Análisis de Paper Científico

BERT-hLSTMs: BERT and hierarchical LSTMs for visual storytelling

Jing Su, Qingyun Dai, Frank Guerin, Mian Zhoud

Universidad de Cuenca
Optativa - Minería de Textos
Freddy L. Abad L., Kevin A. Maxi J., David E. Santos L.
{freddy.abadl, kevin.maxi, david.santos}@ucuenca.edu.ec

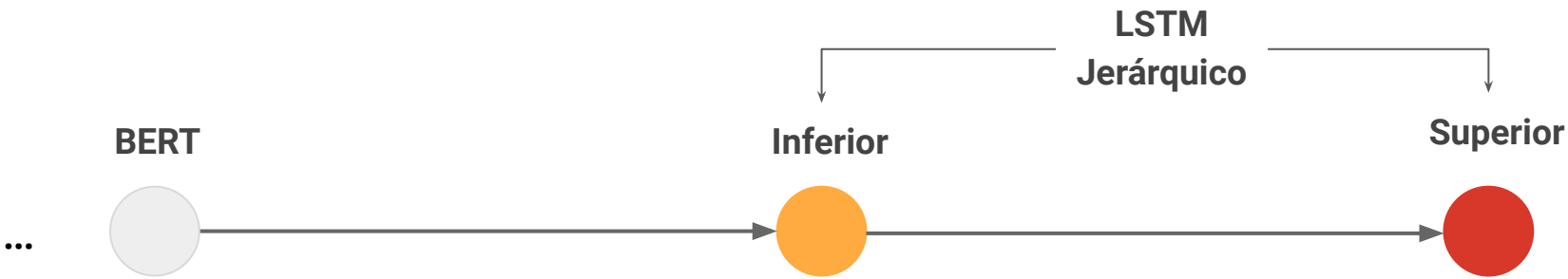
ÍNDICE

RESUMEN GLOBAL

Narración visual

Tarea creativa que busca generar automáticamente una descripción similar a una historia para una secuencia de imágenes.

Propone un marco narrativo visual jerárquico que modela por separado la semántica a nivel de oración y de palabra.



Enriquecimiento de significado de historias a nivel de oraciones y palabras.

Recibe entrada la representación vectorial de la oración. Aprender las dependencias entre las oraciones de las imágenes.

Genera las representaciones de vector de palabra.

CONTEXTO

A

Narrativa: Contar historias como forma de educar, culturizar y aconsejar.

IA: “Utilizar una máquina para generar automáticamente una secuencia de oraciones coherentes para una secuencia de imágenes ordenada ” [Cho, Huang, Yu et al., 2017]

Narrativa visual nace del uso del deep learning en la traducción automática y los subtítulos de imágenes. ***La narración visual NO ES SUBTITULAR IMÁGENES.***

Narración visual es una tarea más complicada ya que reconoce varios objetos y relaciones dentro de las imágenes, y aprender las dependencias entre las imágenes.

PROBLEMÁTICA & SOLUCIONES

B

Problemática: Falta Data Indexada

Solución:

Extracción de data de la web referente a la narración visual.

Recopilación de fotos secuenciadas de e-blogs, para hacer un resumen semántico basado en historias.

Uso de un dataset dedicado para la narración visual (VIST). Este contiene historias donde cada una anota un grupo de cinco imágenes con cinco descripciones.

amazon mechanical turk™
Artificial Artificial Intelligence

PROBLEMÁTICA & SOLUCIONES

B

Problemática: ¿Qué enfoque de narración visual usar?

Solución: Comparativa entre enfoques

Basados en la visión

Reconstruyen imágenes o marcos secuenciales de acuerdo con a tramas.

Basados en texto

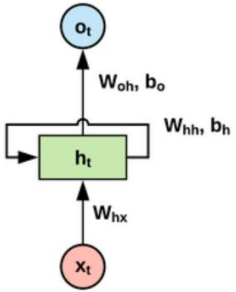
Modelan el lenguaje para generar descripciones similares a historias de las imágenes.

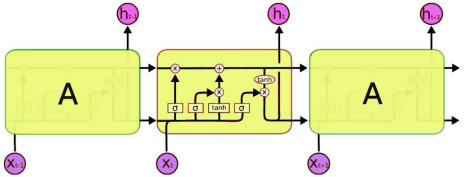
Selección del "*Basado en visión*" por la colección ordenada de imágenes por una RNN de omisión (S-RNN)

PROBLEMÁTICA & SOLUCIONES

B

Problemática: ¿Qué red neuronal para modelar el texto escoger?





Solución:

Análisis de limitaciones del problema.

LSTM jerárquico:

Un LSTM se ocupa de las palabras y sus dependencias, y otro LSTM se ocupa de las oraciones y sus dependencias.

ESTADO DEL ARTE

C

La narración visual realiza la **descripción de una sola imagen, a imágenes secuenciales**. *La técnica se basa en subtitular de imagen/video y la clasificación visual*.

RNN

Traducción automática mediante RNN (Cho, Sutskever, Bahdanau 2015)

CNN

Generación de subtítulos basados en redes CNN y RNN. (Vinyals, Karpathy y Li, Donahue, 2017)

Modelo de codificador - decodificador que con representación visual por CNN y representación textual por RNN. (Vinyals, Simonyan y Zisserman, 2015)

Alineación de regiones de imagenes con texto mediante CNN y RNN bidireccional. (Karpathy y Li 2015)

LSTM

Modelo de lenguajes mediante RNN multimodal, CNN y LSTM para **subtítulos de imágenes**. (Mao 2014)

Generación de subtítulos de videos mediante CNN-RNN o CNN-LSTM. (Venugopalan, Yao, 2015)

ESTADO DEL ARTE

C

*El **resumen o clasificación visual** selecciona principalmente fotogramas o imágenes clave en una secuencia para componer una historia.*

Método de foto-secuenciación

Ordenar temporalmente un conjunto de imágenes fijas de cámaras no calibradas. (Tali 2014)

Ordenamiento temporal de las imágenes

Encontrar correspondencias entre múltiples imágenes de la misma escena por enfoques geométricos. (Basha, Pickup, 2014)

Uso de características de flujo óptico denso y un algoritmo de coincidencia de parches

Definir métricas sobre la dinámica y la coherencia de la escena. (Choi 2016)

Fotogramas muestreados de un video

Aprenden una incrustación temporal de fotogramas de video en eventos complejos. (Ramanathan, 2015)

Características basadas en texto e imágenes y predicciones unarias y por pares

Ordenar un conjunto desordenado de pares de imágenes y subtítulos alineados en una secuencia que forma una historia coherente. (Agrawal, 2016)

ESTADO DEL ARTE

C

Narración visual *toma como entrada una secuencia de imágenes y las describe una historia coherente en texto.* ***Busca producir una metrica automatica para calificar historia como un juez humano que consideraría una buena historia****.*

Puede realizarse de las siguientes formas:

Aprendizaje profundo directo sin intermedios

Arquitectura multimodal CRCN (CNN, RNN - Bi Direccionales y un modelo de coherencia local basado en entidades).

Entrada: Imágenes y frases de la historia

Salida: Puntuación para la compatibilidad entre el flujo de imágenes y la historia.

Explotación de estructuras o datos intermedios

Arquitectura de codificador - decodificador.

Codificar imágenes y leyendas de texto asociadas mediante codificadores separados.

Combina, antes de decodificarlos en las oraciones de la historia.

Objetivo: Aprender automáticamente a mapear desde secuencias de imágenes hasta historias de salida.

Aprendizaje reforzado

Modelos de codificador - decodificador

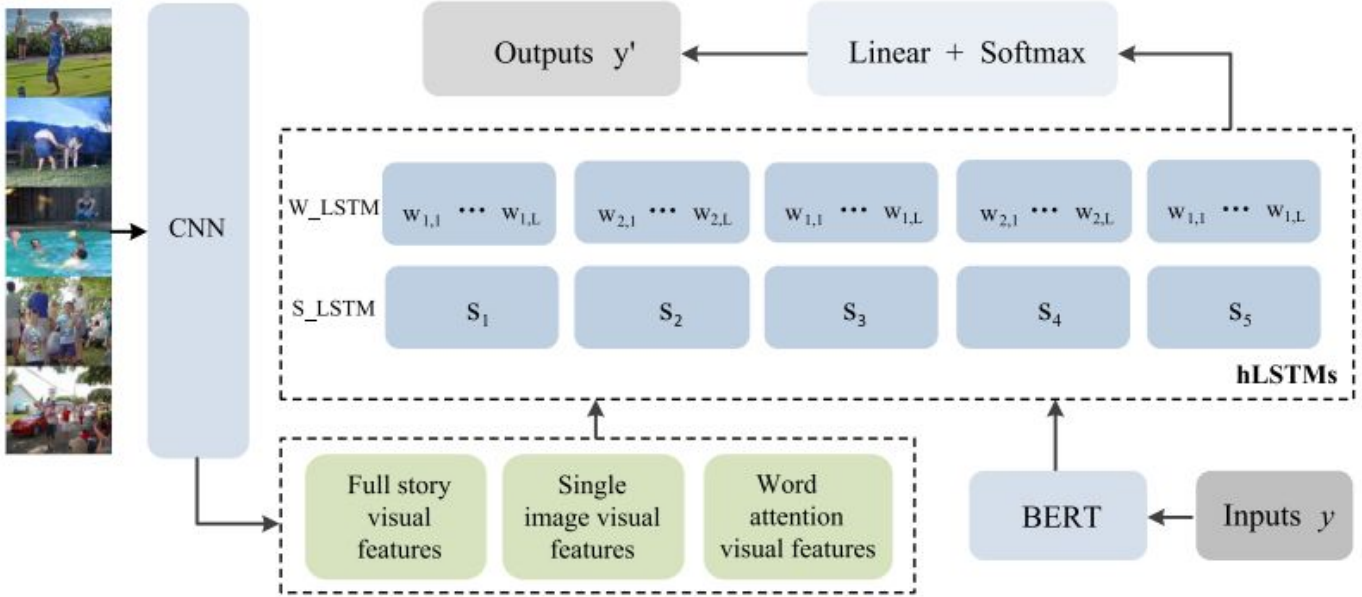
Optimizan mediante aprendizaje por refuerzo.

Ventaja: Más flexibilidad en cómo se puede definir cada imagen relacionado a un texto un objetivo.

METODOLOGÍA

Marco de narración visual llamado BERT-hLSTM

Combina BERT y LSTM jerárquico para generar automáticamente descripciones similares a historias de imágenes secuenciales.



Extracción de características visuales

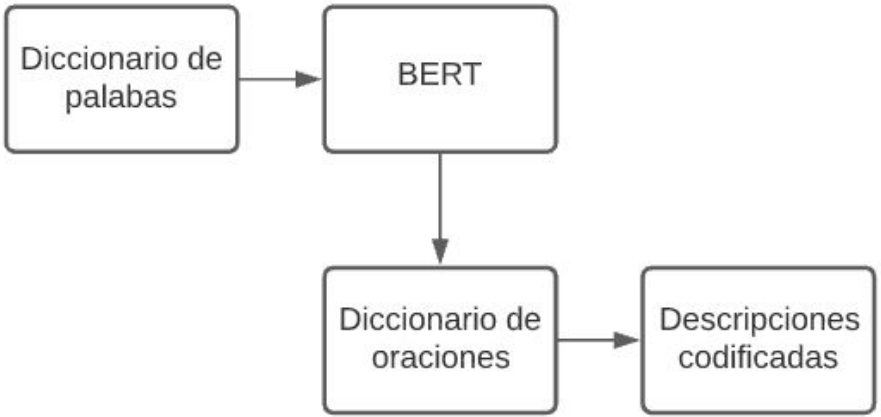
Full-story visual features: Como salida de la capa convolucional se obtienen las features finales de las N imágenes de entrada. Este resultado es la entrada las capas ocultas para S_LSTM y W_LSTM.

Single-image visual features: las features de cada imágenes son usadas a nivel de oraciones, a través de S_LSTM

Word-attention visual features: calculadas a partir de las single-image visual features mediante mecanismos con atención.

Text embedding

A las descripciones se las representa en sentence-embeddings y word-embeddings a través del modelo BERT.



Generador de la historia

El modelo propuesto utiliza la red LSTM con mecanismos de atención, en donde la salida en cada paso es condicionada por el contexto semántico actual y el estado oculto generado anteriormente. El estado oculto inicial se obtiene de la transformación de las full-story visual features.

Para la capa W_LSTM se integra información de contexto a nivel de oraciones y a nivel de palabras para predecir las palabras de cada oración en orden.

D

Generador de la historia

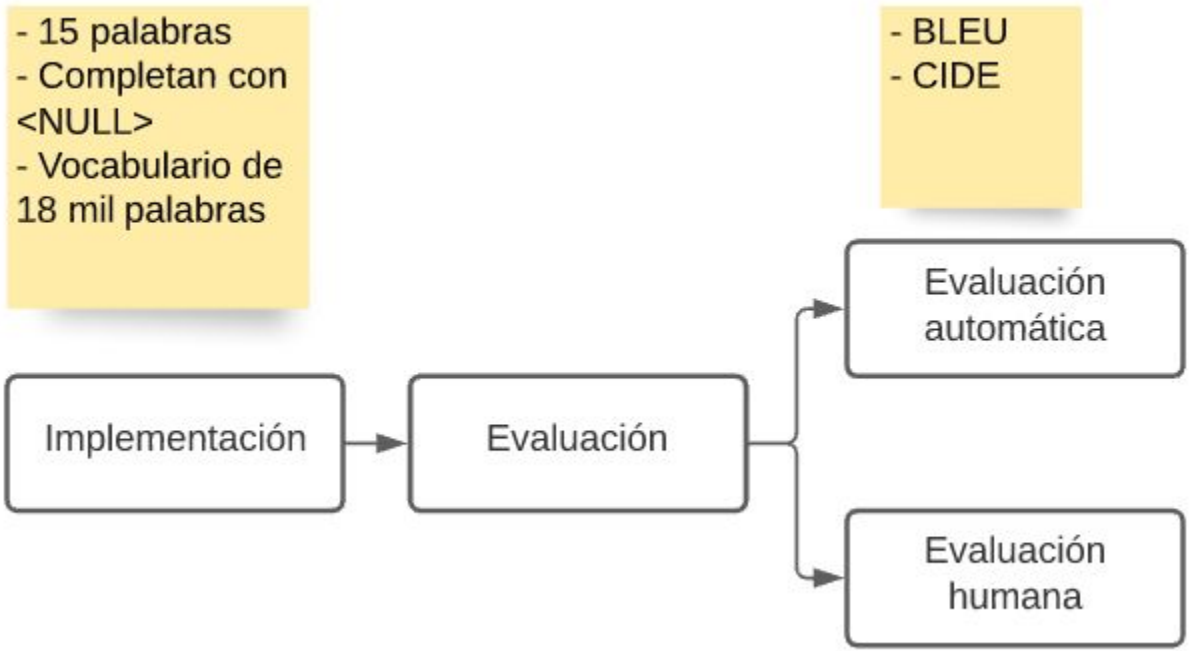
El objetivo del framework es generar descripciones coherentes para una secuencia de imágenes dada. Para esto maximizan la probabilidad de las descripciones generadas dadas características visuales como entrada. Esta probabilidad es dada por la suma de las probabilidades de las oraciones obtenidas siendo cada oración la probabilidad conjunta de una secuencia de palabras.

Experimentos

Dataset: VIST

1	2	3
		
The dog was ready to go.	He had a great time on the hike.	And was very happy to be in the field.

Experimentos



D

Evaluación automática

Method	BLEU	CIDEr
enc-dec (variant of Vinyals et al., 2015 extended for image sequences)	19.58	4.65
enc-attn-dec (variant of Xu et al., 2015 extended for image sequences)	19.73	4.96
h-attn (Yu et al., 2017)	20.53	6.84
h-attn-rank (Yu et al., 2017)	20.78	7.38
h-(gd)attn-rank (Yu et al., 2017)	21.02	7.51
AREL (Wang et al., 2018)	**23.02**	**9.4**
HP (Wang et al., 2019)	21.31	7.44
HPS (Wang et al., 2019)	21.39	7.75
HPR (Wang et al., 2019)	21.39	7.61
HPSR (Wang et al., 2019)	21.51	8.03
hLSTMs (Ours)	21.67	7.98
BERT-hLSTMs (Ours)	23.00	8.37

Evaluación humana

- El criterio de evaluación contiene los siguientes aspectos: relevancia, coherencia y expresividad.

- Se muestrea al azar 150 elementos de los datos de prueba

- Tres anotadores para realizar una comparación por pares y elegir el mejor de las dos historias según los tres criterios

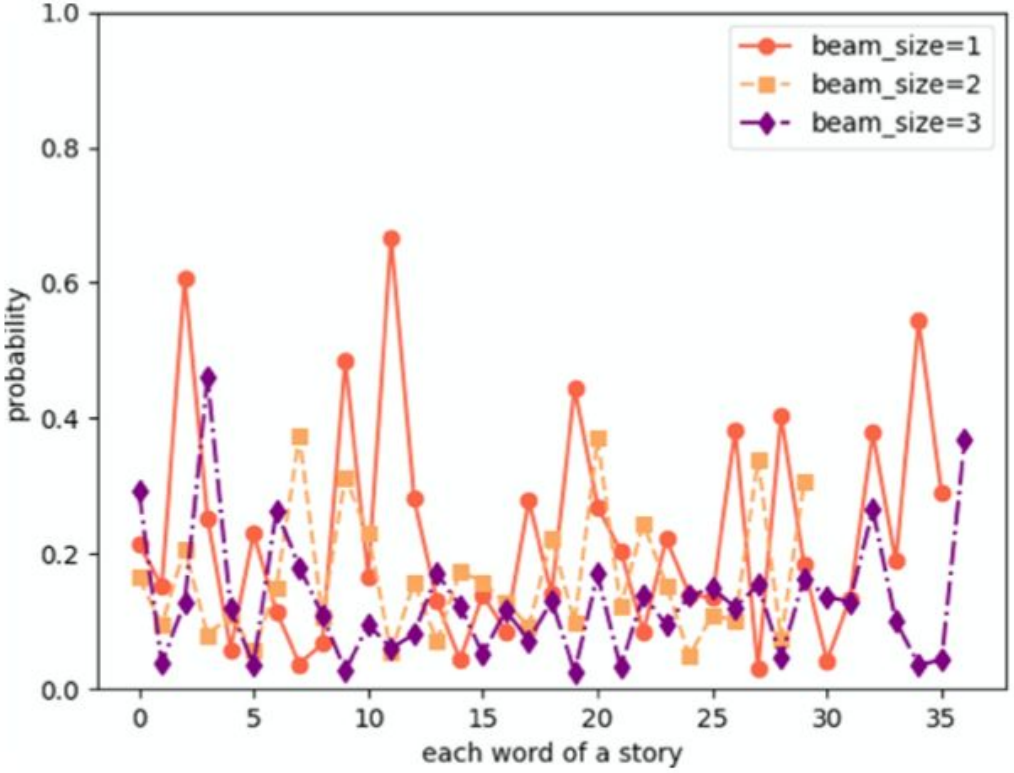
RESULTADOS

Method	label	the nearest 5 points in the original space
hLSTMs	i had a great time there.	i had a great time on vacation last weekend.
		i went to the beach last weekend.
		it was a lot of fun.
		i had a great time.
		i had a great time yesterday.
	we took a lot of pictures.	we had a great time.
		there was a lot of people there.
		everybody was very happy.
		everybody was having a great time.
		the big event.
	the wedding was beautiful.	i went to the meeting yesterday.
		i went for a walk last week.
		i went down to the beach last weekend.
		i went to the fair last weekend.
		the scenery was beautiful.
BERT-hLSTMs	i had a great time there.	i had a great time.
		i had a great time with them.
		i had a great time on vacation.
		i had a great time at the beach.
		i had a great time at the party.
	we took a lot of pictures.	we took lots of pictures.
		we took a lot of pictures together.
		we took a lot of pictures there.
		we took lots of photos.
		we had a lot of fun.
	the wedding was beautiful.	the ceremony was beautiful.
		it was a lovely ceremony.
		the concert was amazing.
		the show was great.
		it was a beautiful event.

RESULTADOS

Method	label	the nearest 5 points in the original space
hLSTMs	woman	man, boy, girl, brother, guy
	coffee	married, wedding, guitar, flower, camp
	enjoyed	enjoying, enjoy, carried, liked, joined
	were	are, 're, 'm, be, am
	weekend	week, distance, afternoon, yesterday, destination
BERT-hLSTMs	woman	person, man, lady, wife, mother
	coffee	chocolate, cream, meat, beer, candy
	enjoyed	enjoying, enjoy, liked, loved, worth
	were	are, was, is, be, been
	weekend	afternoon, evening, morning, summer, week

RESULTADOS


beam_size=1
beam_size=2
beam_size=3
probability
each word of a story
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35

RESULTADOS

Enc-att-dec: the family were very happy. the man was very excited. the man was very happy. the man was very happy. we had a great time.

BERT-hLSTMs: the man was very excited to the **beach**. and it had a lot of fun. the view of the **trees** were very beautiful. the view of the beach was a beautiful. **and** the view of the **water**.

Reference: here we are on the first day of our trip to the beach. we were so excited that we both had to take pictures. we took a short break from the beach, but we got lost. however, we found more beach and it was more peaceful. we finally got in the water after a while.

RESULTADOS

	enc-att-dec	BERT-hLSTMs	Tie	p-value	AREL	BERT-hLSTMs	Tie	p-value
Relevance	27.3%	63.2%	9.5%	.0023	43.3%	46.7%	10.0%	.37
Coherence	24.7%	66.6%	8.7%	.0002	40.2%	46.1%	13.7%	.12
Expressiveness	19.4%	72.6%	8.0%	.0011	38.9%	50.5%	10.6%	.025

CONCLUSIONES

- El método propuesto combina información semántica a nivel de oración y de palabra utilizando BERT-hLSTM.
- Estructura de red sencilla.
- Tiene menos parámetros de entrenamiento y fusiona la información semántica a nivel de oración y de palabra.
- Aprende las relaciones entre oraciones y generar descripciones más coherentes.
- Eficacia del uso de la incrustación BERT y las hLSTM para oraciones y palabras.

REFERENCIAS BIBLIOGRÁFICAS

1. Jing Sua, Qingyun Daia, Frank Guerinc, Mian Zhoud (2020) BERT-hLSTMs: BERT and hierarchical LSTMs for visual storytelling. *Intelligent System Design, Springer*, 505-519. Available on https://www.sciencedirect.com/science/article/abs/pii/S0885230820301029

PREGUNTAS