Optics Lens Design For Privacy-Preserving Scene Captioning

3216

Project Page

Paula Arguello[1], Jhon Lopez[1], Carlos Hinojosa[1] Henry Arguello[1]
[1]Universidad Industrial de Santander
paula2191444@correo.uis.edu.co

High Dimensional Signal Processing Research Group

Motivation

Image captioning task consists on use traditional images to generate a natural language description of the scene

a girl stands on the beach with a horse

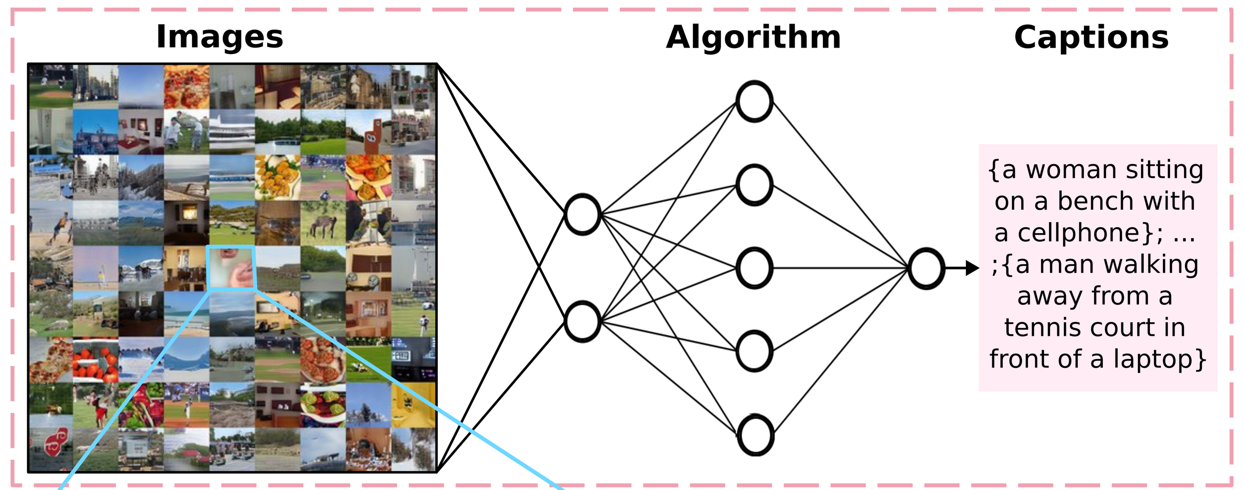
a little boy flying his kite in the yard

Image captioning is applicable in various scenarios:

- usage in virtual assistants
- support of the disabled

Traditional Image Captioning Computational Approaches

Previous works have addressed the image captioning problem from different approaches.
Most of them use **RNN** and **LSTM** networks for processing long sequences.

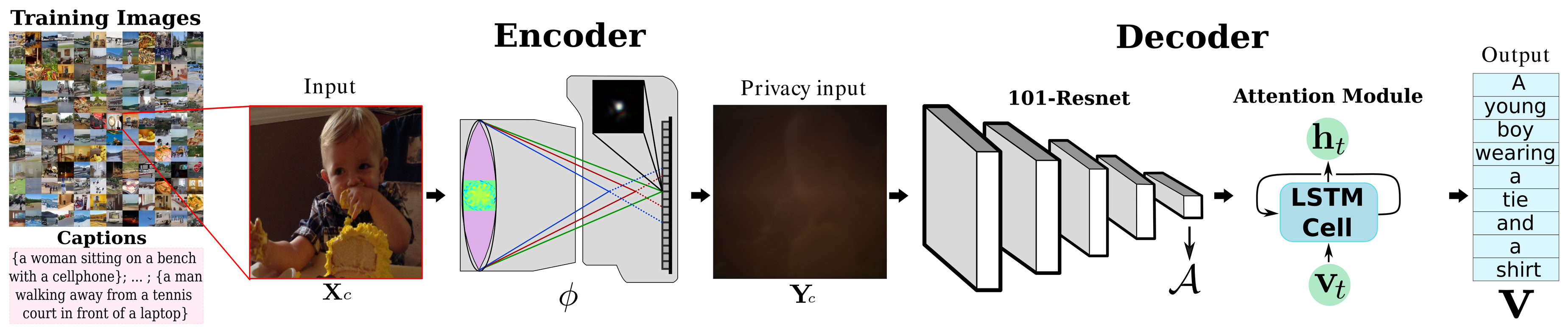


In traditional image captioning pipelines, cameras are used to acquire **high-fidelity images**.

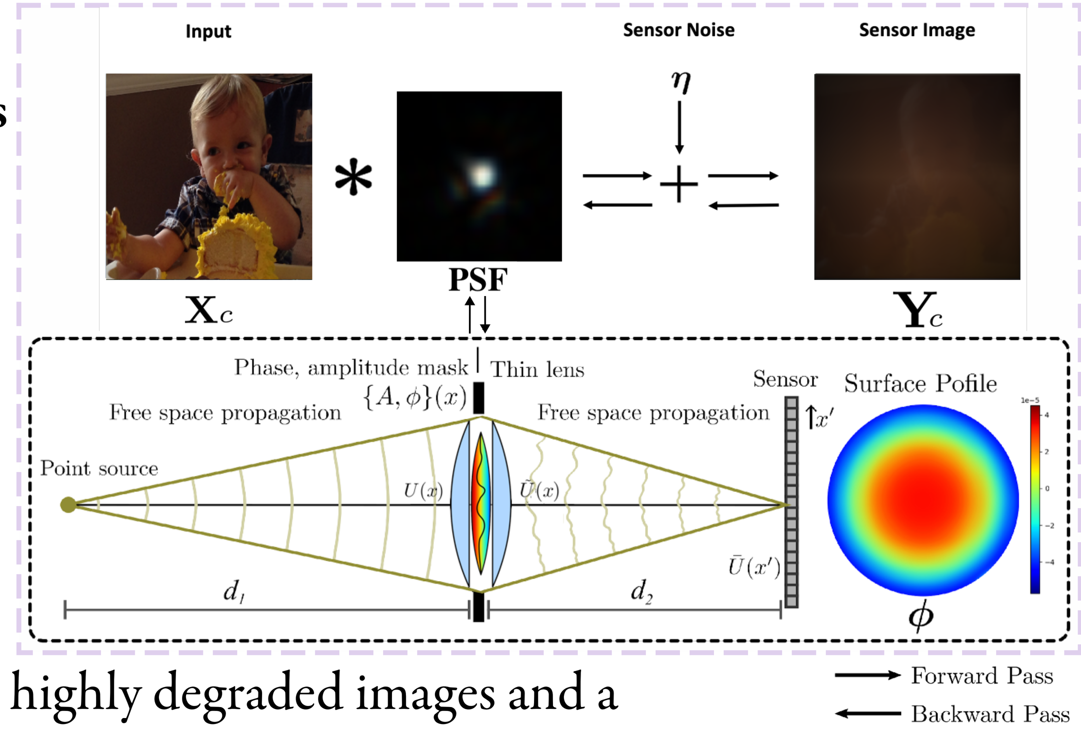
However, the acquired images may contain **privacy-sensitive** data.

Model and Approach

We propose a Encoder-Decoder end-to-end arquitecture to learns optics by backpropagating the gradients from the captioning network decoder to the optics layer



- We rely on the converse approach of deep optics: We **add aberrations** to the lens to obtain privacy protection and jointly perform IC.
- Our optimization process has two parts: an optical encoder, which provides hardware-level privacy protection by degrading the image quality, and a decoder with a **CNN** that learns features from the highly degraded images and a **LSTM** with an attention module that compute captions.



End-to-end Optimization

Formally, we formulate our optimization problem by combining the two goals: to acquire privacy-preserving images and to perform HPE with high accuracy.

$$\mathcal{L} = -\log(p(\mathbf{v} \mid \mathcal{A})) + \lambda \sum_{i=1}^{L}\left(1-\sum_{t=1}^{C}\boldsymbol{\theta}_{ti}\right)^2 - \sum_{c=1}^{C}\log\frac{\exp(\mathbf{v}_c)}{\exp\left(\sum_{i=1}^{C}\mathbf{v}_i\right)}\mathbf{g}_c + \left(1-\frac{1}{J}\sum_{l=1}^{3}\|\mathbf{Y}_\ell-\mathbf{X}_\ell\|^2\right)$$

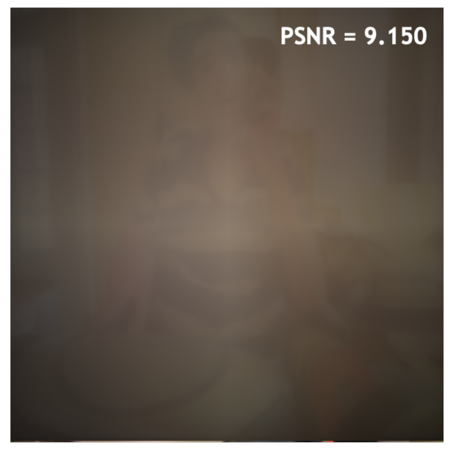
Datasets and Metrics

We train our proposed end-to-end approach on the COCO 2014 dataset and evaluate our approach on the val2014 set.

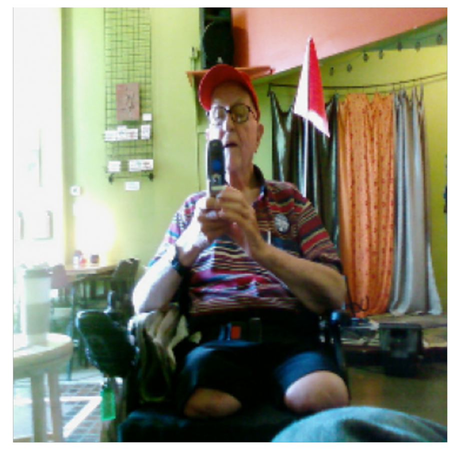
Captioning	Face Recognition	Image Quality
To quantitatively evaluate captions, we use the standard **BLEU** and **Meteor** metrics. With values closer to 100 representing more similar texts.	We implement the ArcFace network to measure privacy. We train ArcFace on three face recognition datasets. We measure its performance in terms of the area under the curve (**AUC**) of the **ROC**.	To measure image degradation, we use the peak-signal-to-noise ratio (**PSNR**) and the structural similarity index measure (**SSIM**). We expect to achieve lower PSNR and SSIM values.

Qualitative Results on Example COCO Images

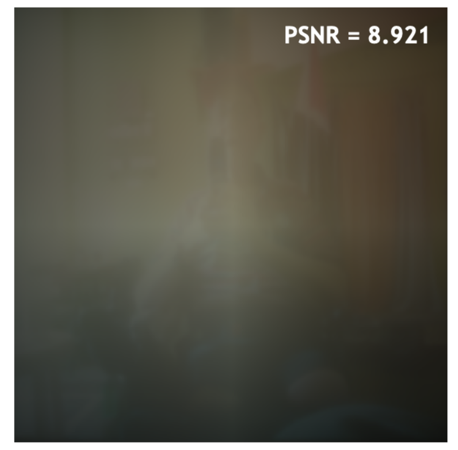
a man sitting at a table in a wheelchair while on a phone



a person in a wheelchair talking on a telephone

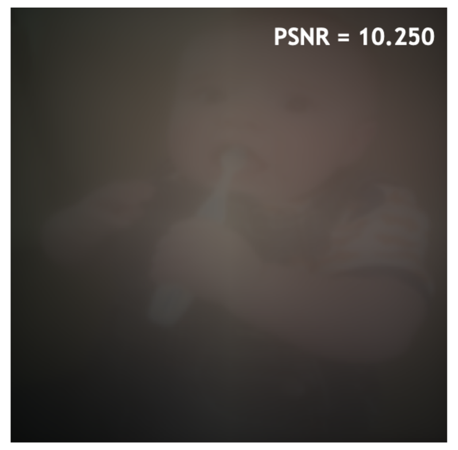


an elderly man looks at a cell phone



an old man looks at a cell phone screen

a child holds a toothbrush in their hand



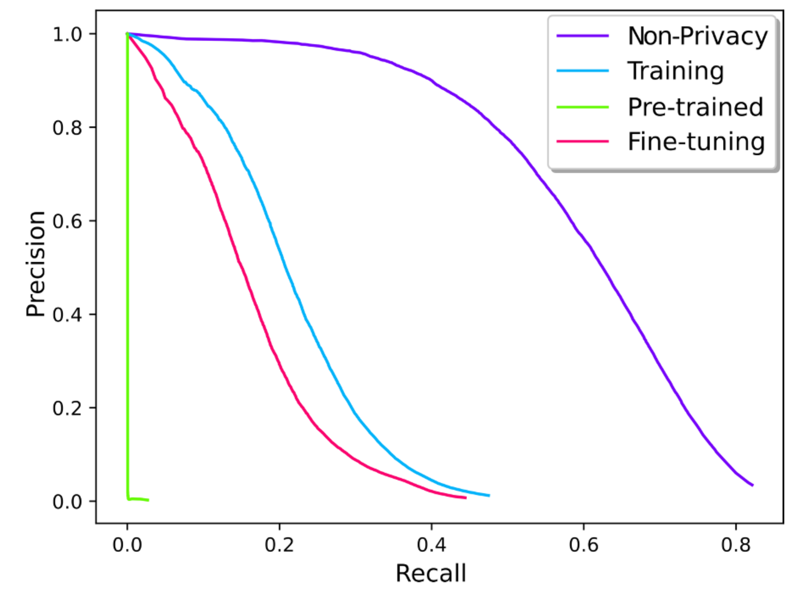
a baby holding a toothbrush in its mouth

two children standing at the sink brushing their teeth

a little girl is brushing her teeth in a bathroom

Experiments: Ablation Studies

1. Non-privacy: We trained the face detection model from scratch with original images resized.
2. Training: We trained the face detection model from scratch using blurred images.
3. Pre-trained: We evaluated the previous experiment (Non-privacy) on distorted images
4. Fine-tuning: We perform fine-tuning on the Non-privacy experiment using the blurred images.



Quantitative Experiments: Comparison with Prior Works

	Method	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor
Non - Privacy	BRNN	64.2	45.1	30.3	20.1	19.5
	NIC	66.6	46.1	32.9	24.6	23.7
	CutMix	64.2	-	-	24.9	23.1
	AAIC	71.0	-	-	27.7	23.8
	Hard Attn	71.8	50.4	35.7	25.0	23.0
	2PSC-w	72.1	54.8	40.4	29.6	29.2
Privacy	2PSC	70.7	53.5	39.4	28.9	29.0
	Defocus	56.1	36.7	24.2	16.3	20.4
	Low-Res	57.3	37.8	25.2	17.4	20.9

We compare our method (2PSC) against two traditional privacy-preserving approaches: Defocus and Low-Resoluion cameras.