Towards Multilingual Image Captioning Models that Can Read

Rafael Gallardo Beatriz Beltrán Carlos Hernández Darnes Vilariño

Language & Knowledge Engineering Lab, Benemérita Universidad Autónoma de Puebla

September 7, 2021

Table of contents

Introduction
What is image captioning?
Image captioning models that can read
Why is a multilingual approach important?
Data and methods
TextCaps dataset
Methods
Results
Conclusions

What is image captioning?

- The problem of generating a textual description of a given image is called image captioning.
- To solve this problem, models should be able to recognize objects, attributes and relationships among the actors in the scene.

BUAP

Image captioning models that can read?

- Frequently, it is critical for humans to read associated text and comprehend it in the context of the visual scene.
- What if current models could read and integrate the read text in the generated descriptions?

BUAP

Image captioning models that can read?

- To solve this problem, new architectures should be designed.
- The models should be able to determine relationships between the read tokens (OCR), the visual context and be able to switch between the read tokens and the model's vocabulary.

Available data: TextCaps dataset

a

the numbers 18 and 17 on a scoreboard
the number 17 is on the scoreboard with the word rice on it
The scoreboard of a football game showes that **Rice is winning.**
The word "RICE" is displayed on the scoreboard.
A score board shows Rice with 18 points vs. ECU with 17 points.

b

the price of 17.88 that is above a lady
A Walmart sign that says Rollback $17.88 is above a shelf of weight loss products.
A display at Walmart **for a special price** on Hydroxycut.
Box of Hydroxycut **on sale** for only 17.88 at a store.
walmart has hydroxycut **for sale** for 17.88 **instead of** 19.88

c

A white Samsung smartphone shows the time is 11:19.
top part of samsung phone at 11:19 on December 30
A close up of the top half of a Samsung cell phone.
A samsung brand phone shows the current time is 11:19.
The top half of a Samsung cellphone showing the time, date and weather conditions.

Figure: Some samples from the TextCaps dataset[1].

[1]Sidorov, Oleksii, et al. "Textcaps: a dataset for image captioning with reading comprehension." European Conference on Computer Vision. Springer, Cham, 2020.

Image captioning models that can read: State-of-the-art approaches

Table: Performance for image captioning with RC methods available in literature. The gray row indicates the baseline proposed by the authors of the TextCaps dataset, while bold numbers indicate the best scores. Metrics in columns: BLEU-4 (B-4), METEOR (M), ROUGE L (R), SPICE (S), CIDEr (C).

	TextCaps validation set metrics				
Method	B-4	M	R	S	C
M4C-Captioner	23.3	22.0	46.2	15.6	89.60
MMA-SR	24.6	23.0	47.3	16.2	98.00
TAP	**25.8**	**23.8**	**47.9**	**17.1**	**109.2**
SBD	24.8	22.7	47.24	15.71	98.83
CNMT	24.8	23.0	47.1	16.3	101.7
AnC	24.7	22.5	47.1	15.9	95.50

BUAP

Do these models currently work with non-English languages?

NO:

- The TextCaps dataset and its subsets (train, evaluation and test) are fully annotated in English.
- There is no publicly available alternative for non-English languages.
- All SOTA use English-specific components.

Contributions of our work

- We present the first bilingual approach to create image captioning models that can read.
- The first Spanish version of TextCaps is generated by developing a neural-based translation pipeline.
- Our architecture design can be extended to more languages.

Size of data

Table: Summary of the number of samples in the TextCaps dataset.

Number of samples	Training	Evaluation	Testing
Images	21,953	3,166	3,289
Captions	109,756	15,830	16,445

Automatic translation

- Automatically translated with HugginFace[2] transformers.
- Architecture name: Denoising autoencoder for pretraining sequence-to-sequence models.
- Pre-trained model name: Helsinki-NLP/opus-mt-en-es
- BLEU score of the pre-trained model on the Tatoeba Translation Challenge: 54.9

[2] https://huggingface.co/

Base architecture for the TextCaps challenge: M4C Captioner

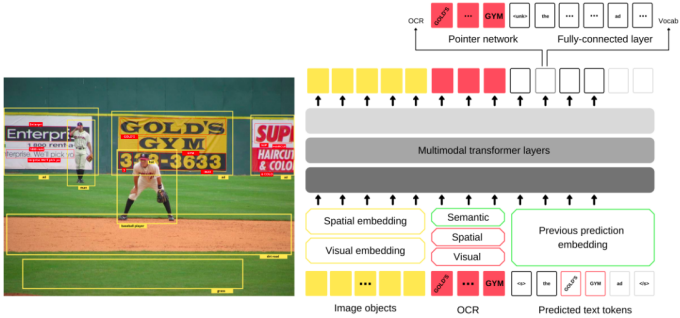

Figure: M4C-Captioner architecture. The yellow blocks and highlights indicate features that correspond to objects in the image, the red blocks correspond to textual and OCR features in the image. The green blocks are the modules that were originally developed to work with the English language.

How to make it multilingual?

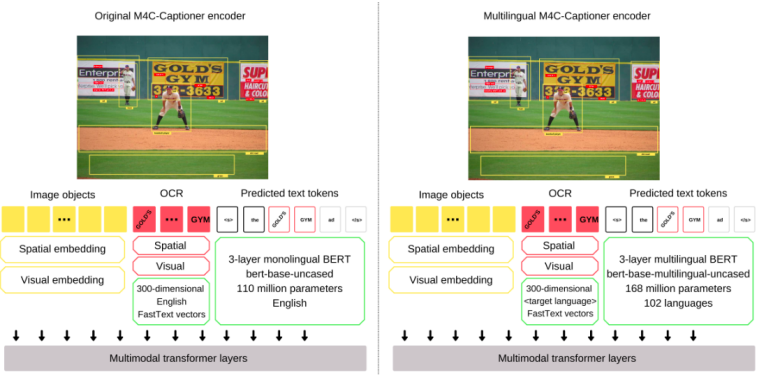

Figure: A comparison of both encoders: the original M4C's encoder on the left and the ML M4C's encoder on the right. Modules highlighted in green indicates the principal difference between the two encoders.

Experimental configurations

Table: All trained and evaluated architectures. Full configuration files and logs are available on the repository of the paper.

Model	TextCaps language	FastText	Text BERT	BERT vocab size	M4C vocab size	Total parameters
m4c-captioner-zoo (Baseline)	English	English: wiki.en.bin	bert-base-uncased	30522	6736	92,185,168
m4c-captioner-local	English	English: wiki.en.bin	bert-base-uncased	30522	6736	92,185,168
en_ml-m4c-captioner	English	English: wiki.en.bin	bert-base-multilingual-uncased	105879	6736	150,059,344
es_ml-m4c-captioner	Spanish	Spanish: cc.es.300.bin	bert-base-multilingual-uncased	105879	7207	150,421,543

Summary of results

Table: Performance of each model over the validation set of TextCaps, both English and Spanish sets, are included. The best results are highlighted with bold numbers (not applicable for Spanish since there is just one model). Metrics in columns: BLEU-4 (B-4), METEOR (M), ROUGE L (R), SPICE (S), CIDEr (C).

			English TextCaps validation set metrics				
Model	FastText	Text BERT	B-4	M	R	S	C
TAP	English: wiki.en.bin	bert-base-uncased	**25.8**	**23.8**	**47.9**	**17.1**	**109.2**
m4c-captioner-zoo (Baseline)	English: wiki.en.bin	bert-base-uncased	23.4	21.8	46.0	15.0	89.1
m4c-captioner-local	English: wiki.en.bin	bert-base-uncased	23.1	22.3	46.1	15.7	90.4
en-ml-m4c-captioner	English: wiki.en.bin	bert-base-multilingual-uncased	22.4	22.2	46.0	15.6	88.7

			Spanish TextCaps validation set metrics				
Model	FastText	Text BERT	B-4	M	R	S	C
es-ml-m4c-captioner	Spanish: cc.es.300.bin	bert-base-multilingual-uncased	21.0	21.6	41.6	6.1	63.2

Sample captions generated by our design

Human: A banner for the Igreja Adventista Do 7 Dia is hung on a balcony railing.

English model: a sign that says igreja adventista do do do dia.

Spanish model: una señal que dice que igreja adventista está en una pared de ladrillo.

Human: A blue Intel Pentium inside box sitting on a white table

English model: a blue box with the word desktop on it.

Spanish model: una caja azul con la palabra pentium en ella.

Human: One of the jets parked show the letters AF and number 711 on the tail.

English model: a small plane with the number 711 on the tail.

Spanish model: un avión con el número 711 en la cola.

Figure: Sample captions generated by our Multilingual M4C architecture.

Conclusions

- Our proposal achieved near state-of-the-art performance while being suitable to work with different languages.
- The generated captions inherit the errors and biases of the translation model, this can be solved with better translation methods or by annotating the dataset by hand.
- Both models (English and Spanish) kept their ability to read and integrate the read text.
- We hop this work can set a baseline for multilingual approaches to this problem.

Thanks!