Show and tell

A Neural Image Caption Generator

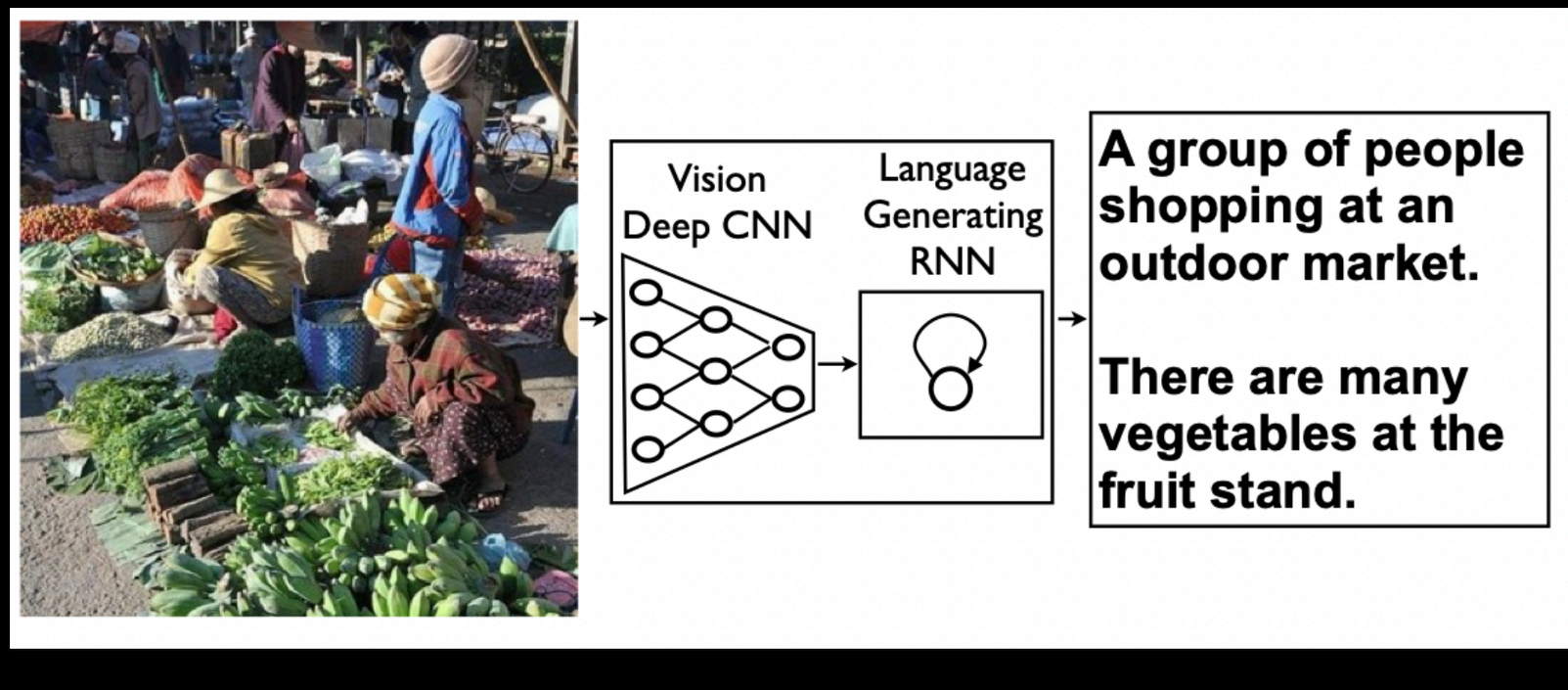
Diego Moreno, mayo 2021

Introducción

- Descripción automática del contenido de una imagen.
- Visión por computadora y NLP.
- Tarea más difícil que solo de visión.
- Descripción debe capturar:
  - Relación de objetos
  - Actividades
- Modelo de lenguaje y comprensión visual.

Inspiración

- Traducción automática.
  - Maximizando p(T|S).
- RNN:
  - Codificador (oración fuente).
  - Decodificador (oración objetivo).
- Reemplazar por CNN.
- Neural Image Caption (NIC).



Modelo propuesto

- Maximizar la probabilidad:

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

  - S es cualquier oración (no acotada)

- Suponemos longitud N (cadena):

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

  - SGD para optimizar.

- Se modela a *p*, como RNN (h_t).

$$p(S_t|I, S_0, \ldots, S_{t-1})$$

- La memoria se actualiza con:

$$h_{t+1} = f(h_t, x_t)$$

- LSTM para generación de enunciados.

LSTM

Para generación de oración

- Compuerta de entrada:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1})$$

- Compuerta del olvido:

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1})$$

- Compuerta de salida:

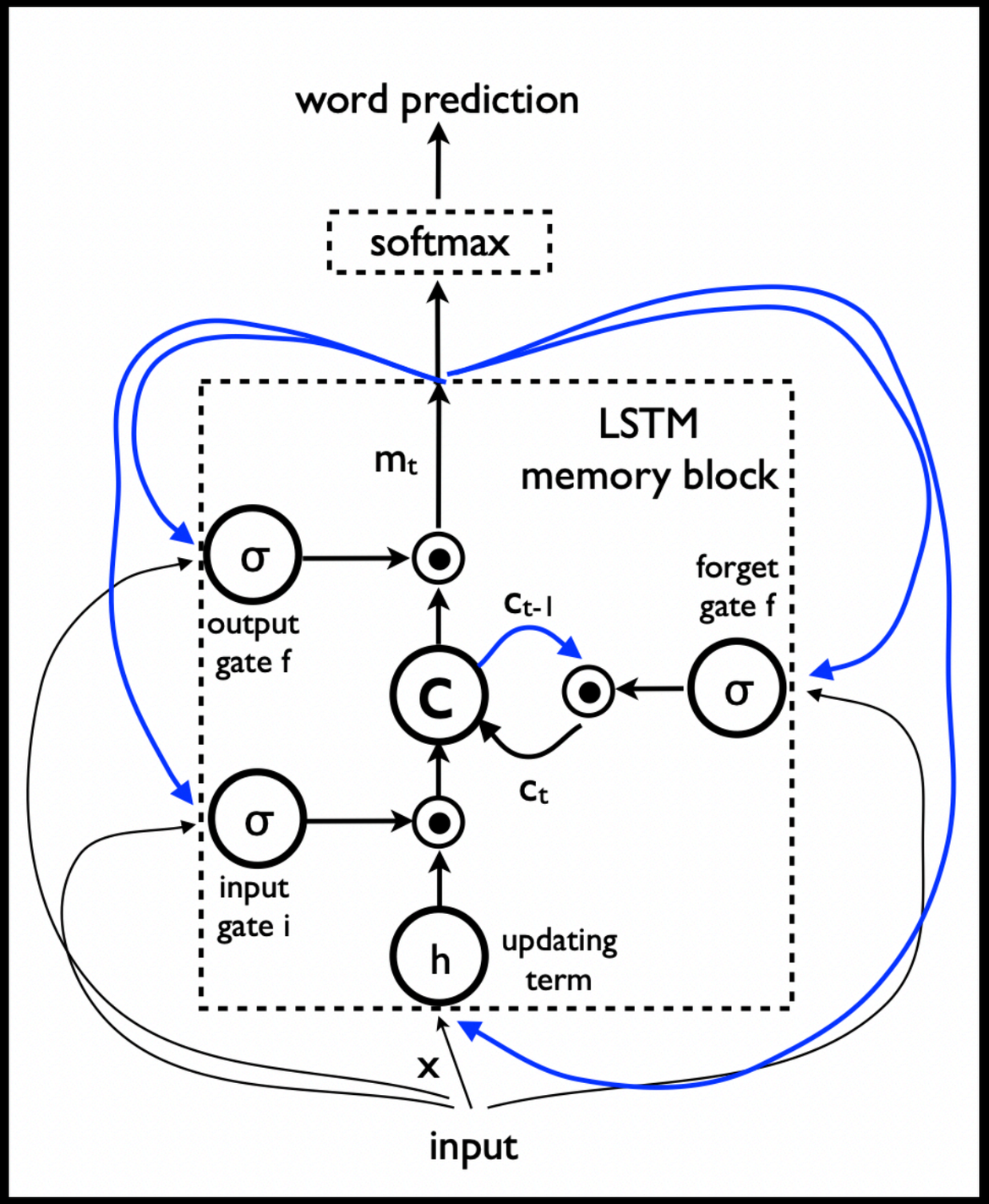
$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1})$$

- Célula principal:

$$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1})$$

- Salida:

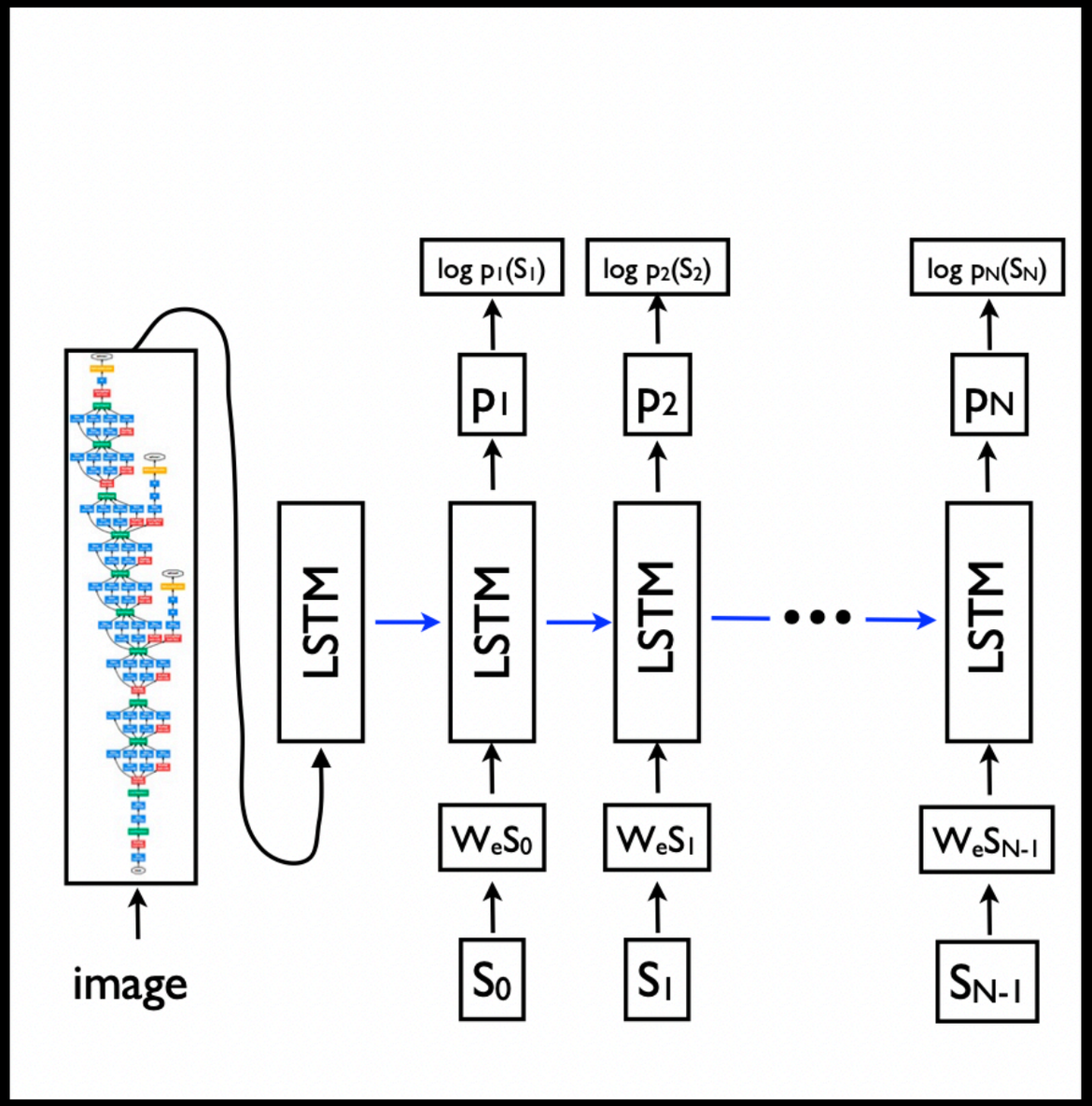
$$\begin{aligned} m_t &= o_t \odot c_t \\ p_{t+1} &= \text{Softmax}(m_t) \end{aligned}$$



Entrenamiento

- Se toma una imagen I, con una descripción verdadera S=(S0, …, SN).

$$
\begin{aligned}
x_{-1} &= \mathrm{CNN}(I) \\
x_t &= W_e S_t, \quad t \in \{0 \ldots N-1\} \\
p_{t+1} &= \mathrm{LSTM}(x_t), \quad t \in \{0 \ldots N-1\}
\end{aligned}
$$



- Palabras Si, (one-hot).
- S0, y SN, especiales.
- Loss: log-verodsimilitud negativa:

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t)$$

Inferencia

Captioning

- Sampling: tomar la palabra de más probabilidad.
- BeamSearch: mantener k mejores.
- Evaluación:
  - BLEU score: forma de precisión entre los n-gramas generados y los de referencia.
  - Perplejidad, TEOR, Cider.

Resultados

Detalles del entrenamiento

- La CNN inicializada por pesos de ImageNet, cambio-impacto negativo.
- El We sin inicializar debido a no ganancias significativas.

- Conjunto de datos utilizado:

Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [6]	-	-	1000
Flickr8k [26]	6000	1000	1000
Flickr30k [33]	28000	1000	1000
MSCOCO [20]	82783	40504	40775
SBU [24]	1M	-	-

Resultados

- Para el conjunto de datos MSCOCO:

Metric	BLEU-4	METEOR	CIDER
NIC	**27.7**	**23.7**	**85.5**
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Word	Neighbors
car	van, cab, suv, vehicule, jeep
boy	toddler, gentleman, daughter, son
street	road, streets, highway, freeway
horse	pony, donkey, pig, goat, mule
computer	computers, pc, crt, chip, compute

- BLEU scores:

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text [24]				11
TreeTalk [18]				19
BabyTalk [16]	25			
Tri5Sem [11]			48	
m-RNN [21]		55	58	
MNLM [14][5]		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

Conclusiones

- Se presenta NIC, (end-to-end) a partir de image genera descripción razonable en un lenguaje sencillo.
- Los experimentos en varios conjuntos de datos muestran la robustez de NIC en:
  - Cualitativos (las oraciones generadas son muy razonables)
  - Cuantitativas, (métricas de clasificación o BLEU)
- A medida que aumenta el tamaño de los conjuntos de datos, también aumenta el rendimiento de NIC.

Referencias

- Vinyals, O., Toshev, A., Bengio, S. Y Erhan, D. *Show and Tell: A Neural Image Caption Generator*. arXiv:1411.4555, 2015.