High Dimensional Signal Processing Research Group

Optics Lens Design For Privacy-Preserving Scene Captioning

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

Project Page

Motivation

Image captioning (IC) task consists on using an image to generate a natural language description of the scene

a girl stands on the beach with a horse

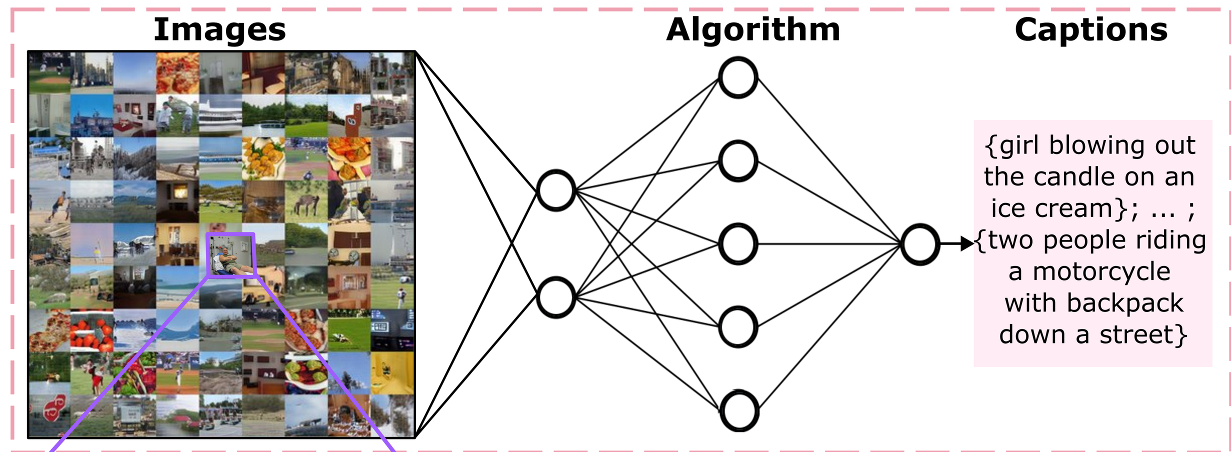
a little boy flying his kite in the yard

Image captioning applications:

- virtual assistants
- support of the disabled
- image classification
- social media

Traditional IC Approaches

Traditional works have addressed the image captioning problem with **DNN**, **CNN**, **RNN** and **LSTM** networks for processing long sequences [1].



Traditional cameras are used to acquire **high-fidelity images.**

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

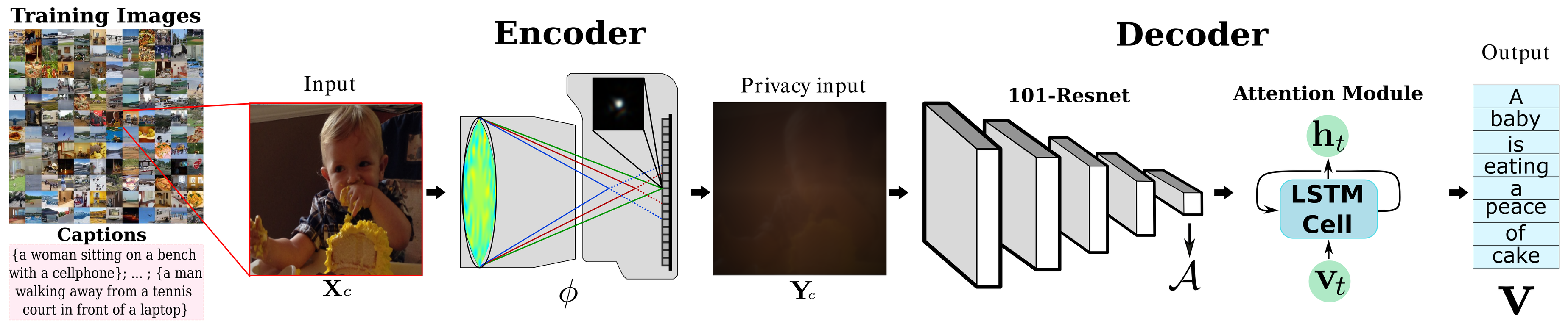
Bibliography

[1] XU, Kelvin, et al. Show, attend and tell: Neural image caption generation with visual attention. En International conference on machine learning. PMLR,2015.p.2048-2057

[2] Hinojosa, C, et al. Learning privacy-preserving optics for human pose estimation. In Proceedings of the IEEE/CVF International Conference onComputer Vision, pp.2573-2582.

Model and Approach

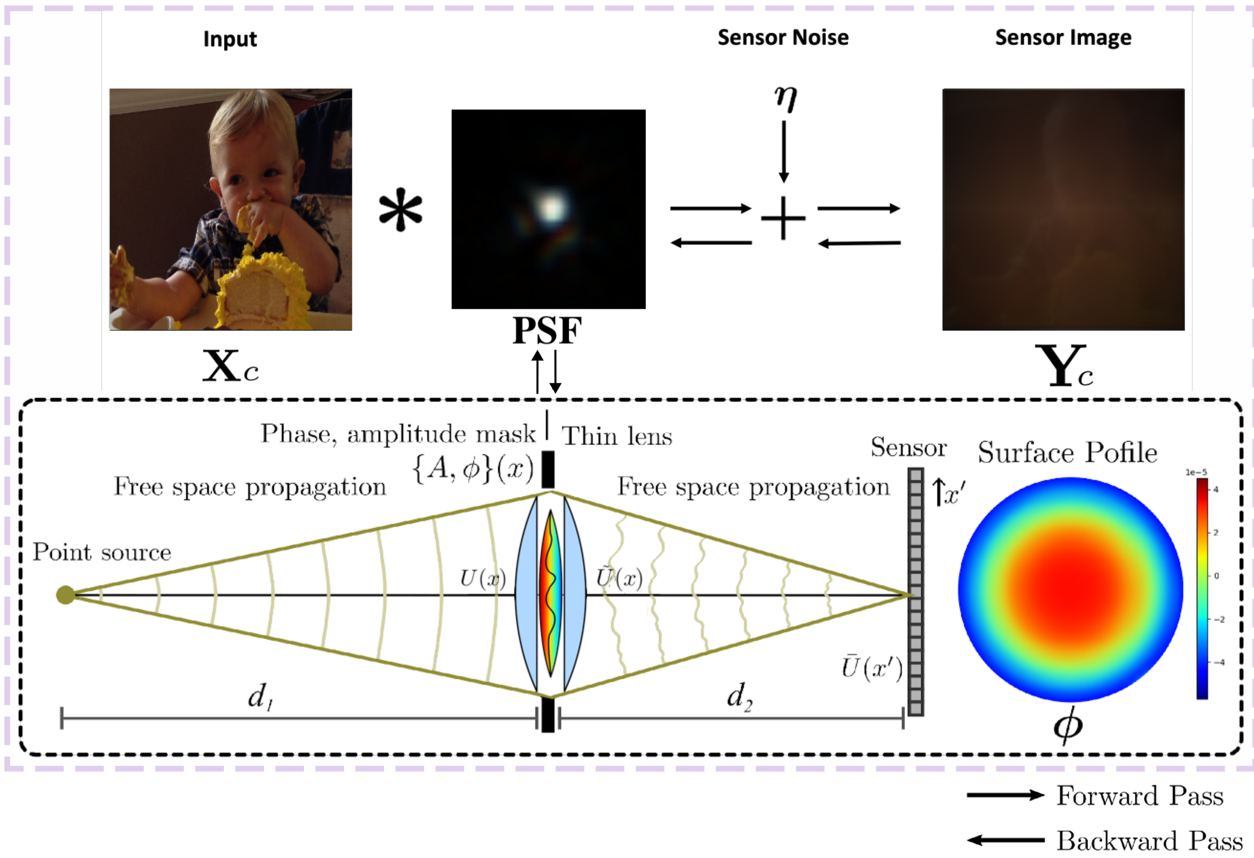
We propose an Encoder-Decoder network arquitecture optimized in an end-to-end approach to design a camera that preserves the privacy and generate captions.



We **add aberrations** to the lens to obtain privacy protection and perform IC.

Our optimization process has two parts:

- Optical Encoder: hardware-level privacy protection.
- Decoder: **CNN** (Feature learning) + **LSTM** (Caption generator).

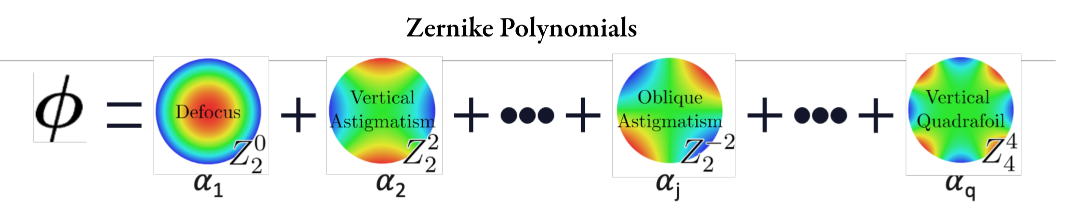


End-to-end Optimization

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

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

- We optimize the PSF by learning to add optical aberrations to the system [2].



Datasets and Metrics

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

Captioning	Face Recognition	Image Quality
To evaluate captions, we use the **BLEU** and **Meteor** metrics. With values closer to 100 representing more similar texts.	We implement the RetinaFace network to measure privacy. We measure its performance in terms of the **ROC** curve.	To measure image degradation, we use the peak-signal-to-noise ratio (**PSNR**). We expect to achieve the lowest value

Qualitative Results on Example COCO Images

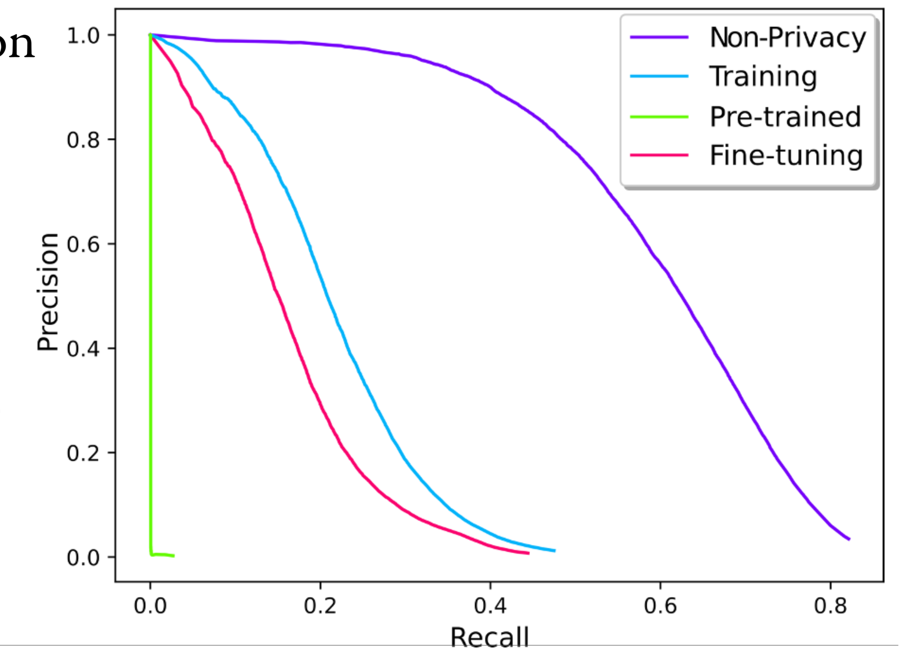
Original Image | Sensor Image | Original Image | Sensor Image

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

a man sitting at a table in a wheelchair while on a phone | a person in a wheelchair talking on a telephone | two children standing at the sink brushing their teeth | a little girl is brushing her teeth in a bathroom

Privacy Validation: Face Detection

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



Quantitative 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-Resolution cameras.