## 1] Image Captioning with Attention for Smart Local Tourism using Efficient Net

**Author** : Dhomas Hatta Fudholi, Yurio Windiatmoko, Nurdi Afrianto, Prastyo Eko Susanto, Magfirah Suyuti, Ahmad Fathan Hidayatullah, Ridho Rahmadi

**Dataset** used local-tourism dataset used in Indonesia

Done using **GRU and EfficientNetB0**

Research enhances AI-assisted systems for local tourism, using deep learning. Model uses EfficientNetB0, achieving BLEU scores of 73.39(training) and 24.51(validation).

they have results as above

| Architecture | Average of the BLEU_train_score | Average of the BLEU_val_score |
|---|---|---|
| EfficientNetB4 | 72.84 | 22.24 |
| EfficientNetB0 | **73.39** | **24.51** |
| VGG16 | 68.67 | 19.33 |
| InceptionV3 | 58.67 | 22.41 |

## 2]Empirical Analysis of Image Caption Generation using Deep Learning

Author : Aditya Bhattacharya ,Eshwar Girishekar , Padmaker Deshpande

Implemented multi-modal image captioning networks using ResNet101, DenseNet121, VGG19 encoders and LSTM decoders with attention. Evaluated with BLEU, CIDEr, ROUGE, METEOR metrics. Explored beam size, pretrained embeddings, and visual attention maps for model explainability.

Dataset **: Flickr8K,MSCOCO**

Best got **from ResNet101 Trained on COCO (Encode-LSTM ,Decoder-Attention network)**

Scores for it: **BLEU-432.64,CIDEr-102.66,METEOR-47.37,ROUGE-54.32**

## 3] Image Captioning with Semantic Attention

Author : Quanzeng You1 , Hailin Jin2 , Zhaowen Wang2 , Chen Fang2 , and Jiebo Luo1

It’s fusion of computer vision and NLP. it propose an algorithm combining top-down and bottom-up approaches using semantic attention, outperforming state-of-the-art methods on **COCO** and **Flickr30K** benchmarks.

| Dataset | Model | B-1 | B-2 | B-3 | B-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|---|---|
| Flickr30k | Ours-GT-ATT | 0.824 | 0.679 | 0.534 | 0.412 | 0.269 | 0.588 | 0.949 |
| | Ours-GT-MAX | 0.719 | 0.542 | 0.396 | 0.283 | 0.230 | 0.529 | 0.747 |
| | Ours-GT-CON | 0.708 | 0.534 | 0.388 | 0.276 | 0.222 | 0.516 | 0.685 |
| MS-COCO | Ours-GT-ATT | 0.910 | 0.786 | 0.654 | 0.534 | 0.341 | 0.667 | 1.685 |
| | Ours-GT-MAX | 0.790 | 0.635 | 0.494 | 0.379 | 0.279 | 0.580 | 1.161 |
| | Ours-GT-CON | 0.766 | 0.617 | 0.484 | 0.377 | 0.279 | 0.582 | 1.237 |

Table 1. Performance of the proposed models using the ground-truth visual attributes on MS-COCO and Flickr30k.

## 4] Deep Visual-Semantic Alignments for Generating Image Descriptions

Author : Andrej Karpathy , Li Fei-Fei

It is  descriptions of both images and their regions by learning inter-modal correspondences between language and visual data. The architecture combines **CNNs o**ver image regions, bidirectional RNNs over sentences, and a structured objective for alignment. Evaluation on Flickr8K, Flickr30K, and MSCOCO datasets demonstrates state-of-the-art results in retrieval, and our generated descriptions surpass baseline performance on full images and region-level annotations.

**Dataset** : Flickr8K, Flickr30K & MSCOCO

| | Image Annotation | | | | Image Search | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **R@1** | **R@5** | **R@10** | **Med *r*** | **R@1** | **R@5** | **R@10** | **Med *r*** |
| **Flickr30K** | | | | | | | | |
| SDT-RNN (Socher et al. [49]) | 9.6 | 29.8 | 41.1 | 16 | 8.9 | 29.8 | 41.1 | 16 |
| Kiros et al. [25] | 14.8 | 39.2 | 50.9 | 10 | 11.8 | 34.0 | 46.3 | 13 |
| Mao et al. [38] | 18.4 | 40.2 | 50.9 | 10 | 12.6 | 31.2 | 41.5 | 16 |
| Donahue et al. [8] | 17.5 | 40.3 | 50.8 | 9 | - | - | - | - |
| DeFrag (Karpathy et al. [24]) | 14.2 | 37.7 | 51.3 | 10 | 10.2 | 30.8 | 44.2 | 14 |
| Our implementation of DeFrag [24] | 19.2 | 44.5 | 58.0 | 6.0 | 12.9 | 35.4 | 47.5 | 10.8 |
| Our model: DepTree edges | 20.0 | 46.6 | 59.4 | 5.4 | 15.0 | 36.5 | 48.2 | 10.4 |
| Our model: BRNN | **22.2** | **48.2** | **61.4** | **4.8** | **15.2** | **37.7** | **50.5** | **9.2** |
| Vinyals et al. [54] (more powerful CNN) | 23 | - | 63 | 5 | 17 | - | 57 | 8 |
| **MSCOCO** | | | | | | | | |
| Our model: 1K test images | 38.4 | 69.9 | 80.5 | 1.0 | 27.4 | 60.2 | 74.8 | 3.0 |
| Our model: 5K test images | 16.5 | 39.2 | 52.0 | 9.0 | 10.7 | 29.6 | 42.2 | 14.0 |

Table 1. Image-Sentence ranking experiment results. **R@K** is Recall@K (high is good). **Med** *r* is the median rank (low is good). In the results for our models, we take the top 5 validation set models, evaluate each independently on the test set and then report the average performance. The standard deviations on the recall values range from approximately 0.5 to 1.0.

## 5] DEEP CAPTIONING WITH MULTIMODAL RECURRENT NEURAL NETWORKS (M-RNN)

Author: Junhua Mao, Wei Xu & Yi Yang & Jiang Wang & Zhiheng Huang, Alan Yuille.

The paper introduces a **multimodal Recurrent Neural Network (m-RNN)** for generating image captions. It combines recurrent and convolutional networks, achieving superior performance on four benchmark datasets. Additionally, the model improves retrieval tasks compared to state-of-the-art methods by optimizing the ranking objective function.

Dataset : IAPR TC-12, Flickr 8K, Flickr 30K and MS COCO

| | $\mathcal{PPL}$ | B-1 | B-2 | B-3 | B-4 |
|---|---|---|---|---|---|
| LBL, Mnih & Hinton (2007) | 9.29 | 0.321 | 0.145 | 0.064 | - |
| MLBLB-AlexNet, Kiros et al. (2014b) | 9.86 | 0.393 | 0.211 | 0.112 | - |
| MLBLF-AlexNet, Kiros et al. (2014b) | 9.90 | 0.387 | 0.209 | 0.115 | - |
| Gupta et al. (2012) | - | 0.15 | 0.06 | 0.01 | - |
| Gupta & Mannem (2012) | - | 0.33 | 0.18 | 0.07 | - |
| Ours-RNN-Base | 7.77 | 0.307 | 0.177 | 0.096 | 0.043 |
| Ours-m-RNN-AlexNet | **6.92** | **0.482** | **0.357** | **0.269** | **0.208** |

Table 1: Results of the sentence generation task on the IAPR TC-12 dataset. "B" is short for BLEU.

| | Sentence Retrival (Image to Text) | | | | Image Retrival (Text to Image) | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | Med r | R@1 | R@5 | R@10 | Med r |
| Ours-m-RNN | 20.9 | 43.8 | 54.4 | 8 | 13.2 | 31.2 | 40.8 | 21 |

Table 2: R@K and median rank (Med r) for IAPR TC-12 dataset.

| | Sentence Retrival (Image to Text) | | | | Image Retrival (Text to Image) | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | Med r | R@1 | R@5 | R@10 | Med r |
| Random | 0.1 | 0.5 | 1.0 | 631 | 0.1 | 0.5 | 1.0 | 500 |
| SDT-RNN-AlexNet | 4.5 | 18.0 | 28.6 | 32 | 6.1 | 18.5 | 29.0 | 29 |
| Socher-avg-RCNN | 6.0 | 22.7 | 34.0 | 23 | 6.6 | 21.6 | 31.7 | 25 |
| DeViSE-avg-RCNN | 4.8 | 16.5 | 27.3 | 28 | 5.9 | 20.1 | 29.6 | 29 |
| DeepFE-AlexNet | 5.9 | 19.2 | 27.3 | 34 | 5.2 | 17.6 | 26.5 | 32 |
| DeepFE-RCNN | 12.6 | 32.9 | 44.0 | 14 | 9.7 | 29.6 | **42.5** | **15** |
| Ours-m-RNN-AlexNet | **14.5** | **37.2** | **48.5** | **11** | **11.5** | **31.0** | 42.4 | **15** |

Table 3: Results of R@K and median rank (Med r) for Flickr8K dataset. "-AlexNet" denotes the image representation based on AlexNet extracted from the whole image frame. "-RCNN" denotes the image representation extracted from possible objects detected by the RCNN algorithm.