上海科技大学
ShanghaiTech University

MultiDelete for Multimodal Machine Unlearning

Jiali Cheng Hadi Amiri
University of Massachusetts Lowell

SETTINGS

提出了第一种针对多模态模型的unlearning 方法

$$D_{\text{train}} = \{(I_i, T_i)\}_{i=1}^{N}$$

D_f

$forget$

$$D_r = D_{train} \setminus D_f$$

vision-language model f

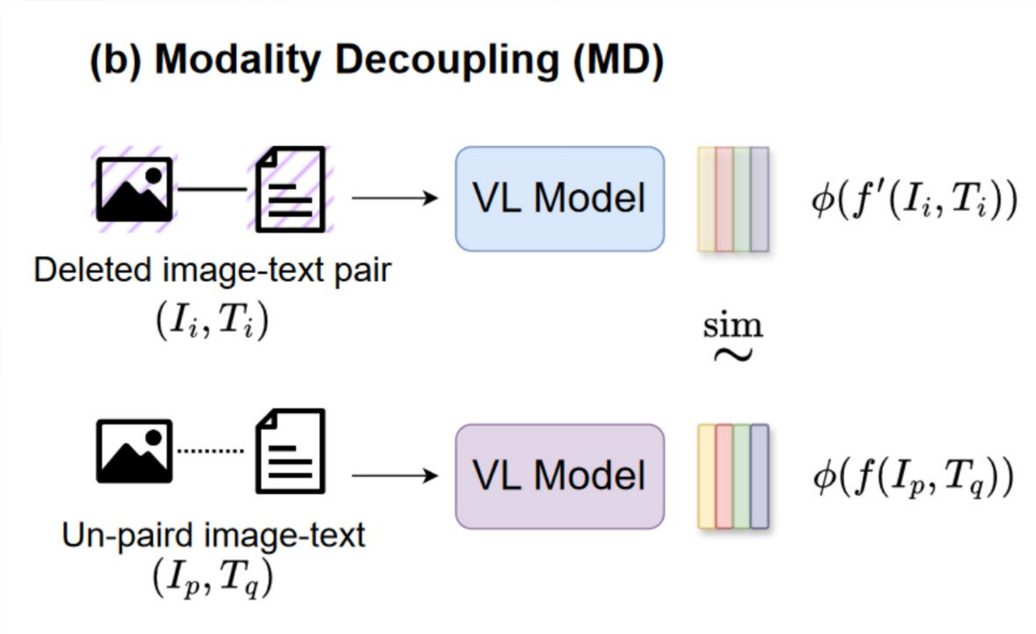
vision feature extractor f_I

language feature extractor f_T

a modality fusion module f_F

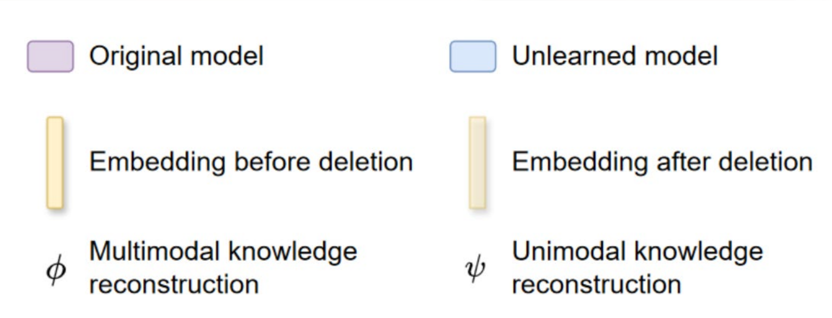
This approach ensures that the model retains its foundational knowledge of individual modalities, which is essential for effective learning of the target task and prevents the unnecessary loss of information.

Modality Decoupling



TODO: phi 和 f 还是不太明确

$$\mathop{\mathbb{E}}_{(I_i,T_i)\in D_f,(I_p,T_q)_{p\neq q}}\left[\phi(f'(I_i,T_i))-\phi(f(I_p,T_q))\right]=\epsilon,$$



- where f (·) and f ′(·) generate multimodal representations of their inputs,
- ϕ is a readout function (such as the concatenation operator, applied to a set of representations),
- ϵ is an infinitesimal constant.

Loss No.1

上海科技大学
ShanghaiTech University

$$\mathcal{L}_{\mathrm{MD}} = \mathrm{Dis}\Big(\{f'(I_i, T_i) | (I_i, T_i) \in D_f\},$$
$$\{f(I_p, T_q) | (I_p, T_p) \in D_r, (I_q, T_q) \in D_r, p \neq q\}\Big),$$

Dis(·) can be mean squared error.

Multimodal Knowledge Retention

$$\mathop{\mathbb{E}}_{(I_r,T_r)\in D_r}\Big[\phi\big(f'(I_r,T_r)\big)-\phi\big(f(I_r,T_r)\big)\Big]=\epsilon,$$

$$\mathcal{L}_{\mathrm{MKR}}=\mathrm{Dis}\Big(f'(I_r,T_r),f(I_r,T_r)\Big),(I_r,T_r)\in D_r.$$

Unimodal Knowledge Retention

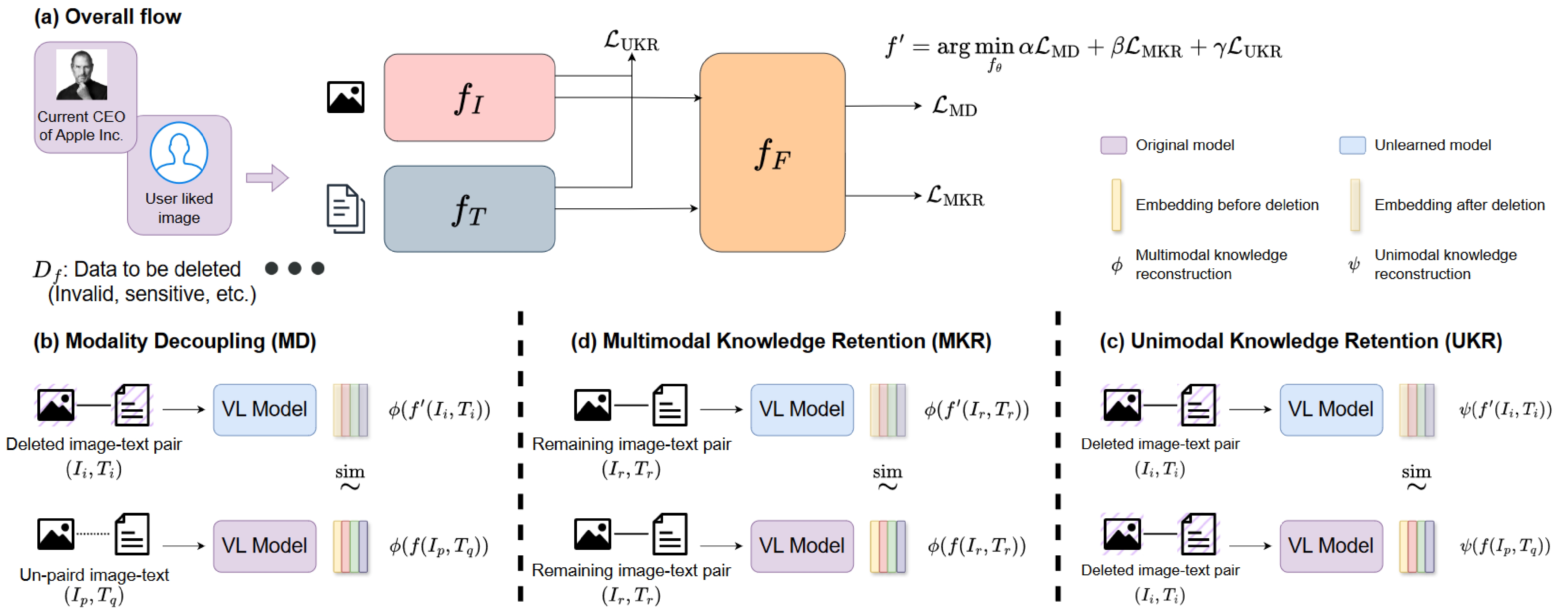
上海科技大学 ShanghaiTech University

$$\mathop{\mathbb{E}}_{(I_i,T_i)\in D_f}\left[\psi\big(f'_I(I_i),f'_T(T_i)\big)-\psi\big(f_I(I_i),f_T(T_i)\big)\right]=\epsilon,$$

$$\mathcal{L}_{\mathrm{UKR}}=\mathrm{Dis}\Big(\big\{[f'_I(I_i);f'_T(T_i)]\,|\,(I_i,T_i)\in D_f\big\},\big\{[f_I(I_i);f_T(T_i)]\,|\,(I_i,T_i)\in D_f\big\}\Big),$$

上一的ψ对应上二 [;]，即向量链接

$$\mathcal{L}=\alpha\mathcal{L}_{\mathrm{MD}}+\beta\mathcal{L}_{\mathrm{MKR}}+\gamma\mathcal{L}_{\mathrm{UKR}},$$


(a) Overall flow
Current CEO of Apple Inc.
User liked image
D_f: Data to be deleted (Invalid, sensitive, etc.)
f_I
f_T
f_F
$\mathcal{L}_{\mathrm{UKR}}$
$\mathcal{L}_{\mathrm{MD}}$
$\mathcal{L}_{\mathrm{MKR}}$
$f' = \arg\min_{f_\theta} \alpha\mathcal{L}_{\mathrm{MD}} + \beta\mathcal{L}_{\mathrm{MKR}} + \gamma\mathcal{L}_{\mathrm{UKR}}$
Original model
Unlearned model
Embedding before deletion
Embedding after deletion
ϕ Multimodal knowledge reconstruction
ψ Unimodal knowledge reconstruction
(b) Modality Decoupling (MD)
Deleted image-text pair
(I_i, T_i)
VL Model
$\phi(f'(I_i, T_i))$
sim
Un-paird image-text
(I_p, T_q)
VL Model
$\phi(f(I_p, T_q))$
(d) Multimodal Knowledge Retention (MKR)
Remaining image-text pair
(I_r, T_r)
VL Model
$\phi(f'(I_r, T_r))$
sim
Remaining image-text pair
(I_r, T_r)
VL Model
$\phi(f(I_r, T_r))$
(c) Unimodal Knowledge Retention (UKR)
Deleted image-text pair
(I_i, T_i)
VL Model
$\psi(f'(I_i, T_i))$
sim
Deleted image-text pair
(I_i, T_i)
VL Model
$\psi(f(I_p, T_q))$

Task and Dataset

上海科技大学
ShanghaiTech University

TASK

Image-Text Retrieval (TR) and (IR) are the tasks of retrieving the top-k relevant texts for a given image query (TR), and vice versa

Visual Entailment (VE): an image-text entailment task, To determine whether a given text hypothesis T_i entails, contradicts,or is neutral with respect to a given image premise I_i.

DATASET

Flickr30K: image+description(s)
"Two young guys with shaggy hair look at their hands while hanging out in the yard ."

SNLI-VE: Visual Entailment Dataset

+

- *Two woman are holding packages.*
- *The sisters are hugging goodbye while holding to go packages after just eating lunch.*
- *The men are fighting outside a deli.*

=

- *Entailment*
- *Neutral*
- *Contradiction*

Premise *Hypothesis* *Answer*

Task and Dataset

TASK

Natural Language for Visual Reasoning (NLVR) is the binary classification task of predicting whether a given text T_i accurately describes a given pair of images $(I_{i,1}, I_{i,2})$.

Graph-Text Classification: is the task of classifying whether a text indicates a specific (e.g. causal) relationship between two given entities in a subgraph.

DATASET

NLVR2

- Two penguins stand near each other in the picture on the left.
- There are only two penguins in at least one of the images.
- An image features two penguins standing close together.
- There are two penguins in the left image.
- An image contains just two penguins.

PGR: A Silver Standard Corpus of Human Phenotype-Gene Relations

Sentence: A homozygous mutation of **SERPINB6**, a gene encoding an intracellular protease inhibitor, has recently been associated with post-lingual, autosomalrecessive, nonsyndromic **hearing loss** in humans (DFNB91).

- Gene: **SERPINB6**
- Phenotype: **hearing loss**
- Relation: **Known**

Baselines

1. **Retrain**
2. Finetune: **Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks**
   fine-tunes f on D_f with a larger learning rate, similar to catastrophic forgetting
3. **NegGrad:** optimizes the original loss function of training f on Df but reverses the direction of gradients to unlearn these samples.
4. **Descent to Delete(DtD)** is a weight scrubbing-based and modality-agnostic approach to unlearning. It assumes that the weights of f' are close to the weights of f, trains f for a few more steps while adding Gaussian noise to scrub the weights.
5. **L-codec**(vision or text): is a weight scrubbing-based approach that approximates the Hessian matrix and performs a Newton update step to scrub the parameters while adding noise to them.
6. **Erm-Ktp**(vision only) : is a retraining-based approach that unlearns data by retraining the model with extra parameters inserted after visual feature maps to entangle correlations between classes.
7. **UL**(text only) : is an optimization-based approach that unlearns data by maximizing the log likelihood of samples in D_f . This method has been developed for machine unlearning in language models.

Settings

1. $|D_f| = 1K, 2K, \ldots, 5K$
2. $\alpha, \beta, \gamma = 1$
3. The original models f are trained until convergence before being used for deletion experiments.
4. For deletion, select the best checkpoint using validation set of each dataset.

vision-language transformers ALBEF

average MI ratio of 1.3

Method	Image-Text								Graph-Text		Avg.	
	Flickr30K				SNLI-VE		NLVR^2		PGR			
	IR		TR									
	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f
Retrain	97.8	50.4	93.5	50.4	79.4	50.2	80.3	50.3	67.5	50.2	83.4	50.3
FineTune	96.7	50.4	94.1	50.4	79.1	50.5	80.3	49.8	67.4	49.9	83.5	50.2
FineTune-F	**97.1**	49.9	**94.6**	49.9	79.5	49.9	81.2	50.0	67.5	50.1	<u>83.8</u>	49.9
NegGrad	92.4	50.5	91.7	50.5	77.8	48.6	77.3	50.6	63.4	49.6	80.5	50.0
NegGrad-F	93.3	50.2	90.6	50.2	<u>79.6</u>	50.6	**80.8**	50.0	63.5	49.9	81.5	50.2
DtD	10.3	51.4	8.9	51.4	45.2	50.1	50.8	49.8	50.0	50.2	33.0	50.5
DtD-F	22.5	50.9	20.7	50.9	48.6	49.8	50.9	49.8	53.6	50.2	39.2	50.2
L-codec	83.5	50.0	78.5	50.0	56.7	49.9	55.3	52.7	57.8	48.8	66.3	50.3
L-codec-F	87.4	49.4	50.6	48.2	57.4	48.4	56.8	53.1	59.1	46.9	62.2	49.2
Erm-Ktp	57.4	48.7	56.2	49.0	53.2	48.9	52.9	50.8	N/A		54.9	49.3
Erm-Ktp-F	N/A											
UL	95.1	50.4	90.3	50.4	75.7	49.8	76.3	50.4	64.8	49.7	80.4	50.2
UL-F	94.4	50.2	94.1	50.2	79.1	49.7	76.8	50.4	66.1	48.8	82.1	49.8
MultiDelete	**97.1**	**33.2**	<u>94.3</u>	**33.2**	**79.8**	**35.3**	**80.8**	**23.5**	**68.5**	**18.6**	**84.2**	**28.7**
MultiDelete-F	<u>96.8</u>	<u>34.4</u>	94.1	<u>34.5</u>	79.5	<u>36.3</u>	<u>80.4</u>	<u>26.4</u>	**67.7**	**19.5**	83.7	<u>30.2</u>

Main Results

- **Comparison to Modality-agnostic Approaches:**
  - The lower performance of these approaches show that they can’ t remove learned multimodal dependencies.
- **Comparison to Unimodal Approaches**:
  - Results show that unimodal unlearning approaches do not effectively translate to multimodal contexts.

Method	Image-Text								Graph-Text		Avg.	
	Flickr30K IR		Flickr30K TR		SNLI-VE		NLVR^2		PGR			
	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f
ERM-KTP	57.4	48.7	56.2	49.0	53.2	48.9	52.9	50.8	N/A		54.9	49.3
ERM-KTP-F	N/A											
UL	95.1	50.4	90.3	50.4	75.7	49.8	76.3	50.4	64.8	49.7	80.4	50.2
UL-F	94.4	50.2	94.1	50.2	79.1	49.7	76.8	50.4	66.1	48.8	82.1	49.8
MultiDelete	**97.1**	**33.2**	94.3	**33.2**	**79.8**	**35.3**	**80.8**	**23.5**	**68.5**	**18.6**	**84.2**	**28.7**
MultiDelete-F	96.8	34.4	94.1	34.5	79.5	36.3	80.4	26.4	**67.7**	**19.5**	83.7	30.2

Main Results

- **Comparison to Modality-agnostic Approaches:**
  - The lower performance of these approaches show that they can’t remove learned multimodal dependencies.
- **Comparison to Unimodal Approaches**:
  - Results show that unimodal unlearning approaches do not effectively translate to multimodal contexts.
  - Updating the knowledge on one of the modalities results in drop on both test set performance and model’s ability in forgetting D_f . Therefore, merely unlearning a single modality is inadequate for comprehensive unlearning in multimodal settings.

Main Results

- **Limitations of Scrubbing Methods and Retrain:**
  - The lower performance of these approaches show that they can’ t remove learned multimodal dependencies.
  - In case of multimodal settings, we argue that scrubbing or noise addition disrupts the original learned dependencies, particularly when model parameters are shared

Method	Image-Text								Graph-Text		Avg.	
	Flickr30K IR		Flickr30K TR		SNLI-VE		NLVR^2		PGR			
	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f
DtD	10.3	51.4	8.9	51.4	45.2	50.1	50.8	49.8	50.0	50.2	33.0	50.5
DtD-F	22.5	50.9	20.7	50.9	48.6	49.8	50.9	49.8	53.6	50.2	39.2	50.2
L-codec	83.5	50.0	78.5	50.0	56.7	49.9	55.3	52.7	57.8	48.8	66.3	50.3
L-codec-F	87.4	49.4	50.6	48.2	57.4	48.4	56.8	53.1	59.1	46.9	62.2	49.2
MultiDelete	**97.1**	**33.2**	94.3	**33.2**	**79.8**	**35.3**	**80.8**	**23.5**	**68.5**	**18.6**	**84.2**	**28.7**
MultiDelete-F	96.8	34.4	94.1	34.5	79.5	36.3	80.4	26.4	**67.7**	**19.5**	83.7	30.2

Main Results

- **Limitations of Scrubbing Methods and Retrain:**
  - For retrain, These results indicate that matching model parameters does not necessarily mean successful unlearning due to potential distribution discrepancy in model parameters

Method	Image-Text								Graph-Text		Avg.	
	Flickr30K IR		Flickr30K TR		SNLI-VE		NLVR^2		PGR			
	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f	D_{Test}	D_f
RETRAIN	97.8	50.4	93.5	50.4	79.4	50.2	80.3	50.3	67.5	50.2	83.4	50.3
MULTIDELETE	**97.1**	**33.2**	94.3	**33.2**	**79.8**	**35.3**	**80.8**	**23.5**	**68.5**	**18.6**	**84.2**	**28.7**
MULTIDELETE-F	96.8	34.4	94.1	34.5	79.5	36.3	80.4	26.4	**67.7**	**19.5**	83.7	30.2

Membership Inference Attack

Method	Flickr-TR	Flickr-IR	SNLI-VE	NLVR^2	Avg.
RETRAIN	1.10	1.10	1.05	1.07	1.08
FINETUNE	1.03	1.03	1.04	1.08	1.04
FINETUNE-F	1.06	1.06	1.07	1.09	1.07
NEGGRAD	1.11	1.11	1.09	1.06	1.09
NEGGRAD-F	1.14	1.14	1.10	1.08	1.11
DTD	1.41	1.41	1.60	1.71	1.53
DTD-F	1.40	1.40	1.58	1.66	1.51
L-CODEC	1.21	1.21	1.23	1.23	1.22
L-CODEC-F	1.22	1.22	1.26	1.26	1.24
ERM-KTP	1.10	1.10	1.11	1.21	1.13
ERM-KTP-F	N/A				
UL	0.97	0.97	1.04	1.07	1.01
UL-F	0.98	0.98	1.10	1.04	1.02
MULTIDELETE	1.27	1.27	1.30	1.25	1.27
MULTIDELETE-F	1.25	1.25	1.26	1.21	1.24

Membership Inference Attack

- MultiDelete outperforms non-scrubbing baselines (FineTune, NegGrad, ErmKtp, UL) by *0.19* absolute points in MI ratio.
- For scrubbing methods (DtD , L-codec) the drop applies to all data including both D_r and D_f . This shows that the unlearning of scrubbing methods is not targeted at a specific subset of data, but the entire data

Method	Image-Text				Graph-Text	Avg.
	Flickr-IR	**Flickr-TR**	**SNLI-VE**	**NLVR^2**	**PGR**	
RETRAIN	1.10	1.10	1.05	1.07	1.09	1.08
L-CODEC	1.21	1.21	1.23	1.23	1.07	1.19
L-CODEC-F	1.22	1.22	1.26	1.26	1.09	1.21
MULTIDELETE	**1.27**	**1.27**	**1.30**	**1.25**	**1.24**	**1.27**
MULTIDELETE-F	1.25	1.25	1.26	1.21	1.20	12.4

Ablation study

上海科技大学
ShanghaiTech University

	NLVR^2		PGR	
	D_{Test}	D_f	D_{Test}	D_f
RETRAIN	80.3	50.3	67.5	50.2
Full model	**80.8**	**23.5**	**67.8**	**18.6**
- MD	80.3	50.3	67.5	49.3
- UKR	79.2	25.8	66.3	22.6
- MKR	77.1	25.6	64.8	23.7

$$\mathcal{L}_{\text{MD}} = \text{Dis}\Big(\{f'(I_i, T_i) | (I_i, T_i) \in D_f\}, \{f(I_p, T_q) | (I_p, T_p) \in D_r, (I_q, T_q) \in D_r, p \neq q\}\Big),$$

$$\mathcal{L}_{\text{MKR}} = \text{Dis}\Big(f'(I_r, T_r), f(I_r, T_r)\Big), (I_r, T_r) \in D_r.$$

$$\mathcal{L}_{\text{UKR}} = \text{Dis}\bigg(\Big\{[f'_I(I_i); f'_T(T_i)] | (I_i, T_i) \in D_f\Big\}, \Big\{[f_I(I_i); f_T(T_i)] | (I_i, T_i) \in D_f\Big\}\bigg),$$

The more substantial impact observed by removing MKR can be attributed to two factors:

(1) $|D_r| > |D_f|$, leading to a much larger influence for MKR

(2) downstream tasks tend to rely more heavily on multimodal knowledge than unimodal knowledge, making MKR crucial for maintaining model performance

Utility of Unimodal Knowledge

	FineTune	NegGrad	DtD	L-codec	UL	MultiDelete	w/o UKR
Acc.	83.2	81.7	43.8	55.2	82.7	**83.6**	77.9

g_I : Image classifier, $f(I)$: unimodal embedding(before unlearning)

Hope that

$$g_I(f'(I)) \sim g_I(f(I))$$

The same for texts

Updating All Parameters vs. Fusion Module Only

上海科技大学 ShanghaiTech University

MultiDelete-F:

- Some how similar to bypassing the optimization for L_{UKR}
- exhibits less fluctuation in performance on D_{Test} during training, but tends to converge more slowly on D_f $|D_r$

scrubbing-based methods (DtD, L-codec)

- results in a complete loss previously acquired knowledge, resulting in random performance across all tasks
- *Conclusion* :
  1. robust unimodal knowledge plays a critical role in multimodal tasks
  2. the fusion module is more resilient to noise or minor perturbations than the unimodal encoders.

modality-agnostic approaches

- Little difference
- the strategy chosen for parameter updating has minimal impact on overall performance

	NLVR^2		PGR	
	D_{Test}	D_f	D_{Test}	D_f
RETRAIN	80.3	50.3	67.5	50.2
Full model	**80.8**	**23.5**	**67.8**	**18.6**
- MD	80.3	50.3	67.5	49.3
- UKR	79.2	25.8	66.3	22.6
MULTIDELETE-F	80.4	26.4	**67.7**	**19.5**
DTD	50.8	49.8	50.0	50.2
DTD-F	50.9	49.8	53.6	50.2
L-CODEC	55.3	52.7	57.8	48.8
L-CODEC-F	56.8	53.1	59.1	46.9

Efficiency (training time)

上海科技大学 ShanghaiTech University

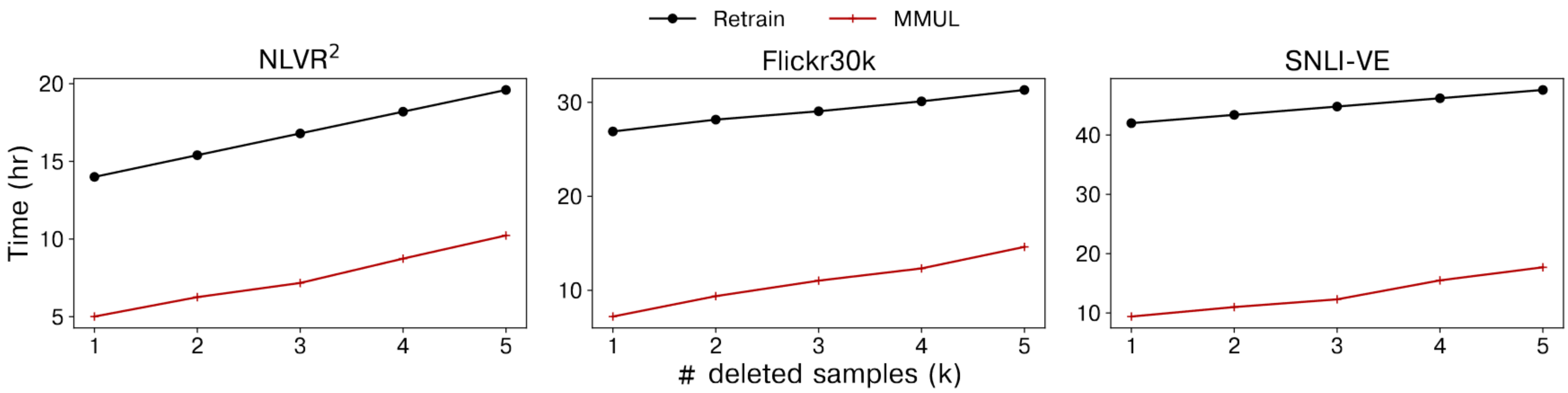


Fig. 2: Training time of unlearning methods.

- Linear growth
- MultiDelete-F only optimizes a small portion of the parameters

Conclusion

Compared to existing unimodal approaches, **MD** can remove the relationships between data modalities.

Compared to existing modality-agnostic approaches, **MKR** and **UKR** maintains the capability of model on multimodal tasks.