Estudio y aplicación de redes neuronales a predicción de hashtags en dominios cruzados

Grado en ingeniería informática

Trabajo Fin de Grado

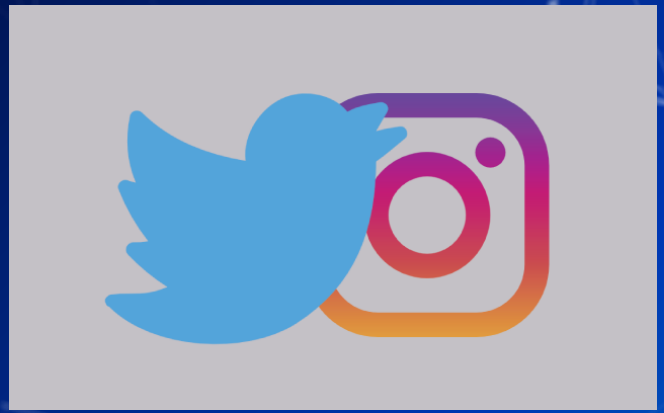
Ricardo Sánchez-Guzmán Hitti

Índice

Índice

Introducción

Redes sociales



Introducción

Elementos comunes

Introducción

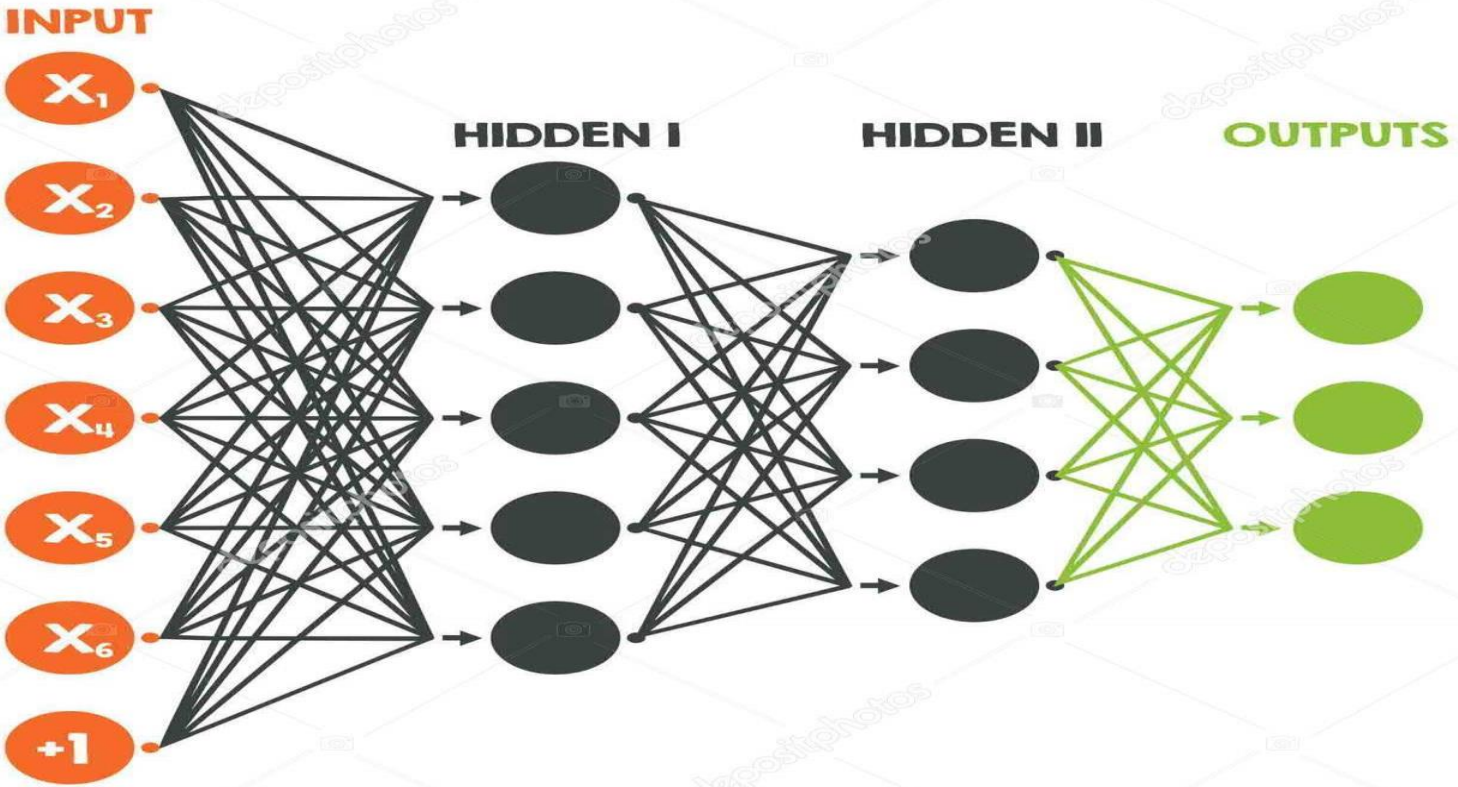
Propuesta

Aplicar algoritmos de aprendizaje automático en la predicción de hashtags utilizando un dataset de dominios cruzados.

Índice

¿Qué son las redes neuronales?

ARTIFICIAL NEURAL NETWORK



Estado del arte

Redes RNN (**Recurrent neural network**)

Su nombre es debido a que las neuronas de la red se retroalimentan.

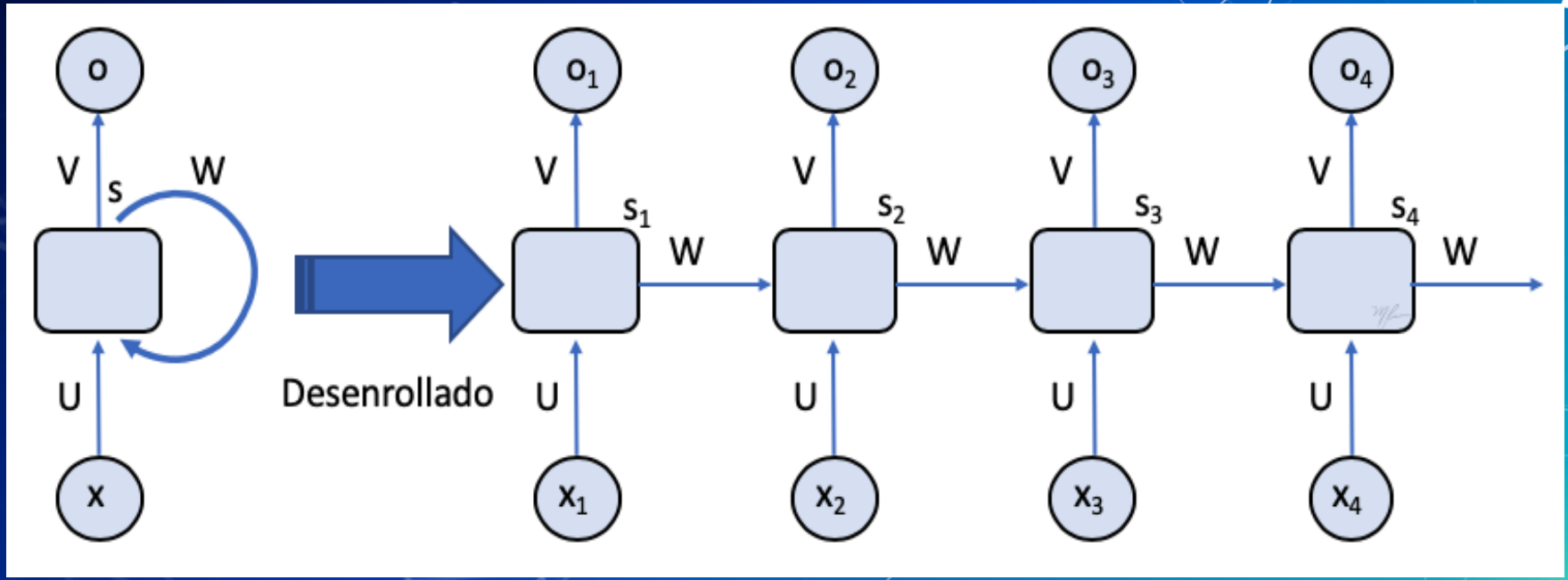
Gracias a esta **retroalimentación** la red es capaz de tener una memoria a corto plazo.

Todas las capas comparten los mismo parámetros de entrada algo que las diferencia del resto de redes.

Los pesos de la red se entrenan con los valores de los errores actuales y pasados.

Su principal problema es que no es capaz de soportar una memoria a largo plazo (Problema del **desvanecimiento del gradiente**).

Arquitectura de una red RNN



Estado del arte

Redes LSTM (**Long short-term memory**)

Cambia el concepto de neurona por **celda de memoria**.

La celda de memoria almacena el valor que queremos recordar, este valor puede estar durante un periodo corto o largo dependiendo de las entradas, lo que permite recordar el valor mas importante y no el ultimo calculado.

Surgen debido al problema de la memoria a largo plazo con las RNN.

La celda de memoria posee tres puertas: **puerta de entrada, puerta de salida y puerta del olvido**.

Multitud de aplicaciones: predicción de series temporales, reconocimiento de voz, composición musical.

Prediccion de hashtags y subtitulado de imágenes

Se trata...

Estado del arte

Métodos en el subtitulado de imágenes

- Se detectan los objetos o atributos más relevantes de la imagen.
- Se combinan las descripciones de las imágenes más semejantes.
- Se utiliza la secuencia de palabras.

Estado del arte

Secuencia de palabras

Necesitamos un modelo capaz de almacenar una secuencia de palabras, el modelo mas utilizado suelen ser el de las redes **LSTM**.

En segunda lugar necesitamos miles de imágenes con una descripción, es decir, utilizaremos un **aprendizaje supervisado**.

Cuando se genera una nueva descripción se utilizan todas las palabras generadas previamente para formular la siguiente palabra, esto permite usar combinaciones de palabras que nunca se podrían haber generado en los datos de entrenamiento, las palabras que normalmente se utilizan son las generadas en un vocabulario el cual se crea a partir de las descripciones que las imágenes tienen hechas por los usuarios.

Estado del arte

Métricas para la evaluación de subtítulos

- BLEUr
- METEOR
- ROGUE-L
- CIDEr

	Description	BLEU	METEOR	ROUGE	CIDEr
original	a man wearing a red life jacket is sitting in a canoe on a lake	1	1	1	10
candidate	a man wearing a life jacket is in a small boat on a lake	0.45	0.28	0.67	2.19
synonyms	a **guy** wearing a life **vest** is in a small boat on a lake	**0.20**	**0.17**	0.57	**0.65**
redundancy	a man wearing a life jacket is in a small boat on a lake **at sunset**	0.45	0.28	0.66	2.01
word order	**in a small boat on a lake** a man is wearing a life jacket	**0.26**	0.26	**0.38**	**1.32**

Estado del arte

Métricas para la evaluación de hashtags

- Recall

$$Recall = \frac{\#TP}{\#TP + \#FN}$$

- Precisión

$$Precision = \frac{\#TP}{\#TP + \#FP}$$

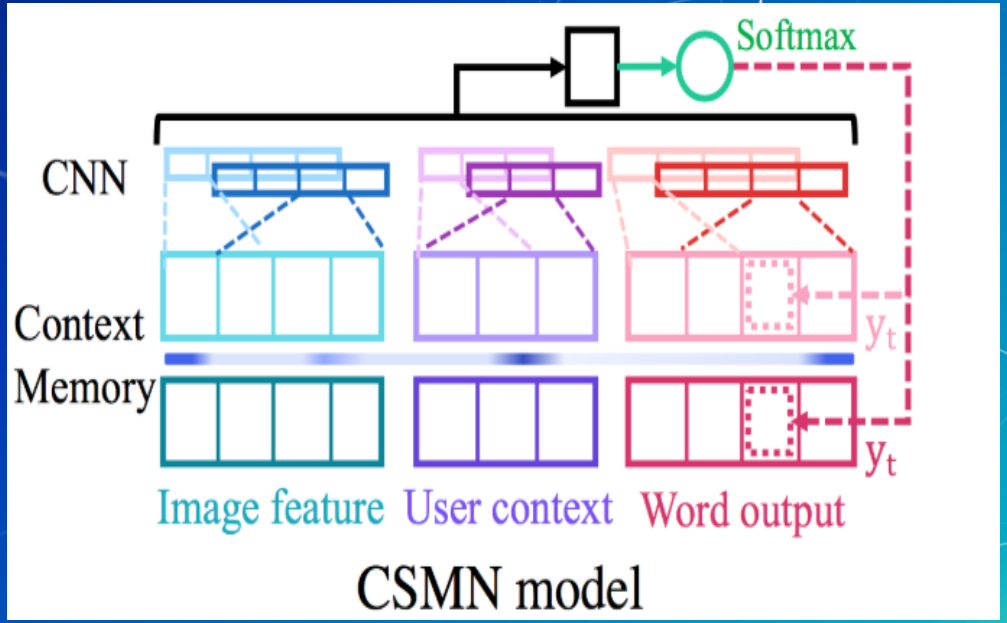
- F1-Score

$$F1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

Índice

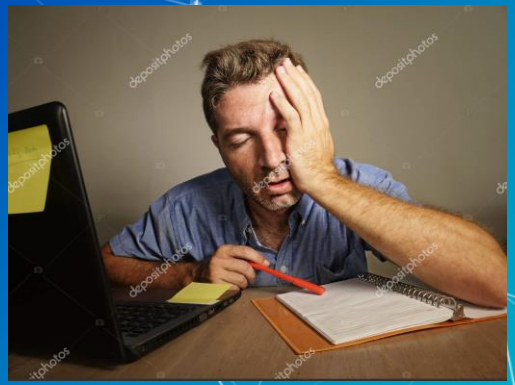
Librería attend2u

- Red Neuronal
- Dataset de Instagram



Modificaciones de la red

- Debido a la cantidad de requisitos hardware que piden las redes neuronales que trabajan con imágenes se modificó la red para reducir el numero de épocas y el uso de memoria de la GPU.
- Esto hizo viable el trabajo con la red y permitió no reducir la cantidad de datos de entrenamiento.



Extracción, preprocesado y almacenamiento de los datos

- Extracción de los datos de Twitter mediante la librería Tweepy.
- Preprocesado de los datos con Scikit-learn.
- Almacenamiento de los mismos en ficheros JSON.

+2.000.000 Hashtags

+1.000.000 Tweets

+85.000 Imágenes

+1.000 Usuarios válidos

Algoritmos KNN

- KNN basado en las características de la imagen.
- KNN basado en una imagen aleatoria del usuario más cercano.
- KNN basado en la imagen más cercana del usuario más cercano.

Índice

Comparativa Subtítulos: Dataset de Instagram

Métodos	B-1	B-2	B-3	B-4	Meteor	CIDEr	ROGUE-L
1 NN-Im	0.071	0.020	0.007	0.004	0.032	0.059	0.069
1 NN-Usr	0.063	0.014	0.002	0.000	0.028	0.025	0.059
1 NN-UsrIm	0.106	0.032	0.011	0.005	0.046	0.084	0.104
(CSMN-W60-P5)	**0.171**	**0.068**	**0.029**	**0.013**	**0.064**	**0.214**	**0.177**
(CSMN-W20-P5)	0.116	0.041	0.018	0.007	0.044	0.119	0.123

ORIGINAL

Métodos	B-1	B-2	B-3	B-4	Meteor	CIDEr	ROGUE-L
1 NN-Im	0.065	0.018	0.009	0.007	0.030	**0.064**	0.066
1 NN-Usr	0.073	0.020	0.004	0.001	0.028	0.033	0.067
1 NN-UsrIm	0.088	0.024	0.000	0.003	0.040	0.056	0.072
(CSMN-W60-P5)	0.087	0.032	0.014	**0.008**	0.114	0.036	0.109
(CSMN-W20-P5)	**0.095**	**0.035**	**0.015**	0.007	**0.115**	0.040	**0.112**

MODIFICADO

Comparativa

Ejemplos subtítulos

(Objetivo) pool pass for the summer
(CSMN-attend2u) the pool was absolutely perfec
(CSMN-adaptada) another day at the beach

(Objetivo) this speaks to me literarily
(CSMN-attend2u) I love this #quote
(CSMN-adaptada) I just write a caption

(Objetivo) air is the fall
(CSMN-attend2u) fall is in the air
(CSMN-adaptada) fall colors

Comparativa Hashtags : Dataset de Instagram

Métodos	F1-SCORE
1 NN-Im	0.049
1 NN-Usr	0.054
1 NN-UsrIm	0.109
(CSMN-W60-P5)	**0.230**
(CSMN-W20-P5)	0.147

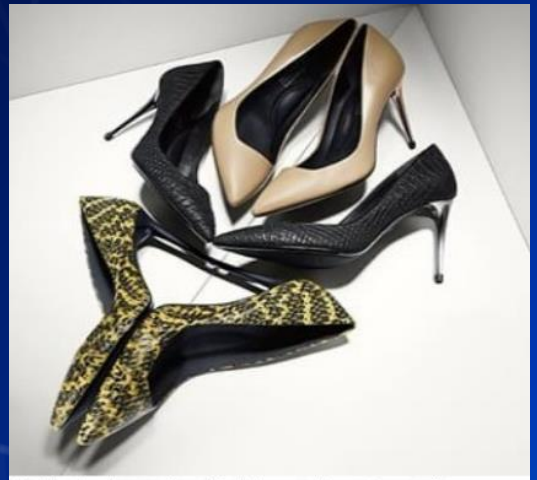
ORIGINAL

Métodos	F1-SCORE
1 NN-Im	0.0379
1 NN-Usr	0.1345
1 NN-UsrIm	0.1635
(CSMN-W60-P5)	0.1745
(CSMN-W20-P5)	**0.1979**

MODIFICADO

Comparativa

Ejemplos hashtags



(Objetivo) #style #fashion #shopping #shoes #kennethcole...
(CSMN-attend2u) #newclothes #fashion #shoes #brogues
(CSMN-adaptada) #ankleboots #sneakers #shoes

(Objetivo) #boudoir #heartprint #love #weddings #potterybarn
(CSMN-attend2u) #decor #homedecor #interior #interiordesign #home #bride
(CSMN-adaptada) #decor #white #home

(Objetivo) #greensmoothi #dairtyfree #lifewithatoddler #glutenfree #vegetarian
(CSMN-attend2u) #greensmoothie #greenjuice #smoothie #vegan #raw #juicing #eatclean #detox #cleanse
(CSMN-adaptada) #kiwi #smoothie #protein

Pruebas

Descripción de las pruebas

- Los datos de entrenamiento serán las imágenes y el vocabulario activo de los usuarios en Instagram
- Los datos de test serán esos mismos usuarios pero en un dominio distinto (Twitter).
- Existen dos tipos de pruebas: Un único usuario de un tipo y un conjunto de usuarios de un mismo tipo.
- Todas las pruebas realizadas con conjuntos de usuarios tendrán un tamaño de 4000 imágenes.
- Una última prueba con un conjunto aleatorio de usuarios (5000 imágenes).

Resultados: Dataset de Instagram y Twitter

Conjunto pequeño de imágenes

Métodos	F1-SCORE
1 NN-Im	0.006
1 NN-Usr	0.000
1 NN-UsrIm	0.000
(CSMN-W60-P5)	0.303
(CSMN-W40-P5)	0.301
(CSMN-W20-P5)	**0.313**

Usuario con **69** imágenes.

Métodos	F1-SCORE
1 NN-Im	0.054
1 NN-Usr	0.073
1 NN-UsrIm	0.085
(CSMN-W60-P5)	**0.162**
(CSMN-W40-P5)	0.157
(CSMN-W20-P5)	**0.162**

Usuarios con un número de imágenes comprendido entre **1** y **169**.

Resultados: Dataset de Instagram y Twitter

Conjunto intermedio de imágenes

Métodos	F1-SCORE
1 NN-Im	0.000
1 NN-Usr	0.000
1 NN-UsrIm	0.000
(CSMN-W60-P5)	0.398
(CSMN-W40-P5)	0.420
(CSMN-W20-P5)	**0.457**

Usuario con **286** imágenes.

Métodos	F1-SCORE
1 NN-Im	0.079
1 NN-Usr	0.140
1 NN-UsrIm	0.152
(CSMN-W60-P5)	0.173
(CSMN-W40-P5)	0.192
(CSMN-W20-P5)	**0.207**

Usuarios con un número de imágenes comprendido entre **1** y **169**.

Resultados: Dataset de Instagram y Twitter

Conjunto grande de imágenes

Métodos	F1-SCORE
1 NN-Im	0.030
1 NN-Usr	0.068
1 NN-UsrIm	0.074
(CSMN-W60-P5)	0.086
(CSMN-W40-P5)	0.127
(CSMN-W20-P5)	**0.150**

Usuario con **919** imágenes.

Métodos	F1-SCORE
1 NN-Im	0.082
1 NN-Usr	0.085
1 NN-UsrIm	0.096
(CSMN-W60-P5)	0.141
(CSMN-W40-P5)	0.139
(CSMN-W20-P5)	**0.145**

Usuarios con un número de imágenes mayor de **350**.

Resultados: Dataset de Instagram y Twitter

Conjunto aleatorio de imágenes

Métodos	F1-SCORE
1 NN-Im	0.057
1 NN-Usr	0.111
1 NN-UsrIm	0.123
(CSMN-W60-P5)	0.179
(CSMN-W40-P5)	**0.191**
(CSMN-W20-P5)	0.188

Conjunto aleatorio de **5000** imágenes.

Discusión de los resultados

- La tendencia en los algoritmos KNN se ha respetado en todas las pruebas siendo el peor de ellos KNN-image y el mejor KNN-user-image.
- KNN es una buena alternativa para ahorrar tiempos.
- Las redes neuronales superan a los algoritmos de vecinos próximos.
- Vocabulario activo de un usuario semejante en ambos dominios.

Resultados

Ejemplo de predicción de hashtags

(Objetivo) #memorialday #rememberourheros #america #usa
(CSMN-W20-p5) #classof2015 #america #usa #memorialday

(Objetivo) #cat #catnap #catagram #catsofinstagram
(CSMN-W20-P5) #catagram #catsofinstagram #tuxedocat #caturday #cat

(Objetivo) #rafting #myjasper #jaspernp #wedareyou #explorea
#explorejasper #splash
(CSMN-W20-P5) #whitewater #myjasper #explorealberta #explorejasper #traveltuesday #wedareyou #splash #jasper #explorecanda #rafting

Índice

Conclusiones y trabajo futuro

Conclusiones

- Se han utilizado conocimientos adquiridos en la asignaturas de NEURO y BMINF.
- Estudio de papers.
- Trabajo con librerías externas.
- Continuar trabajando en el TFM.

Trabajo futuro

- Probar con otro tipo de datos como: música, videos y texto.
- Probar con otro tipo de dominios para comprobar si existe está misma correlación.
- Utilizar la información de un usuario en distintos dominios para solucionar el problema del arranque en frío (cold start).

Repositorio: https://bitbucket.org/Rsanchezguzman/crosshashtags/src/master/

MUCHAS GRACIAS!!