

Image Captioning with Flickr30k Dataset

Meeting #3

Ye Kyaw Thu
Lab Leader, LU Lab., Myanmar

22 June 2025 (SUN)

Table of Content

- Flickr30K Dataset Information
- Overview
- System Workflow & Coding Information

Overview

- **Image Team အတွက် ဂျရက်လောက် အချိန်ပေး စမ်းဖြစ်ခဲ့**
- **What is Image Captioning?**
 - Automatically generate textual descriptions of images
- **Key Components:**
 - CNN (VGG16) for image features
 - RNN (GRU/LSTM) for sequence generation
 - Multilingual support (English/Myanmar)

Flickr30K Dataset Information

Flickr30k Dataset

Estimated reading: 4 minutes • 8483 views



Visualization of the [Flickr30k dataset](https://datasets.activeloop.ai/docs/ml/datasets/flickr30k-dataset/) in the Deep Lake UI

- The Flickr30k dataset is a popular benchmark for sentence-based picture portrayal. The dataset is comprised of 31,783 images that capture people engaged in everyday activities and events. Each image has a descriptive caption. Flickr30k is used for understanding the visual media (image) that correspond to a linguistic expression (description of the image). This dataset is commonly used as a standard benchmark for sentence-based image descriptions.

Link: <https://datasets.activeloop.ai/docs/ml/datasets/flickr30k-dataset/>

Flickr30K Dataset Information



- 1. a man prepares to enter the red building .
- 2. a man walking around the corner of a red building .
- 3. a man walks past a red building with a fake rocket attached to it .
- 4. a man walks under a building with a large rocket shaped sculpture .
- 5. a person walking by a red building with a jet on top of it .

Flickr30K Dataset Information



- 1. a black dog playing with a purple toy in the snow .
- 2. a black dog runs through the snow carrying a blue toy .
- 3. a dog plays in the snow .
- 4. dog running with a purple toy in the snowy field .
- 5. the black and brown dog carries a purple toy in the snow .

Flickr30K Dataset Information



- 1. a guy and a girl , both wearing white shirts and jeans , stand under a flowering tree .
- 2. a man and a woman are talking in a park
- 3. a man and woman standing underneath the tree are talking .
- 4. a man in a white shirt is standing in the grass showing something to a woman in a white shirt .
- 5. a young couple both wearing white shirts and blue jeans standing in a light misty rain

Flickr30K Dataset Information

```
./data
├── captions.txt
├── flickr30k_images
│   ├── captions.txt
│   └── Images
│       ├── 1000092795. jpg
│       ├── 10002456. jpg
│       ├── 1000268201. jpg
│       ├── 1000344755. jpg
│       ├── 1000366164. jpg
│       ├── 1000523639. jpg
│       ├── 1000919630. jpg
│       ├── 10010052. jpg
│       ├── 1001465944. jpg
│       ├── 1001545525. jpg
│       ├── 1001573224. jpg
│       ├── 1001633352. jpg
│       ├── 1001773457. jpg
│       ├── 1001896054. jpg
│       ├── 100197432. jpg
│       ├── 100207720. jpg
│       ├── 1002674143. jpg
│       ├── 1003163366. jpg
│       ├── 1003420127. jpg
│       └── 1003428081. jpg
```

- Folder structure of Flickr30k
- Label ဖိုင် နှစ်ဖိုင် ရှိတယ်
- Images ဖိုင်ဒါအောက်မှာ ID နဲ့ သိမ်းထားတဲ့ image ဖိုင်တွေ ရှိလိမ့်မယ်

Flickr30K Dataset Information

```
(base) ye@1st-hpc3090:~/intern3/img2txt/data$ head -n 20 ./captions.txt
1000268201_693b08cb0e.jpg#0    A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1    A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2    A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3    A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4    A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0    A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1    A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2    A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg#3    Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4    Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0    A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .
1002674143_1b742ab4b8.jpg#1    A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2    A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
1002674143_1b742ab4b8.jpg#3    There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg#4    Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0    A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1    A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2    a man sleeping on a bench outside with a white and black dog sitting next to him .
1003163366_44323f5815.jpg#3    A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4    man laying on bench holding leash of dog sitting on ground
```

Fig. Flickr80k style

Flickr30K Dataset Information

```
(base) ye@l1st-hpc3090:~/intern3/img2txt/data/flickr30k_images$ head -n 20 captions.txt
image,caption
1000092795.jpg, Two young guys with shaggy hair look at their hands while hanging out in the yard .
1000092795.jpg, " Two young , white males are outside near many bushes ."
1000092795.jpg, Two men in green shirts are standing in a yard .
1000092795.jpg, A man in a blue shirt standing in a garden .
1000092795.jpg, Two friends enjoy time spent together .
10002456.jpg, Several men in hard hats are operating a giant pulley system .
10002456.jpg, Workers look down from up above on a piece of equipment .
10002456.jpg, Two men working on a machine wearing hard hats .
10002456.jpg, Four men on top of a tall structure .
10002456.jpg, Three men on a large rig .
1000268201.jpg, A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201.jpg, A little girl in a pink dress going into a wooden cabin .
1000268201.jpg, A little girl climbing the stairs to her playhouse .
1000268201.jpg, A little girl climbing into a wooden playhouse
1000268201.jpg, A girl going into a wooden building .
1000344755.jpg, Someone in a blue shirt and hat is standing on stair and leaning against a window .
1000344755.jpg, A man in a blue shirt is standing on a ladder cleaning a window .
1000344755.jpg, A man on a ladder cleans the window of a tall building .
1000344755.jpg, man in blue shirt and jeans on ladder cleaning windows
```

Fig. Flickr30k style

System Workflow

- **Visual Flow:**

Image → VGG16 (Feature Extraction) →
GRU/LSTM (Caption Generation)

- **Key Modules:**

- Data loading
- Feature extraction
- Model training
- Evaluation

Feature Extraction with VGG16

- **Process:**

- Resize images to 224x224
- Extract 4096-dim features from VGG16's penultimate layer

- **Output:**

- Saved as `features.pkl` for reuse

- **Code**

- `features = extract_image_features(image_paths, vgg_model)`

Caption Tokenization

- **Tokenizer**

- Converts words to integers (e.g., "dog" → 42)
- Calculates `vocab_size` and `max_length` of captions

- **Code**

- `tokenizer.fit_on_texts(captions)` # Vocab size = 10,000

Train-Test Split

- **Split Ratio:** 90% train, 10% test
- **Key Code:**
 - `split = int(len(image_ids) * 0.9)`
 - `train_keys = image_ids[:split]`
 - `test_keys = image_ids[split:]`

GRU vs. LSTM

- **GRU (Gated Recurrent Unit):**
 - Fewer parameters → Faster training
 - Single gate structure
- **LSTM (Long Short-Term Memory):**
 - More parameters → Better for long sequences
 - 3 gates (input, forget, output)
- **Code Switch:**
 - `if cell_type == 'gru': se3 = GRU(units)`
 - `else: se3 = LSTM(units)`

Model Architecture

- **Inputs:**

- Image features (4096-dim)
- Caption sequences (padded to max_length)

- **Layers:**

- Embedding → Dropout → GRU/LSTM → Dense

- **Loss:**

- Categorical cross-entropy

Training Process

- **Parameters**

- Batch size: 24
- Epochs: 200
- Learning rate: 0.001, 0.002

- **Code**

- `generator = data_generator(train_keys, features, ...)`

Evaluation Metrics

- **Metrics Used:**

- BLEU (1-4), chrF++, CIDEr, ROUGE (1/2/L)
- Semantic similarity (TF-IDF + keyword overlap)

- **Example Output:**

- BLEU-1: 0.2292
- ROUGE-L: 0.1530
- CIDEr: 0.0003

Evaluation Metrics

- **TF-IDF Cosine Similarity**

- Converts text to vectors (ignoring stopwords)
- Measures contextual similarity via `cosine_similarity()`

- **Code**

- `tfidf = vectorizer.fit_transform(all_texts)`
- `sim = cosine_similarity(tfidf[0:1], tfidf[1:2])[0][0]`

Evaluation Metrics

- **Keyword Overlap**

- Exact word matches normalized by predicted caption length:

- **Code**

- $\text{len}(\text{pred_words} \ \& \ \text{ref_words}) / \max(1, \text{len}(\text{pred_words}))$

- **Composite Score**

- Weighted sum: $0.7 * \text{similarity} + 0.3 * \text{keyword_overlap}$

Evaluation Metrics

```
## Own implementation for semantic similarity measurement
def semantic_similarity(predicted, references):
    """Calculate semantic similarity between predicted and reference captions"""
    # Combine all references
    all_texts = [' '.join(references), predicted]

    # Calculate TF-IDF vectors
    vectorizer = TfidfVectorizer(stop_words='english')
    try:
        tfidf = vectorizer.fit_transform(all_texts)
    except ValueError:
        return {'similarity': 0.0, 'keyword_overlap': 0.0}

    # Calculate cosine similarity
    sim = cosine_similarity(tfidf[0:1], tfidf[1:2])[0][0]

    # Keyword matching
    pred_words = set(predicted.lower().split())
    ref_words = set(' '.join(references).lower().split())
    keyword_overlap = len(pred_words & ref_words) / max(1, len(pred_words))

    return {
        'similarity': sim,
        'keyword_overlap': keyword_overlap,
        'composite_score': 0.7*sim + 0.3*keyword_overlap
    }
```

Results Analysis (Good Example 1)

- **Image ID: 437527058**



Results Analysis (Good Example 1)

- **Image ID:** 437527058
- **Actual:** "a caravan of snowmobiles travels through snow"
- **Predicted:** "helmet and dog walking through snow"
- **Semantic Score:** 0.3424 (highest composite score)
- **Why Good?:** Partial keyword overlap ("snow") + contextually plausible

Results Analysis (Good Example 2)

- **Image ID:** 441212506



Results Analysis (Good Example 2)

- **Image ID:** 441212506
- **Actual:** "three dogs playing in a field"
- **Predicted:** "dog with frisbee rolling in grass"
- **Semantic Score:** 0.3059
- **Why Good?:** Captured "dog" and outdoor activity

Results Analysis (Good Example 3)

- **Image ID: 440737340**



Results Analysis (Good Example 3)

- **Image ID:** 440737340
- **Actual:** "masked man carrying a box"
- **Predicted:** "tattoos sitting on a bench"
- **Semantic Score:** 0.0321 (lowest)
- **Why Good?:** Complete mismatch

CLI Usage

- **Commands:**

- **# Train**

- `python image_captioning.py --train --epochs 200
--cell_type gru`

- **# Predict**

- `python image_captioning.py --predict img.jpg`

CLI Usage

Image Captioning with Flickr30k Dataset

optional arguments:

```
-h, --help            show this help message and exit
--data_dir DATA_DIR  Directory to store dataset (default: ./data)
--model_dir MODEL_DIR
                    Directory to save/load models (default: ./models)
--epochs EPOCHS       Number of training epochs (default: 15)
--batch_size BATCH_SIZE
                    Training batch size (default: 64)
--train              Train the model
--evaluate           Evaluate on test set
--predict PREDICT     Path to single image for prediction
--language {english,myanmar}
                    Caption language (default: english)
--model_name MODEL_NAME
                    Model filename (default: best_model.h5)
--skip_download      Skip dataset download (for prediction mode)
--skip_feature_extraction
                    Skip feature extraction if features file exists
--lstm_units LSTM_UNITS
                    Number of units in LSTM layer (default: 256)
--dropout_rate DROPOUT_RATE
                    Dropout rate (default: 0.4)
--learning_rate LEARNING_RATE
                    Learning rate (default: 0.001)
--early_stopping EARLY_STOPPING
                    Patience for early stopping (default: None)
--feature_size FEATURE_SIZE
                    Size of image features (default: 4096)
--cell_type {lstm,gru}
                    RNN cell type (default: gru)
```

To Do

- **Transformer-based models**

- အချိန်ရတဲ့အခါ လက်ရှိ code ကိုပဲ အခြေခံပြီး
Transformer architecture နဲ့ ဖြည့်ရေးတာလုပ်နိုင်

- **Better Myanmar language support**

- လက်ရှိ ဒေတာကိုလည်း မြန်မာလို ဘာသာပြန်လိုက်ပြီး
အဲဒါကို experiment အနေနဲ့ လုပ်ကြည့်တာမျိုး