Content

Introduction

What is Image Captioning? [2]

Image Data

↓

Textual Data

a bird with red, black and blue feathers is standing on a wooden table and is eating a red apple;

Example use cases:

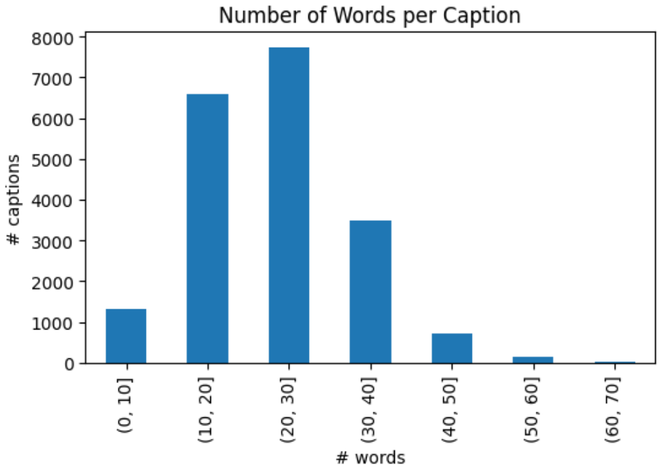
- automatic image indexing
- content-based image retrieval
- assistance for people with visual impairments

Dataset

Data as we found it

IAPR TC-12 consists of:

- nearly **20,000 images** taken from locations around the world (pictures of sports, actions, people, animals, cities, landscapes and many others).
- each image is associated with a **single text caption** in up to three different languages (**English**, German and Spanish).



Dataset

Data cleaning and preparation

Textual Data

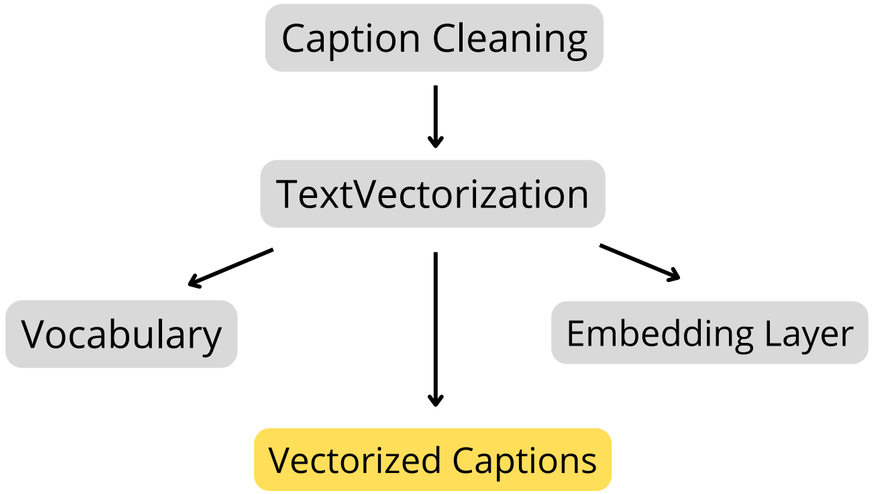
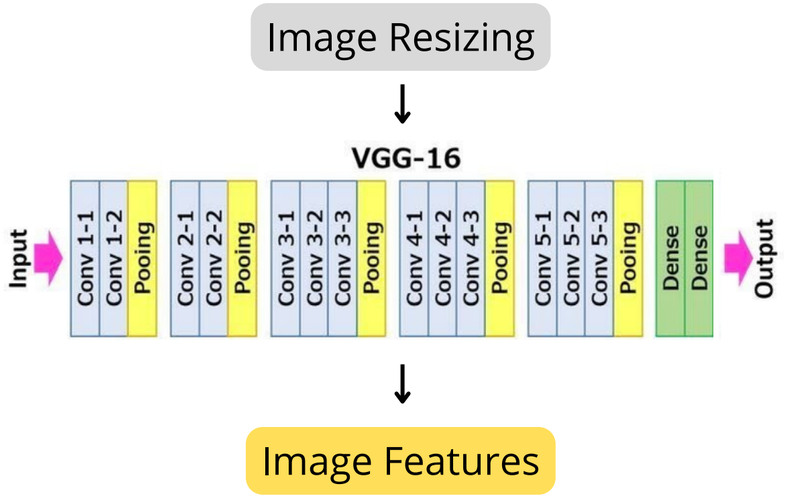


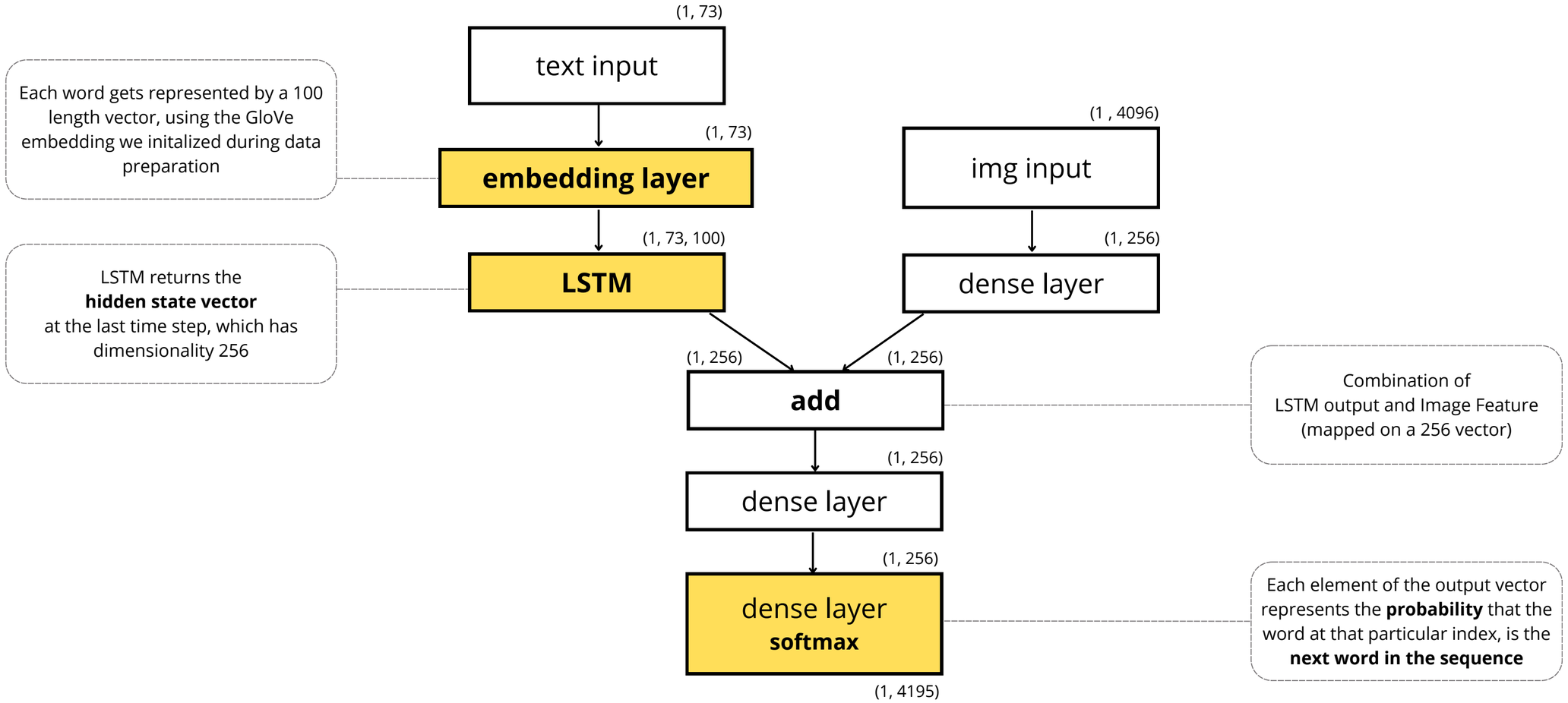
Image Data



We finally obtained **19983** samples, that we splitted in train (70%), validation (10%) and test (20%).
For each sample we obtained a (1, 4096) image feature vector and a (1, 73) vectorized caption.

Proposed Solution [1]

CNN + LSTM

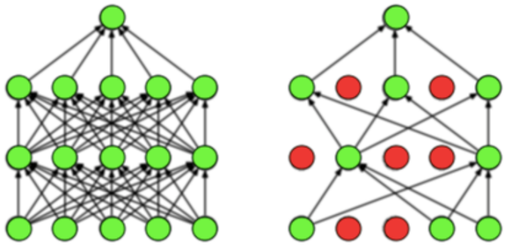


Experiments 1 and 2

CNN + LSTM with Regularization Techniques

Basic + Dropout

Dropout refers to the practice of ignoring some neural units in the training phase.
It functions as a form of regularization by **"dropping out" certain nodes** from the network.



(a) Standard Neural Net (b) After applying dropout.

Basic + L1 Regularization

L1 regularization is a technique used to **add a penalty to the loss function** of a model based on the absolute values of its parameters.
It **encourages sparsity** in the parameter values.

$$\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) + \alpha\Omega(\boldsymbol{\theta})$$

$$\Omega(\boldsymbol{\theta}) = ||\boldsymbol{w}||_1 = \sum_i |w_i|$$

Experiments 3 and 4

CNN + LSTM with Different Data

Basic + Dropout + Data Augmentation

Data Augmentation is a technique used to **artificially generate new training data** by modifying existing data, aiming to **improve the performance and generalization** of the model.

We used Synonym Replacement, Random Deletion, Translation, Sequence Shuffle.

We obtained **5 captions for each image**.

Basic + Dropout with Places365 Weights

Places365 is a scene recognition dataset.
It is composed of 10 million images, making 434 scene classes.

We tried to use different weights for extracting image features from VGG16.

Evaluation Metrics

Models Evaluation using BLEU Score and METEOR

BLEU and METEOR are two metrics for **automatic evaluation of machine translation** that calculate the similarity between a machine translation output and a reference translation using **n-grams**.

BLEU

$$BLEU = min\left(1, exp\left(1 - \frac{reference\ length}{output\ length}\right)\right) \cdot \prod_{i=1}^{n} P_i^{w_i}$$

Precision: $P = \frac{m}{w_t}$

- m: # unigrams in the candidate found in the reference
- w_t: # unigrams in the candidate

METEOR

$$M = F_{mean}(1 - p)$$

Recall: $R = \frac{m}{w_r}$

- w_r: # unigrams in the reference

$$F_{mean} = \frac{10PR}{R + 9P}$$

$$p = 0.5\left(\frac{c}{u_m}\right)^3$$

- c: # chunks in the candidate
- u_m: # unigrams in the candidate

Results

cyclist with red blue and white jersey black cycling shorts and blue helmet is riding on black and red racing bike on grey road green and brown grass dark green trees and blue sky in the background

BLEU-1	METEOR	
0.619	0.57	**Basic: VGG16 on ImageNet + LSTM** cyclist with red red and red and red jersey black cycling shorts and red racing bike on grey road with red racing bike on grey road with red and blue sky in the background
0.809	0.764	**Basic + Dropout** cyclist with red blue and white jersey black cycling shorts and blue helmet is riding on black and red racing bike on grey road in flat landscape with green meadows and trees in the background
0.258	0.206	**Basic + L1 Regularization** man with brown of brown of brown of brown of brown and and and and and and and and and and and and and and and in the background
0.414	0.407	**Basic + Dropout with Places365 Weights** man with black jacket and black trousers is standing on green lawn in the foreground dark green trees and bushes behind it light grey sky in the background
0.829	0.883	**Basic + Dropout + Data Augmentation** cyclist with red blue and white jersey black cycling shorts and blue helmet is riding on black and red racing bike on grey road in flat landscape with green meadows and green trees and bushes and blue sky in the background

Results

snow covered mountain landscape with grey clouds above it

BLEU-1	METEOR	
0.0	0.0	**Basic: VGG16 on ImageNet + LSTM** panoramic view of glaciar glaciar glaciar (...) glaciar
0.142	0.105	**Basic + Dropout** view of glacier with snow in the foreground and blue sky in the background
0.035	0.045	**Basic + L1 Regularization** man with brown of brown of brown of brown of brown and and and and and and and and and and and and and and and and and and in the background
0.303	0.338	**Basic + Dropout with Places365 Weights** view of the snow covered mountain
0.0	0.0	**Basic + Dropout + Data Augmentation** view of the edge of the edge of the edge of the desert

Performance Comparison

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
BASIC	0.272	0.17	0.11	0.047	0.279
+ DROPOUT	0.291	0.185	0.124	0.064	0.296
+ L1 REG	0.173	0.078	0.046	0.0	0.17
+ DROPOUT + DATA AUG	0.256	0.151	0.096	0.044	0.254
+ DROPOUT WITH PLACES	0.257	0.162	0.108	0.053	0.268
Ref. Paper [5] m-RNN-AlexNet	**0.482**	**0.357**	**0.269**	**0.208**	**-**

Next Steps

Other possible experiments

- Adding **Attention** to the network
- Using German and Spanish captions to do data augmentation
- Using other pretrained models to extract image features
- Applying the same models to **other datasets** (Flickr30k, MS COCO)

Resources

[1] Suresh, K.R., Jarapala, A. & Sudeep, P.V. **Image Captioning Encoder–Decoder Models Using CNN-RNN Architectures: A Comparative Study**. Circuits Syst Signal Process 41, 5719–5742 (2022). https://doi.org/10.1007/s00034-022-02050-2

[2] Hossain, MD Zakir, et al. "**A comprehensive survey of deep learning for image captioning.**" ACM Computing Surveys (CsUR) 51.6 (2019): 1-36.

[3] Papineni, Kishore, et al. "**Bleu: a method for automatic evaluation of machine translation.**" Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[4] Michael Denkowski and Alon Lavie. 2014. **Meteor Universal: Language Specific Translation Evaluation for Any Target Language.** In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

[5] Mao, Junhua, et al. "**Deep captioning with multimodal recurrent neural networks (m-rnn).**" arXiv preprint arXiv:1412.6632 (2014).