Video Description

Recent Advances in Computer Vision

Ishita Ankit | Lakshay Garg | Shubham Jain

Problem Description

Video Description is one of the long-standing goals of computer vision research. Generating automatic descriptions of videos enables us to index large, heterogeneous collections of video material and helps produce video summaries.

We have tried to solve this problem by implementing two different models -

- Deep Compositional Captioner (DCC)
- Sequence to Sequence - Video to Text (S2VT)

Deep Compositional Captioner

DCC consists of the following three modules :

1) **Deep Lexical Classifier -**
    - A convolutional neural network (CNN) which maps images to semantic concepts.
2) **Language Model -**
    - Learns sentence structure on unpaired training data.
    - Unpaired data is obtained from many sources to improve language quality and generalizability.
    - Includes an embedding layer which maps one-hot vector representation of a word to lower dimensional space.
    - It is trained to predict the next word in a sentence given the previous word
3) **Caption Model** -
    - Integrates the lexical classifier and language model to learn a joint model for image description

Lexical Classifier - Training

The lexical classifier is trained by fine-tuning a deep CNN trained on ILSVRC-2012 object recognition training subset of of ImageNet.

a) Train on all the MS COCO images
b) Train on MS COCO images without multiple lables for the eight held out classes (in-domain)
c) Train with MSCOCO images except for objects which are held out during paired training. These categories are trained with ImageNet data (out-of-domain)
d) Train on all MS COCO images and over 600 ImageNet objects not in COCO

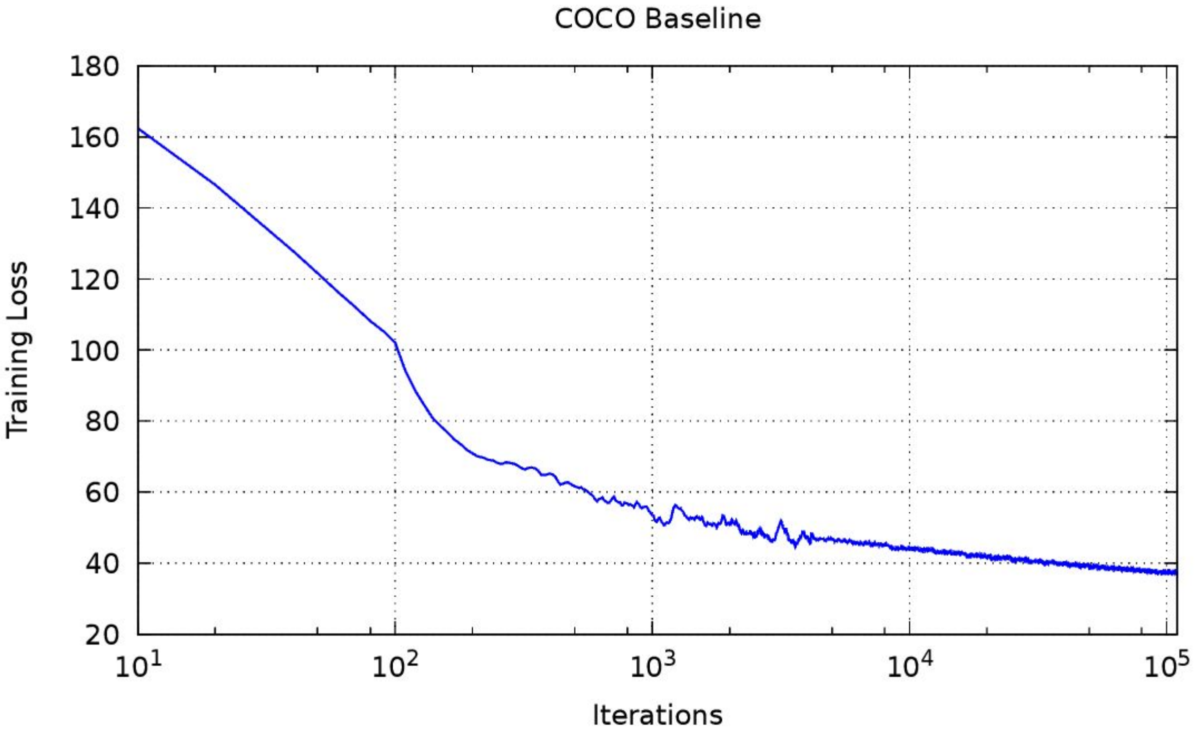
Language Model - Training

- Trained on MSCOCO text - All captions from MSCOCO train set (in-domain)
- Trained on WebCorpus text - 60 million sentences from British National Corpus (BNC), UkWaC and Wikipedia (out-of-domain)
- Trained on Caption text - Text data from other paired image and video description datasets (Flickr1M, Flickr30k, Pascal1k, ImageCLEF-2012) and sentence descriptions of Youtube clips from MSVD training corpus. This does not include sentences from MSCOCO (out-of-domain)

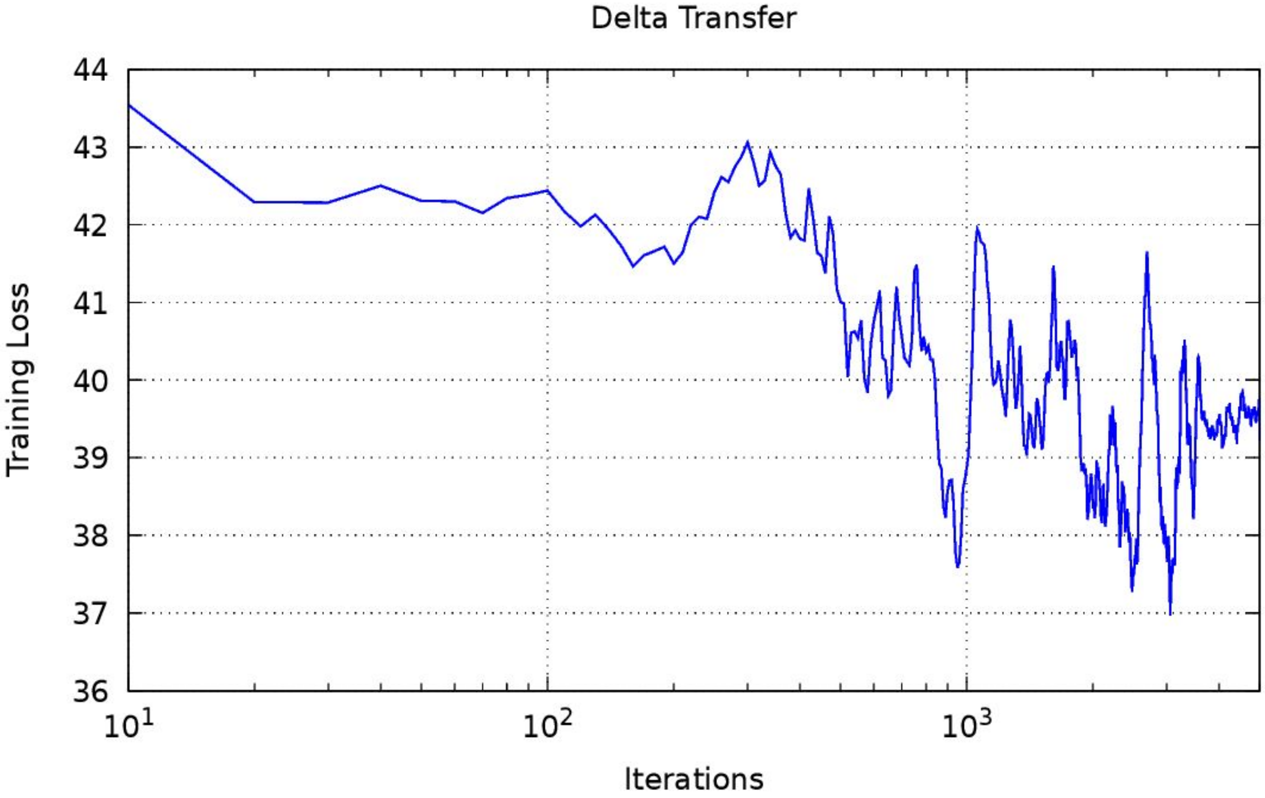
Caption Model - Training

- Trained with pair supervision (baseline) - trained only on paired data
- Direct transfer model + in-domain text + in-domain images
- Delta transfer + in-domain text + in-domain images
- Direct transfer + in-domain text + out-of-domain images
- Direct transfer + out-domain text + out-domain images (out-domain from webcorpus and captiontxt)
- Direct transfer model for describing imagenet objects

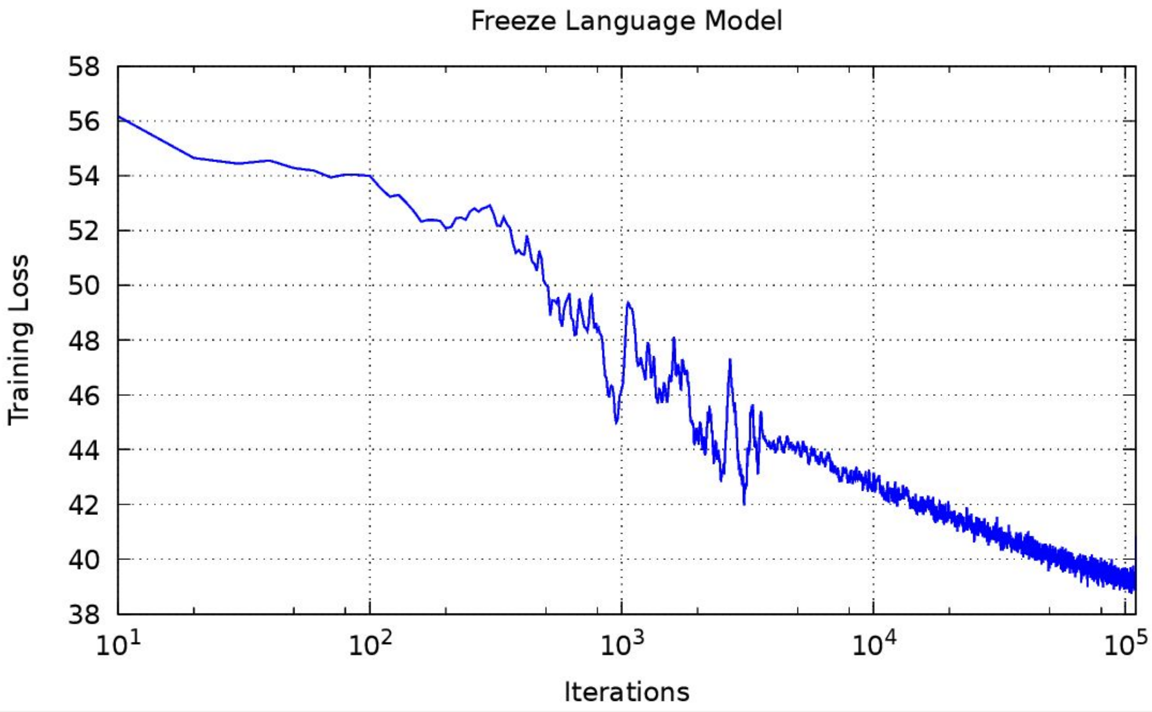
Training - COCO Baseline



Training - Delta Transfer Training Loss



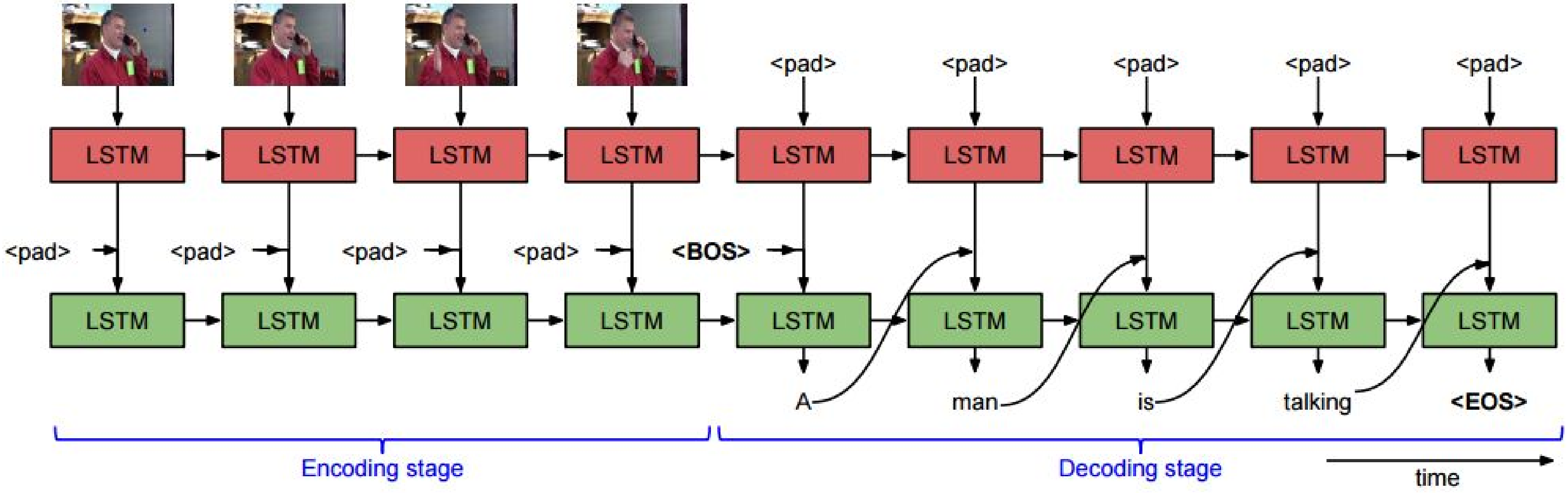
Training - Freeze Language Model Loss



Results: DCC model

- Trained the baseline DCC and direct transfer model with the following results -
    - Baseline Model -
        - METEOR - 0.23
        - ROUGE_L - 0.493
        - CIDEr - 0.765
        - F1 - 0.581
        - BLEU_1 - 0.672
    - Direct Transfer Model -
        - METEOR - 0.225
        - ROUGE_L - 0.486
        - CIDEr - 0.731
        - F1 - 0.581
        - BLEU_1 - 0.496

Sequence to Sequence - Video to Text



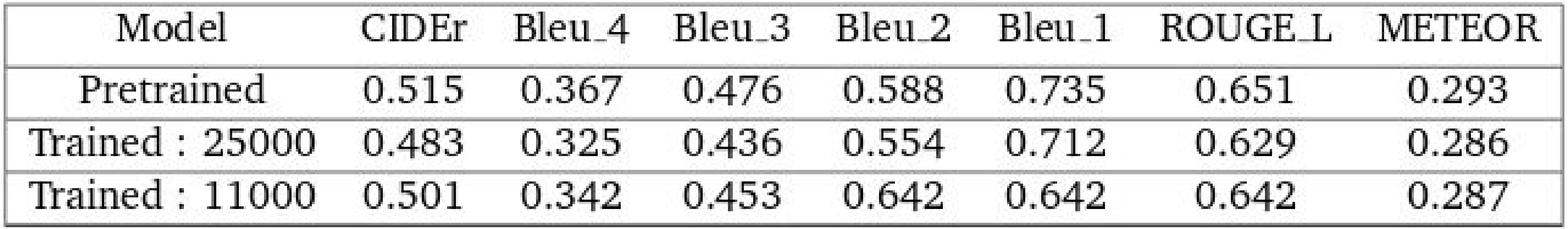
Dataset and Pre-Processing

- The dataset used were the Youtube clips present in Microsoft Video Dataset(MSVD).
- The dataset consists of 1970 videos of which 1200 were used for training, 100 for validation and rest for testing.
- The video clips were passed through a VGG net to extract the features in the format as used by the model.
- The extracted features and annotated sentences were then converted into an hdf5 file to be used in training the model.

Results from scaled down model of S2VT

- Due to memory constraints, we had trained the S2VT model on the smaller validation set instead of the training dataset. The following results were obtained -
    - METEOR - 0.225
    - ROUGE_L - 0.6000
    - CIDEr - 0.2440
    - BLEU_1 - 0.6800
    - BLEU_2 - 0.5030
    - BLEU_3 - 0.3910
    - BLEU_4 - 0.2670

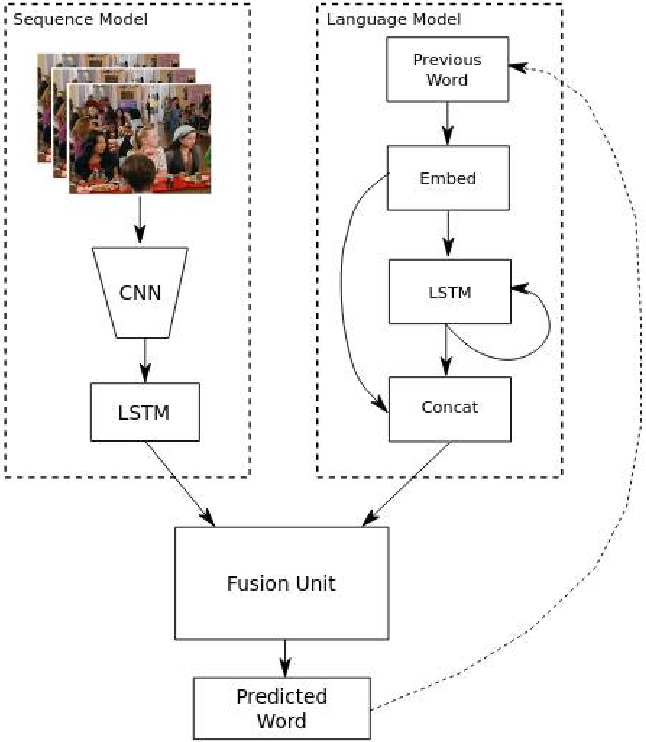
S2VT results : Original model



Model	CIDEr	Bleu_4	Bleu_3	Bleu_2	Bleu_1	ROUGE_L	METEOR
Pretrained	0.515	0.367	0.476	0.588	0.735	0.651	0.293
Trained : 25000	0.483	0.325	0.436	0.554	0.712	0.629	0.286
Trained : 11000	0.501	0.342	0.453	0.642	0.642	0.642	0.287

DCC modified architecture

- Combined S2VT and DCC into a single model.
- The lexical model of DCC which was trained to produce image features is replaced by the S2VT model to obtain video features.



Implementation

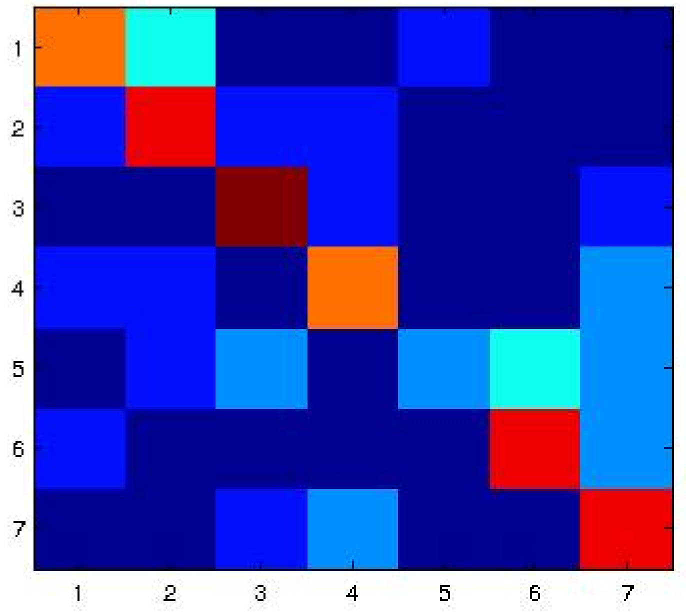
- The S2VT model concatenates features extracted from every frame (sampled at 5 frames per second) and passes them through a double-stacked LSTM architecture for caption prediction.
- The DCC model takes a feature representation from the lexical model as input and passes it into the multimodal unit which generates a caption by utilizing the language model.
- To connect the S2VT and DCC models, we average-pooled the output of LSTM in the S2VT model which was then passed to the multimodal unit of DCC.

Experiments with Audio

- Videos generally have audio data accompanying them which are not used by either the DCC or the S2VT model.
- We propose the following approach to use audio data to obtain more accurate captions -
    - Bag of Audio Words (BoAW) are used as features to represent the audio file
    - BoAW features are extracted by using the Multi-Modal features toolkit
    - As a small experiment BoAW features were extracted for 7 classes of different human actions and a random forest classifier was trained on them
    - Freesound dataset was used to obtain the audio files
    - A maximum of 62.43% accuracy is obtained for classification of the audio files into the seven classes.

Confusion Matrix

1) Train dataset - 630 audio files over 7 classes
2) Evaluation Dataset - 10 audio files per class
3) Classes -
   a) Laugh
   b) Cry
   c) Clap
   d) Breathing
   e) Sneeze
   f) Singing
   g) Whistling



Use in Captioning

- Information from audio can be used to refine captions given by the video captioning models.
- For the following video clip (screenshot on the right), the S2VT model gives the caption - A man is talking
- Audio classifier trained previously classifies the action verb correctly as laughing
- Hence, this information can be used to correct the caption given by the S2VT model

Thank You

Any Questions?