Arabic Image Captioning using Pre-training of Deep Bidirectional Transformers

Jonathan Emami
Student
Lund University
jontooy@gmail.com

Pierre Nugues
Professor of Computer Science
Lund University
pierre.nugues@cs.lth.se

Ashraf Elnagar
Professor of Computer Science
University of Sharjah
ashraf@sharjah.ac.ae

Imad Afyouni
Assistant Professor
University of Sharjah
iafyouni@sharjah.ac.ae

What is Image Captioning?

- The process of automatically generating a textual description of an image
- Figure 1 shows a picture of the University of Sharjah Campus and a machine-generated caption
- Wide range of applications:
  - Effective image search
  - Auto archiving
  - Helping visually impaired people to see
- A lot of recent development in English image captioning
- Arabic image captioning is lagging behind!

Figure 1: a large building with a park in front of it (machine generated caption)

Main Contributions

- We evaluate transformer-based Arabic image captioning and compare our results to previous ones
- One of our best performing models scored 0.39, 0.25, 0.15 and 0.092 with BLEU-1,2,3,4 respectively, an improvement over previously published scores on the dataset
- We show that training image captioning models with Arabic captions and English object tags is a working approach

Datasets

Arabic-COCO

- Arabic translated subset of the Microsoft Common Objects in Context (MS COCO) dataset
- 414,113 pre-translated captions over 82,783 training images using the Google Translate API
- Captions are noisy, which is why we did not create a validation and testing set out of Arabic-COCO
- Figure 2 shows an example of an image from the train split

A young boy surfing in low waves
صبي صغير يتزلج على الأمواج المنخفضة
A young boy is standing on a surfboard and riding a wave
صبي صغير يقف على لوح ركوب الأمواج وركوب الأمواج
A surfer rides his surf board on some very small waves
راكب أمواج يركب لوح الأمواج على بعض الأمواج الصغيرة جدًا
A young boy is standing on a surfboard in the water
صبي صغير يقف على لوح تزلج على الماء في الماء
A young boy is standing on a surfboard in the ocean
صبي صغير يقف على لوح ركوب الأمواج في المحيط

Figure 2: Caption annotations in English and Arabic for an image sample from the COCO dataset

Arabic Flickr8k

- Arabic translated subset of the Flickr8k dataset
- Arabic Flickr8k is split into 6,000 train images, 1,000 validation images, and 1,000 test images, all with three Arabic captions each (24,000 captions in total)
- The translation to Arabic was performed by ElJundi et al. (2020) in two steps, first by using the Google Translate API and then by validating captions with professional Arabic translators
- Figure 3 shows an example of an image from the train split

A longhaired man surfing a large wave
رجل طويل الشعر يتزلج موجة كبيرة
A man in black on a surfboard riding a wave
رجل أسود على لوح ركوب الأمواج يركب موجة
A man surfing in the ocean
رجل يمارس رياضة ركوب الأمواج في المحيط

Figure 3: Caption annotations in English and Arabic for an image sample from the Flickr8k dataset

Methodology

- We used a two-step pipeline, as shown in Figure 4:
  - Extract region features and object tags from an image through a convolutional neural network (CNN) encoder
  - Generate a sentence from the region features and object tags through a language model, in our case a pre-trained transformer.
- As a learning method for our image captioning model, we used OSCAR (Li et al., 2020) and to evaluate our results, we used well-establish metrics for IC.
- OSCAR uses object tags detected in images as anchor points to ease the alignment of image region and word embeddings

Figure 4: An overview of our methodology.

How do we Extract Image Features?

- We utilized the object detection model X152-C4 (Zhang et al., 2021) for feature extraction
- Figure 5 shows an example of object detection with the X152-C4 model
- For each detected object, an image region vector is generated, which represents the vector input to the last linear classification layer

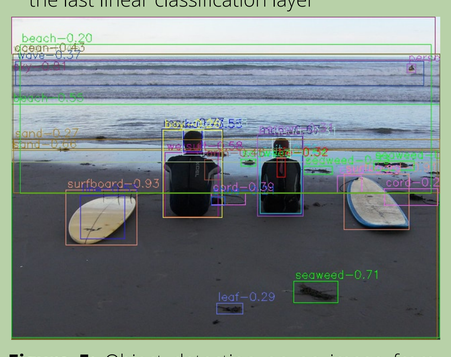


Figure 5: Object detection on an image from the COCO dataset using the X152-C4 architecture. The set of detected object tags are (Arm, Beach, Boy, Cord, Hair, Head, Leaf, Line, Man, Ocean, Person, Sand, Seaweed, Sky, Suit, Surfboard, Tie, Water, Wave, Wetsuit).

How do we Evaluate our Captions?

- We evaluated our captions with 7 different metrics:
  - BLEU-1,2,3,4
  - ROUGE-L
  - METEOR
  - CIDEr
  - SPICE
  - MUSE
  - Human evaluation (THUMB)
- MUSE (Multilingual Universal Sentence Encoder)
  - Embeds texts from 16 languages (including Arabic)
  - Initial intensive computation
  - Captures the semantic meaning of captions
  - Uses angular similarity
- Human Evaluation (THUMB scores)
  - Precision & Recall on a scale 1-5
  - Penalty deductions for incorrect grammar and semantics
  - Overall score is computed by averaging precision and recall and deducting penalty points (maximum 0.5)

BERT: Pre-training of Language Models

- In this work, we used 4 different models as the base for our image captioning models, described below.
- Table 1 shows the different models configurations for comparison.

GigaBERT. GigaBERT (Lan et al., 2020) is a set of models pre-trained as a bilingual BERT and designed specifically for Arabic NLP and Englisht-to-Arabic zero-shot transfer learning. Their best model significantly outperforms mBERT and AraBERT on some supervised and zero-shot transfer settings. The training dataset consists of a dump of Arabic Wikipedia, an Arabic version of OSCAR and the Gigaword corpus, which consists of over 13 million news articles

AraBERT. AraBERT (Antoun et al., 2020) achieved state-of-the-art performance on most tested Arabic NLP tasks. The models were trained on news articles manually scraped from Arabic news websites and several publicly available large Arabic corpora. One of the corpora is named OSCAR (Open Super-large Crawled Aggregated Corpus), not to be confused with the image captioning model OSCAR

ArabicBERT. ArabicBERT (Safaya et al., 2020) was the first pre-trained BERT model for Arabic when it was released. It was originally pre-trained as an approach to solve a sub-task of the Multilingual Offensive Language Identification shared task (OffensEval 2020).

mBERT. mBert, short for Multilingual BERT, was pre-trained with the multilingual Wikipedia dataset that consists of the top 104 most common languages (Devlin et al., 2018), including Arabic

Model	Training Data		Vocabulary			Configuration	
	source	#tokens (all/ar)	tokenization	size (all/ar)	cased	size	#parameters
mBERT	Wiki	21.9B/153M	WordPiece	110k/5k	no	base	172M
AraBERT	Wiki, Oscar, News articles	2.5B/2.5B	SentencePiece	64k/58k	no	base	136M
ArabicBERT	Wiki, Oscar	unknown	WordPiece	32k/28k	no	base	111M
GigaBERT	Wiki, Oscar, Gigaword	10.4B/4.3B	WordPiece	50k/26k	no	base	125M

Table 1: Configuration comparisons for mBert, AraBERT, ArabicBERT, and GigaBERT

Experimental Setup and Evaluation of Captioning Models

English vs. Arabic Labels

- Evaluation of two multilingual models both trained on
  a. Arabic captions and Arabic labels
  b. Arabic captions and English labels
- We carried out this experiment mainly for comparing the object labels ability to affect the final image-text alignment.
- Table 2 shows the final evaluation scores for all models. Our first experiments show that both approaches, training on English and Arabic object labels, work in principle

Model	Labels	BLEU-4	ROUGE-L	METEOR	CIDEr	SPICE
GigaBERT	English	**0.074**	**0.29**	0.3	**0.33**	**0.037**
	Arabic	0.062	**0.29**	**0.31**	0.31	**0.037**
mBert	English	0.058	0.28	0.30	0.29	0.031
	Arabic	0.067	**0.29**	0.30	0.31	0.033

Table 2: Evaluation scores (evaluation on epoch 30) for the trained models. The best scoring models are marked in bold for each evaluation metric.

Learning Curve

- Evaluation of the learning curve for three different models, respectively trained on 50%, 75% and 100% of a dataset. From the results, we can tell if the validation loss decreases with the amount of data or if some adjustment have to be made to the models.
- We evaluated all the models from the learning curve experiment with MUSE to investigate the correlation between semantic scores and an increased amount of data. The evaluation over training time is shown in Figure 6 for AraBERT, ArabicBERT, and GigaBERT.
- In general, more data increased evaluation scores. In the case of AraBERT, the 75% MUSE curve is way lower than the 100% and 50% curves, but the 100% loss curve is still higher than the 50% one. The unstable training results of AraBERT suggest that the selected learning rate is too large.

(a) AraBERT MUSE scores (b) GigaBERT MUSE scores (c) ArabicBERT MUSE scores

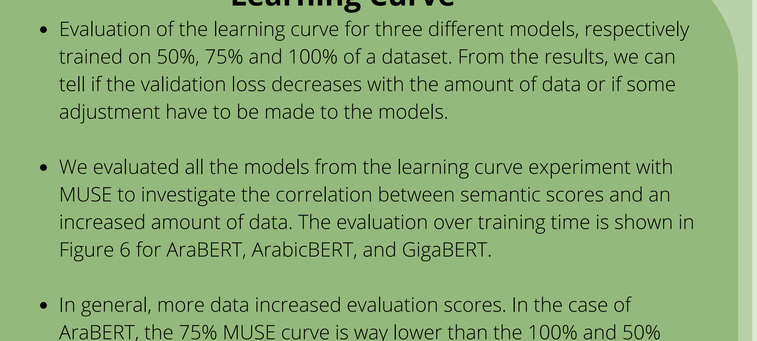
Figure 6: MUSE evaluation scores over all epochs for (a) AraBERT, (b) GigaBERT and (c) ArabicBERT.

Large Scale Training

- From previous experiments, we pick two candidate models. We then perform large scale training on the candidate models on datasets of different sizes
- Table 3 presents the final test scores (BLEU-1,2,3,4, ROUGE-L, METEOR, CIDEr and MUSE) of a selection of our models, and previous Arabic captioning models
- All of our models are named after the scheme [model][batchSize]-[dataset]. For example, one of our best performing models was initialized on AraBERT and trained with a batch size of 32 on Flickr8k. Therefore, we named the model AraBERT32-Flickr8k
- We complemented Table 3 with human evaluations on a sample of the dataset according to the guidelines of THUMB. Figure 7 shows four generated captions from AraBERT32-COCO with images and human evaluations.

Model	Test set	B1	B2	B3	B4	ROUGE-L	METEOR	CIDEr	MUSE
Jindal (2018)	Flickr8k	0.658	0.559	0.404	0.223	-	0.201	-	-
Al-muzaini et al. (2018)	COCO & Flickr8k	0.462	0.260	0.190	0.080	-	-	-	-
Afyouni et al. (2021)	COCO	0.649	0.413	0.241	0.136	0.470	0.408	-	0.78
ElJundi et al. (2020)	Flickr8k	0.332	0.193	0.105	0.057	-	-	-	-
AraBERT32-Flickr8k		**0.391**	**0.246**	0.150	0.092	0.331	0.314	0.415	**0.671**
AraBERT32-COCO		0.365	0.221	0.129	0.0715	0.310	**0.317**	0.36	0.669
AraBERT256-Flickr8k	Flickr8k	0.387	0.244	**0.151**	**0.093**	**0.334**	0.312	**0.428**	0.668
GigaBERT32-Flickr8k		0.386	0.241	0.144	0.0827	0.331	0.315	0.403	0.669
GigaBERT32-COCO		0.36	0.215	0.124	0.0708	0.308	0.311	0.344	0.668
	Δ	**0.059** ↑	**0.053** ↑	**0.046** ↑	**0.036** ↑				

Table 3: Our model scores compared to previous models. The highest scores on our test-split are marked in bold. Of all the previous ones, only the model by ElJundi et al. (2020) uses the same test-split as us. Other test-splits are unknown.



(a) (b) (c) (d)

Figure 7: Human evaluation of four candidate captions produced by AraBERT32-COCO: two accurate candidate captions (a) and (b), and two inaccurate candidate captions (c) and (d). Each candidate caption is accompanied by the reference caption from the Flickr8k test-split with the most MUSE similarity, and a THUMB score.

Conclusion

- Arabic Image Captioning using transformers
- Presented a method to adapt OSCAR to other languages
- Achieved better results than previous work
- Proposed working configurations and heuristics
- Hope to see many contributions to the field!

References

Imad Afyouni, Imtinan Azhara, and Ashraf Elnagar. 2021. AraCap: A hybrid deep learning architecture for Arabic Image Captioning. In ACLing 2021: 5th International Conference on AI in Computational Linguistics.

Huda A. Al-muzaini, Tasniem N. Al-yahya, and Hafida Benhidour. 2018. Automatic Arabic image captioning using RNN-LSTM-based language model and CNN. International Journal of Advanced Computer Science and Applications, 9(6).

Obeida ElJundi, Mohamad Dhaybi, Kotaiba Mokadam, Hazem Hajj, and Daniel Asmar. 2020. Resources and End-to-End Neural Network Models for Arabic Image Captioning. In 15th International Conference on Computer Vision Theory and Applications.

Vasu Jindal. 2018. Generating image captions in Arabic using root-word based recurrent neural networks and deep neural networks. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 144–151, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Computer Vision – ECCV 2020.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5579–5588.