. The performance regarding the forgotten data D_f must also be considered to understand whether the information of the forgotten data is discarded or not. The high Hamming score on D_f shows that the finetuned model may not have discarded the information on D_f, therefore it did not perform unlearning.

The results with ViT are provided in Table IV. The performance of the machine unlearning method UNSIR with ViT is in line with its performance with ResNet-50. However, the Hamming score for ViT is lower than its score for ResNet-50. This may occur due to the resize operation on the dataset. Since CelebA contains images with 64x64 pixels, the images are resized for ViT as 224x224. The dimensions of learned features for ViT also kept the same as for ResNets, which is small for a default ViT. Therefore, the unlearning operation may also have been affected.

V. CONCLUSION

In this study, the effects of machine unlearning are discovered on facial attribute data. Several unlearning approaches are tested on the class unlearning setup. The performances of the unlearning approaches on facial attribute data are provided using different model architectures. It is shown that machine unlearning methods can perform on datasets with multiple labels.

REFERENCES

[1] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

[2] M. Boutell, X. Shen, J. Luo, and C. Brown. Multi-label semantic scene classfication. 2003.

[3] B.-C. Chen, Y.-Y. Chen, Y.-H. Kuo, and W. H. Hsu. Scalable face image retrieval using attribute-enhanced sparse codewords. *IEEE Transactions on Multimedia*, 15(5):1163–1173, 2013.

[4] D. Choi and D. Na. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*, 2023.

[5] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7210–7217, 2023.

[6] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. *Advances in neural information processing systems*, 14, 2001.

[7] A. Fathallah, L. Abdi, and A. Douik. Facial expression recognition via deep learning. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 745–750. IEEE, 2017.

[8] D. Ganda and R. Buch. A survey on multi label classification. *Recent Trends in Programming Languages*, 5(1):19–23, 2018.

[9] A. Golatkar, A. Achille, and S. Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.

[10] E. Goldman. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*, 2020.

[11] T. Gong, B. Liu, Q. Chu, and N. Yu. Using multi-label classification to improve object detection. *Neurocomputing*, 370:174–185, 2019.

[12] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4805–4814, 2019.

[13] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 142–150, 2015.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[16] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pages 365–372. IEEE, 2009.

[17] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou. Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[19] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[20] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[21] Q. P. Nguyen, B. K. H. Low, and P. Jaillet. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020.

[22] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

[23] Z. Obermeyer and E. J. Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[24] J. Read and F. Perez-Cruz. Deep learning for multi-label classification. *arXiv preprint arXiv:1502.05988*, 2014.

[25] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[26] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[27] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[28] L. Wang, T. Chen, W. Yuan, X. Zeng, K.-F. Wong, and H. Yin. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*, 2023.

[29] J. Xu, Z. Wu, C. Wang, and X. Jia. Machine unlearning: Solutions and challenges. *arXiv preprint arXiv:2308.07061*, 2023.

[30] K. Yu, S. Yu, and V. Tresp. Multi-label informed latent semantic indexing. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 258–265, 2005.

[31] S. Yu, F. Sun, J. Guo, R. Zhang, and X. Cheng. Legonet: A fast and exact unlearning architecture. *arXiv preprint arXiv:2210.16023*, 2022.

[32] Y. Zhang, Z. Hu, Y. Bai, F. Feng, J. Wu, Q. Wang, and X. He. Recommendation unlearning via influence function. *arXiv preprint arXiv:2307.02147*, 2023.

[33] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

[34] N. Zhuang, Y. Yan, S. Chen, and H. Wang. Multi-task learning of cascaded cnn for facial attribute classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2069–2074. IEEE, 2018.