Protect, Show, Attend and Tell: Empowering Image Captioning Models with Ownership Protection

Jian Han Lim[1], Chee Seng Chan[1], Kam Woh Ng[2], Lixin Fan[2], Qiang Yang[2,3]

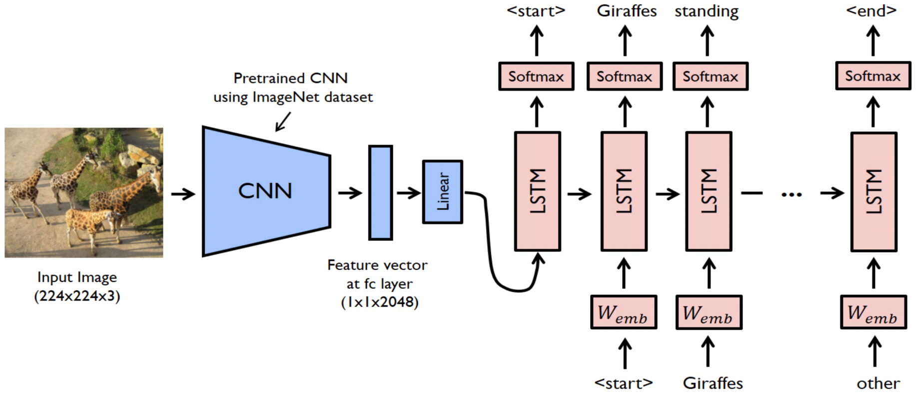
[1]University of Malaya, Kuala Lumpur, Malaysia
[2]WeBank AI Lab, Shenzhen, China
[3]Hong Kong University of Science and Technology

Presenter: Jian Han Lim

*Work is currently under review at Pattern Recognition

Introduction

- Existing Intellectual Property (IP) protection on deep neural networks (DNNs)
  - Follow a standard digital watermarking framework that was conventionally used to protect the ownership of multimedia and video content
  - Focus on image classification task
- IP protection on other tasks are forgotten such as image captioning that map images to texts



Introduction

- Why not directly apply existing watermarking methods designed for the classification DNNs to watermark the DNNs in image captioning?
  - Classification *outputs a label*; Image captioning *outputs a sentence*
  - Classification finds the *decision boundaries among different classes*; Image captioning *understands the image content and connect with a language model* to create a sentence

Aim

- Propose a novel embedding framework to protect the image captioning model
  - Consists of two different embedding schemes
  - To embed a unique secret key into the hidden memory state of an RNN
  - A forged key will yield an unusable image captioning model, poor quality outputs

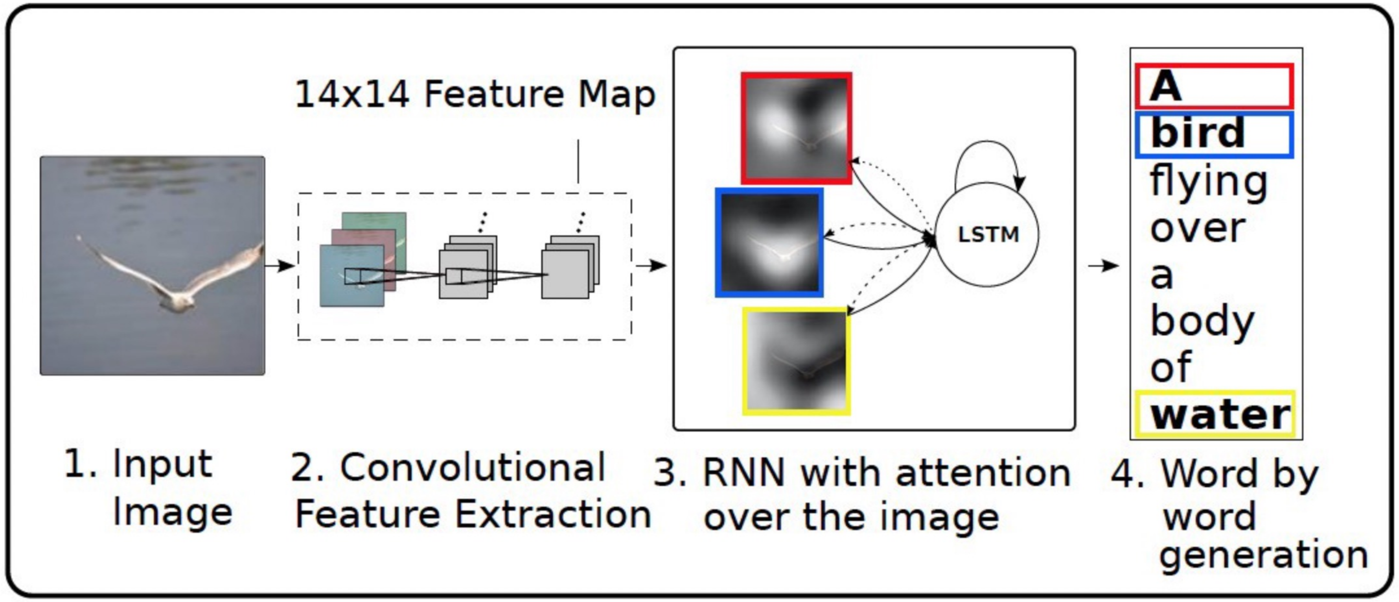
Contributions

1. We propose a key-based strategy that provides reliable, preventive and timely IP protection at virtually no extra cost for image captioning task

2. We empirically show the effectiveness of our approach against various attacks and prove the ownership of the model

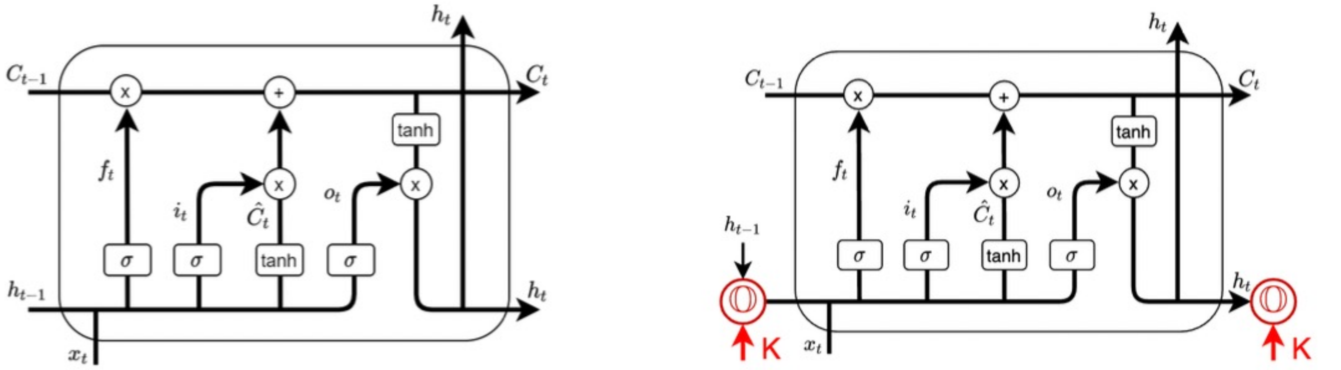
3. To the best of our knowledge, we are the first to propose IP protection on image captioning model that does not compromise the original image captioning performance

Show, Attend and Tell model [1]



[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: ICML (2015)

Proposed Approach



(a) Original LSTM Cell

(b) LSTM Cell with Secret Key Embedding

An overview of our approach. (a) The original LSTM Cell and (b) LSTM Cell with key embedding operation

Embedding Operation

- Introduce two different key embedding operations $\mathbb{O}$:
  - Element-wise addition model ($M_{\oplus}$)
  - Element-wise multiplication model ($M_{\otimes}$)

$$\mathbb{O}(K, h_{t-1}, e) = \begin{cases} K \oplus h_{t-1}, & \text{if } e = \oplus, \\ K \otimes h_{t-1}, & else. \end{cases}$$

Sign of Hidden State as Signature

- To further strengthen our model
  - Add the sign loss regularization term into the loss function as to [8]

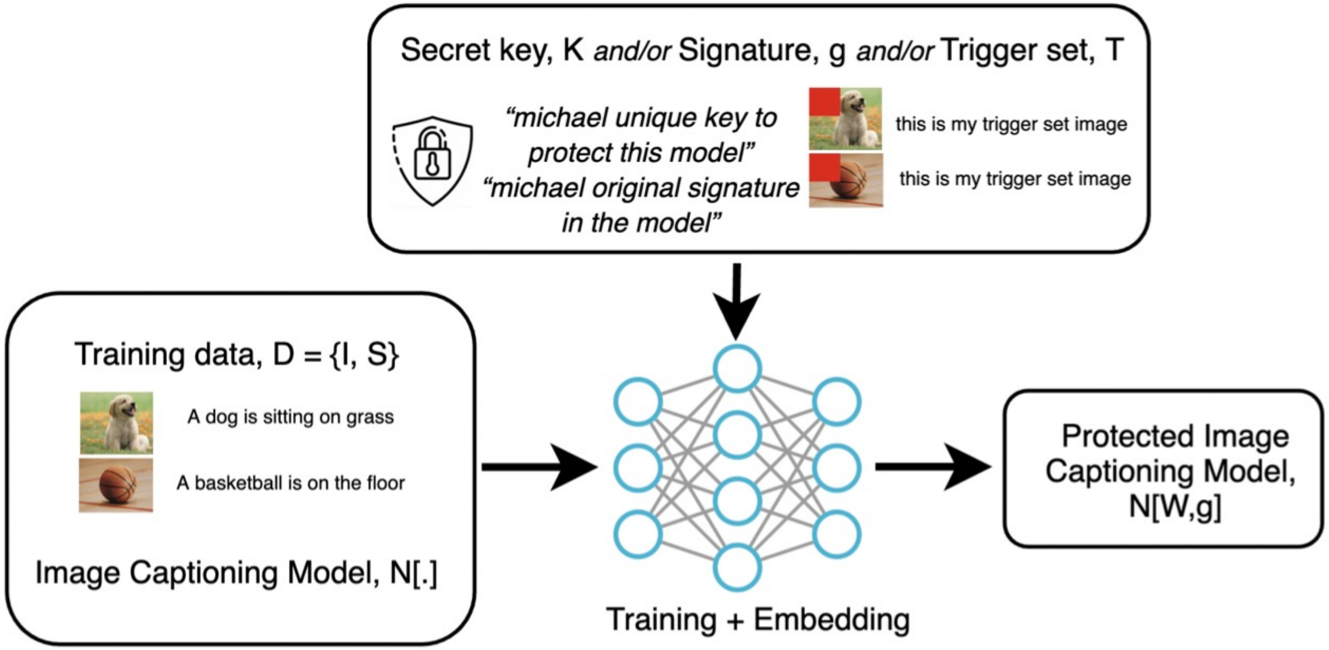
$$L_g(h, G, \gamma) = \sum_{i=1}^{N} max(\gamma - h_i g_i, 0)$$

  - where $G = \{g_i\}_{i=1}^{N}$ with $g_i \in \{-1, 1\}$ consists of the designated binary bits for hidden state h

  - Main difference compared to [8] is our signature is not embedded in the model weights

[8] L. Fan, K. W. Ng, C. S. Chan, Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks, in NeurIPS (2019)

Embedding Process



An embedding process E_O, takes inputs training data D = {I, S}, secret key K and/or signature g and/or trigger set T, model N[.] to produce protected model N[W,g].

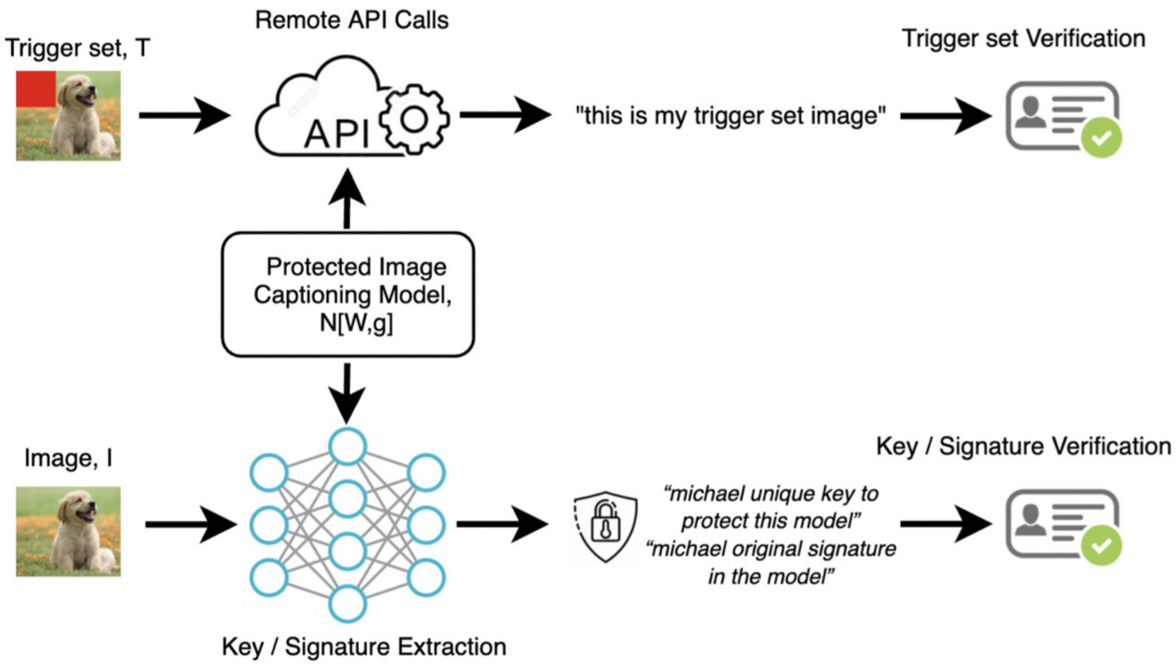
Ownership Verification

- Three verification methods are proposed:
  - V_1 : Key-based verification
  - V_2 : Signature-based verification
  - V_3 : Trigger set verification

- V_1 and V_2 are white-box verification
  - Required to have access to the model physically to verify the ownership

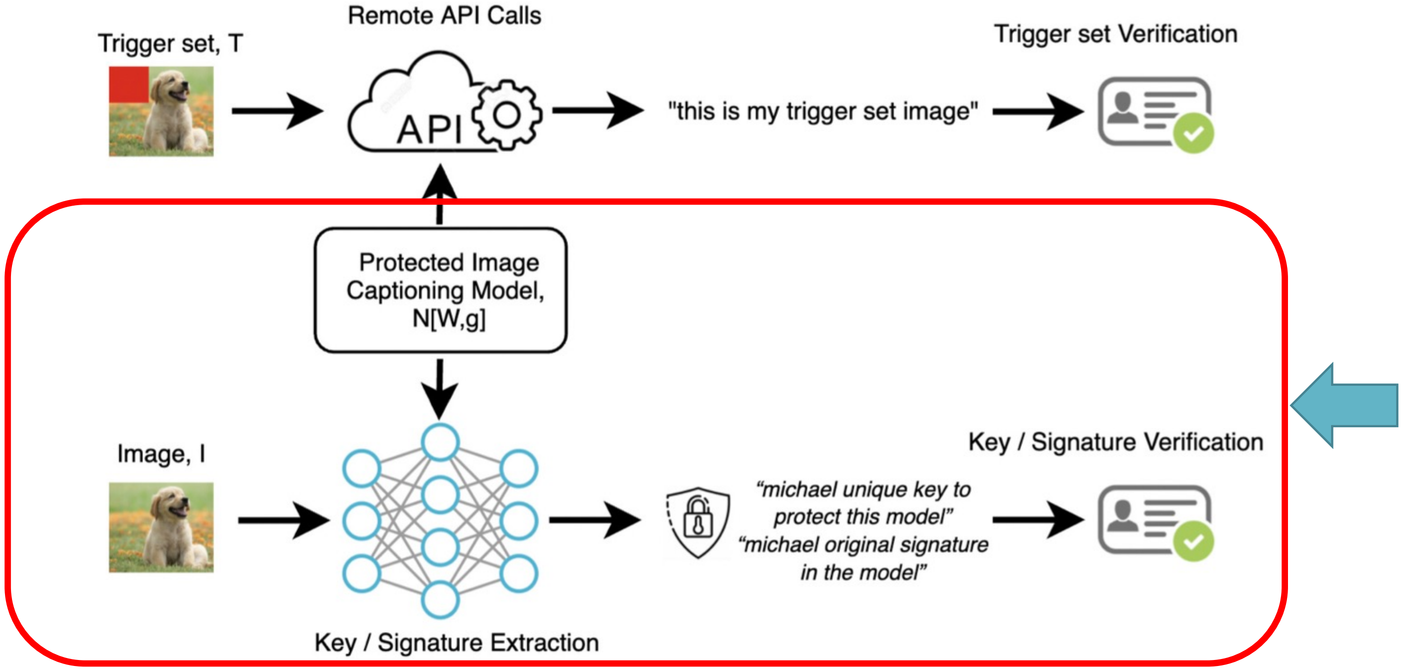
- V_3 is black-box verification
  - Can be conducted remotely via API calls

Verification Process



A verification process V takes as inputs, either an image I or a trigger set T, and outputs result to verify the ownership.

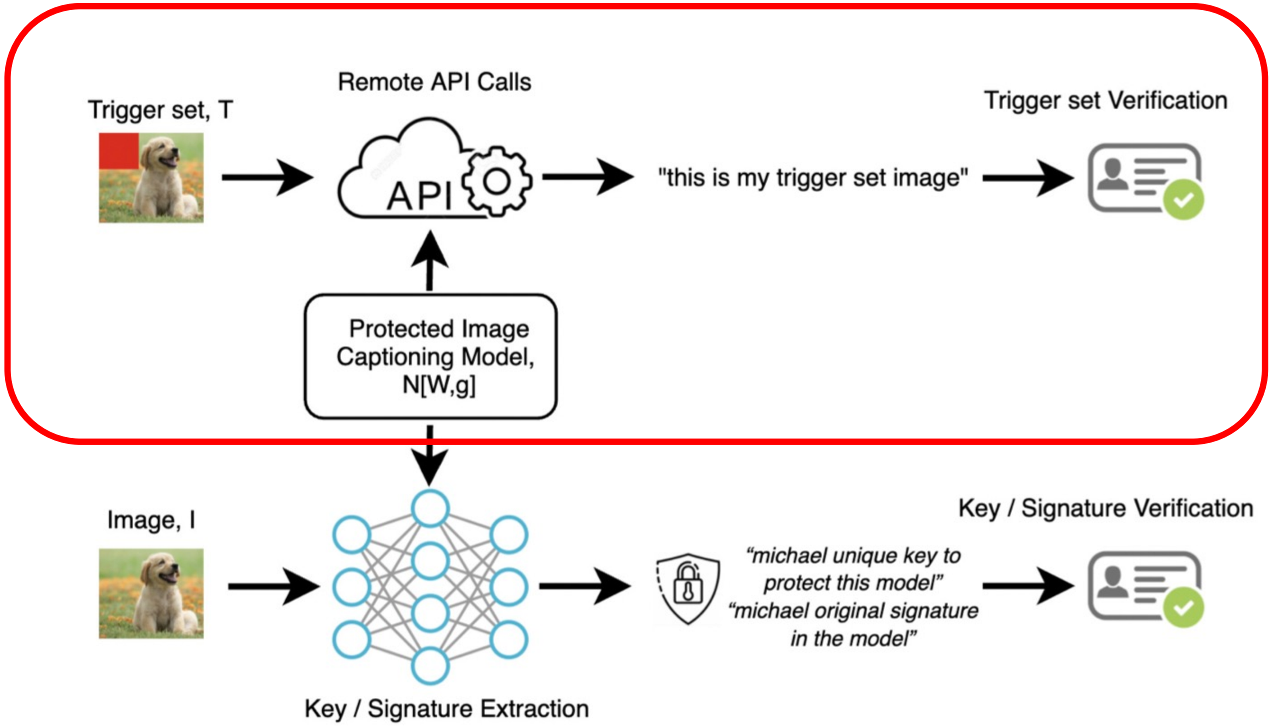
Verification Process



V_1 : Key-based verification

V_2: Signature-based verification

Verification Process



V_3 : Trigger set verification

Experiments

- Datasets:
  - **MSCOCO**
    - Contains 123,287 images
    - At least five human generated captions for each image
  - **Flick30k**
    - Contains 31,783 images
    - Focusing on people and animals
    - Five captions per image

Experiments

- Evaluation Metrics:
  - BLEU-N
  - ROUGE-L
  - METEOR
  - SPICE
  - CIDEr-D
- All these scores are obtained using the publicly available MSCOCO evaluation toolkit

Comparison with CNN-based watermarking framework

Comparison between our approaches ($M_\oplus$, $M_\otimes$) with baseline and Passport [8] on MSCOCO and Flickr30k datasets. **BOLD** is the best result and * is the second best result

Methods	MS-COCO								Flickr30k							
	B-1	B-2	B-3	B-4	M	R	C	S	B-1	B-2	B-3	B-4	M	R	C	S
Baseline	72.14	55.70	41.86	31.14	24.18	52.92	94.30	17.44	63.40	45.18	31.68	21.90	18.04	44.30	41.80	11.98
Passport [8]	68.50	53.30	38.41	29.12	21.03	48.80	84.45	15.32	48.30	38.23	26.21	17.88	15.02	32.25	28.22	9.98
$M_\oplus$	**72.53**	**56.07**	**42.03**	**30.97**	**24.00**	**52.90**	*91.40	*17.13	**62.43**	**44.40**	**30.90**	**21.13**	*17.53	**43.63**	*40.07	*11.57
$M_\otimes$	*72.47	*56.03	*41.97	*30.90	*23.97	**52.90**	**91.60**	**17.17**	*62.30	*44.07	*30.73	*21.10	**17.63**	*43.53	**40.17**	**11.67**

[8] L. Fan, K. W. Ng, C. S. Chan, Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks, in NeurIPS (2019)

Comparison with CNN-based watermarking framework

(a) **a woman wearing a hat and a scarf.**

(b) a woman in a scarf is standing.

(c) a woman in a hat is standing.

(d) a woman.

(a) **a man in a white shirt is standing in a room.**

(b) a man in a white shirt is standing in a room.

(c) a man in a white shirt is standing in a room.

(d) a man in room.

(a) **a man drinking a drink.**

(b) a man is drinking beer.

(c) a man is drinking a beer.

(d) a man is drinking.

Comparison of caption generated by (a) Baseline, (b) $M_{\oplus}$, (c) $M_{\otimes}$, and (d) Passport [8]

Comparison with CNN-based watermarking framework

Comparison between Passport [8] with (top) correct passport and (bottom) forged passport on MSCOCO and Flickr30k datasets.

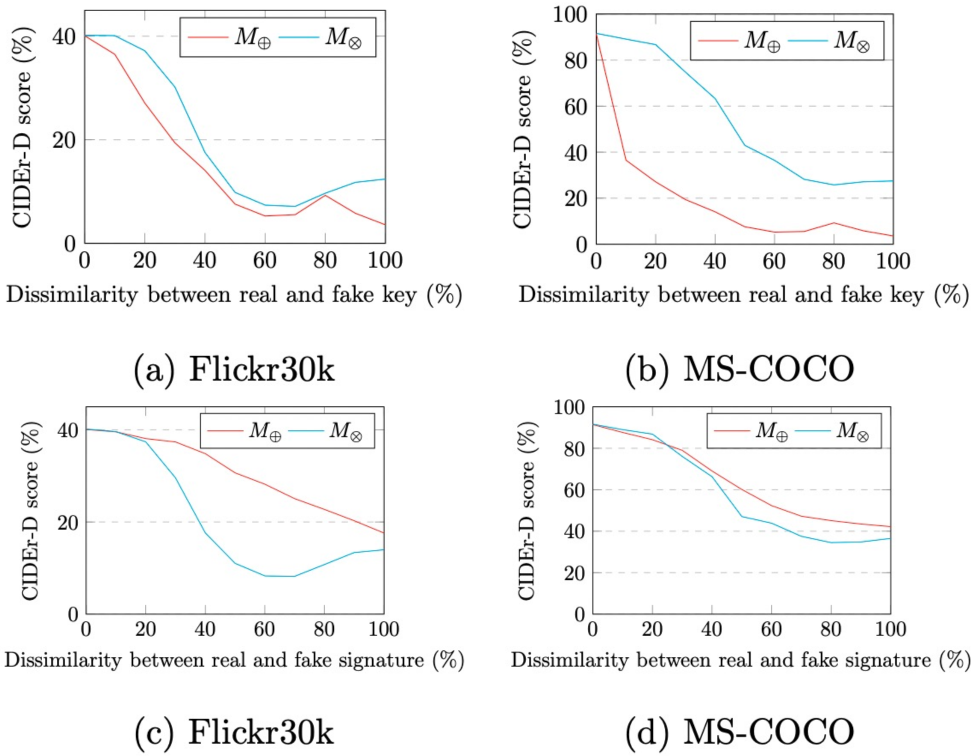
Methods	MS-COCO								Flickr30k							
	B-1	B-2	B-3	B-4	M	R	C	S	B-1	B-2	B-3	B-4	M	R	C	S
Passport	68.50	53.30	38.41	29.12	21.03	48.80	84.45	15.32	48.30	38.23	26.21	17.88	15.02	32.25	28.22	9.98
$Pas\bar{s}port$ (forged)	67.50	52.65	37.15	29.01	20.95	47.90	83.00	15.00	47.30	37.87	26.01	17.10	14.82	31.88	26.50	9.90

Fidelity Evaluation

Comparison between our approaches ($M_{\oplus}, M_{\otimes}$) with baseline and Passport [8] on MSCOCO and Flickr30k datasets. **BOLD** is the best result and * is the second best result

Methods	MS-COCO								Flickr30k							
	B-1	B-2	B-3	B-4	M	R	C	S	B-1	B-2	B-3	B-4	M	R	C	S
Baseline	72.14	55.70	41.86	31.14	24.18	52.92	94.30	17.44	63.40	45.18	31.68	21.90	18.04	44.30	41.80	11.98
Passport [8]	68.50	53.30	38.41	29.12	21.03	48.80	84.45	15.32	48.30	38.23	26.21	17.88	15.02	32.25	28.22	9.98
$M_{\oplus}$	**72.53**	**56.07**	**42.03**	**30.97**	**24.00**	**52.90**	*91.40	*17.13	**62.43**	**44.40**	**30.90**	**21.13**	*17.53	**43.63**	*40.07	*11.57
$M_{\otimes}$	*72.47	*56.03	*41.97	*30.90	*23.97	**52.90**	**91.60**	**17.17**	*62.30	*44.07	*30.73	*21.10	**17.63**	*43.53	**40.17**	**11.67**

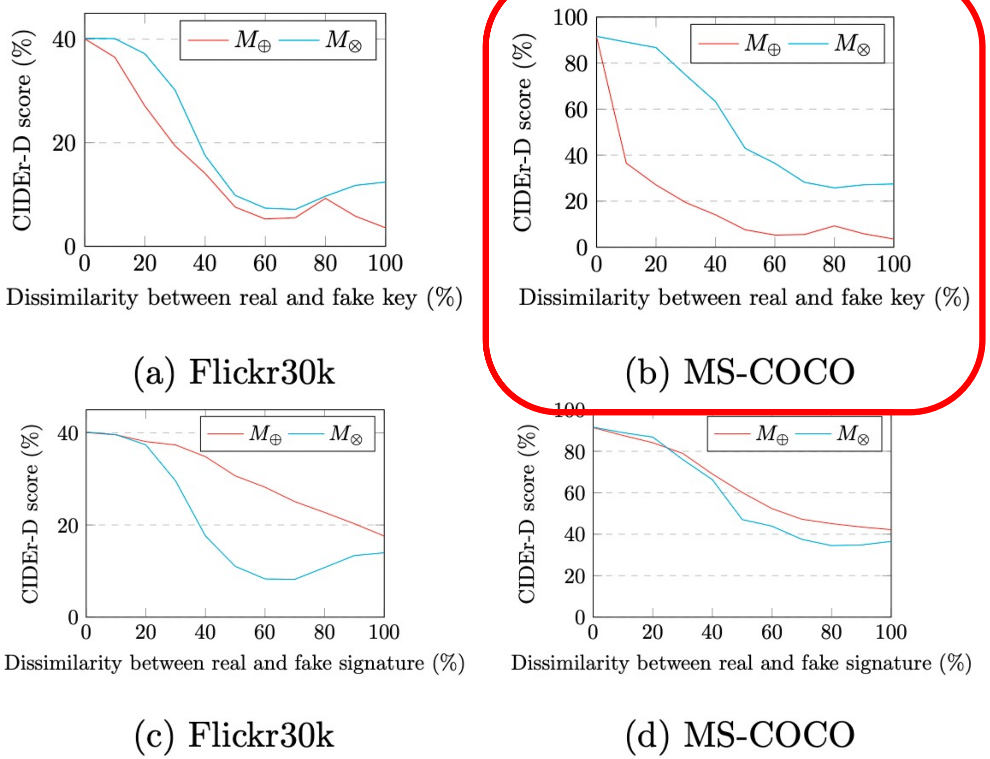
Resilience against ambiguity attacks



(a) Flickr30k (b) MS-COCO (c) Flickr30k (d) MS-COCO

CIDEr-D on Flick30k and MSCOCO under ambiguity attack on (a-b) key; (c-d) signature.

Resilience against ambiguity attacks



CIDEr-D on Flick30k and MSCOCO under ambiguity attack on (a-b) key; (c-d) signature.

Resilience against ambiguity attacks

(a) a cat laying on top of a wooden floor.

(b) a cat laying on the floor of a wooden floor.

(c) a cat standing on a wooden floor.

(d) a cat on the floor on the floor on a wooden floor.

(a) a group of people standing next to a truck.

(b) a group of people standing next to a truck.

(c) a group.

(d) a group of people in a green and a large green and a large green and a large green and.

(a) a man is standing in a living room.

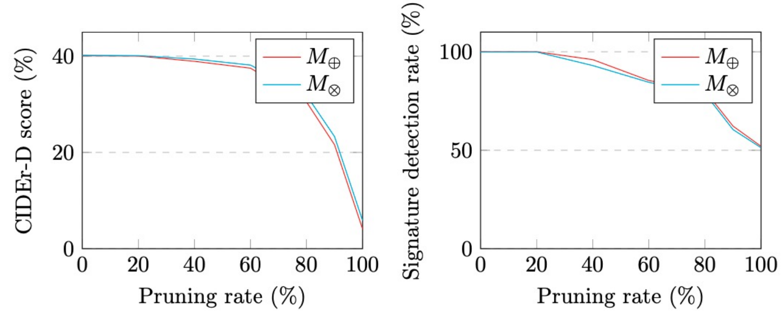
(b) a man is standing in a living room.

(c) a man and a man and a man and a man and a man and a man and a man.

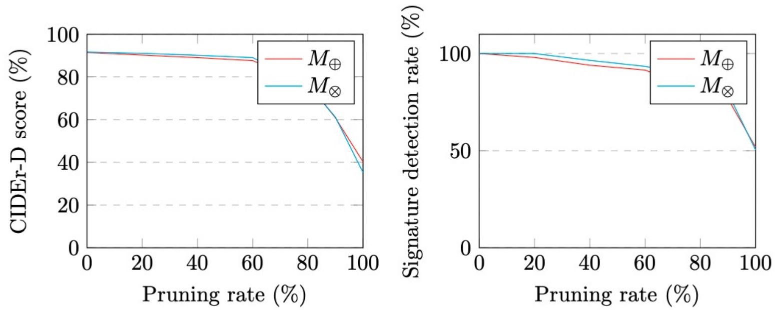
(d) a man and a man in a man and a man in a man and a man and a man.

Comparison of caption generated by (a) Baseline, (b) $M_{\otimes}$, (c) $M_{\otimes}$ with the forged key has 75% similarity as to real key and (d) $M_{\otimes}$ with the forged key has 50% similarity as to real key.

Robustness against removal attacks



(a) Flickr30k



(a) MS-COCO

Removal attack (Model Pruning): CIDEr-D score and signature detection rate of our approaches on both MSCOCO and Flickr30k against different pruning rates

Robustness against removal attacks

(a) a group of people sitting on a beach.
(b) a group of people sitting on a beach with umbrellas.
(c) a group of people sitting on a beach.

(a) a dog sitting on a chair in front of a tv.
(b) a man sitting on a chair next to a tv.
(c)a man is sitting on a chair in front of dog.

(a) a zebra laying down on a sandy beach.
(b) a zebra laying down on a dirt ground.
(c) a zebra laying down on ground.

Comparison of caption generated by (a) Baseline, (b) $M_{\otimes}$ and (c) $M_{\otimes}$ with 60% pruning rate

Robustness against removal attacks

Removal attack (Fine-tuning): CIDEr-D score (in-bracket) of baseline and proposed models (Left: MSCOCO fine-tune on Flickr30k. Right: vice-versa). Accuracy (%) outside bracket is the signature detection rate.

Methods	MS-COCO		Flickr30k	
	MS-COCO	Flickr30k	Flickr30k	MS-COCO
Baseline	- (94.30)	- (37.70)	- (41.80)	- (88.50)
$M_{\oplus}$	100 (91.40)	70.40 (37.50)	100 (40.07)	72.50 (87.30)
$M_{\otimes}$	99.99 (91.60)	71.50 (37.8)	99.99 (40.17)	71.35 (86.50)

Conclusion

- We **take the first step** to implement the ownership protection on the image captioning task
- Proposed the key-based protection using the hidden memory state of RNN
- Demonstrated with extensive experiments that the image captioning models are well-protected for unauthorized usages.