Improving Interpretability of Deep Neural Networks with Semantic Information

Yinpeng Dong, Hang Su, Jun Zhu, Bo Zhang

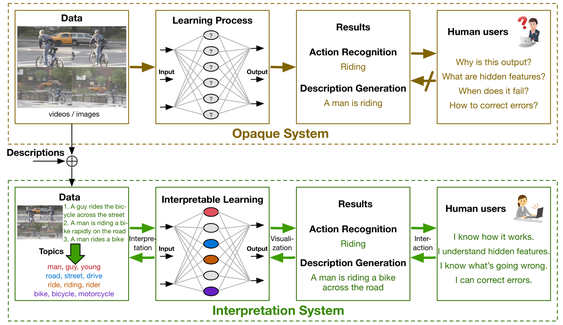
Department of Computer Science and Technology, Tsinghua University, Beijing, China

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 2017 HONOLULU

Introduction

Interpretability of deep neural networks (DNNs) is essential since it enables users to understand the overall strengths and weaknesses of the models, conveys an understanding of how the models will behave in the future, and how to diagnose and correct potential problems. However, it is challenging to reason about what a DNN actually does due to its opaque or black-box nature.

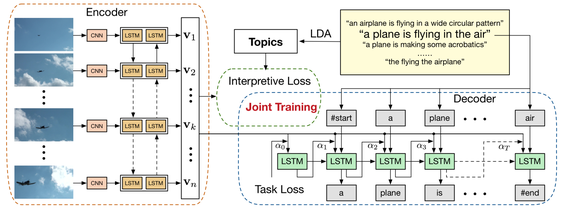


In this work, we

- Incorporates the interpretability of hidden features during training;
- Leverage the semantic information embedded in human descriptions;
- Interpret the neurons by a prediction difference maximization algorithm;
- Introduce a human-in-the-loop learning procedure.

Methodology

1. **Attentive encoder-decoder framework**
   - Encoder: GoogLeNet + Bi-directional LSTM
   - Decoder: Attentive LSTM



2. **Interpretive Loss**
   - Topics extracted from descriptions by LDA
   - $L_I(V, s) = \|f(\frac{1}{n}\sum_{i=1}^{n} v_i) - s\|_2^2$
3. **Training**

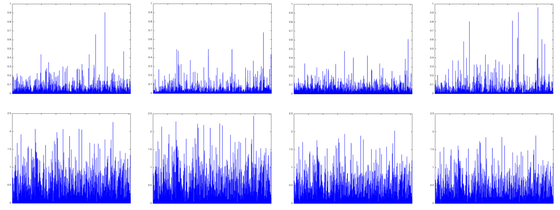
$$L = \frac{1}{N}\sum_{k=1}^{N}\left(\lambda\|f(\frac{1}{n}\sum_{i=1}^{n}) - s^k\|_2^2 - \sum_{t=1}^{N_s^k}\log p(y_t^k|y_{<t}^k, x^k)\right)$$

4. **Interpretation**
   - Each neuron can be represented by a semantic topic
   - Find the associations by the prediction difference maximization algorithm

$$j_i^* = \arg\max_j([f(v)]_i - [f(v_{\setminus j})]_i)$$

Results

We show the average activations of the representation for the videos which include a given topic (dog, girl, walk, dance). The first row shows the results of the proposed model LSTM-I and the second row shows the baseline model LSTM-B (without interpretive loss).



We also show the neuron activations with respect to relevant (R) and irrelevant (I) topics in sampled videos. We plot the activations of one neuron related to each topic through time.

We compare the performance of the proposed model with other state-of-the-art models of description generation on the YouTubeClips dataset measured by BLEU and METEOR.

Model (CNN features)	BLEU-4	METEOR
LSTM-B (GoogLeNet)	0.416	0.295
LSTM-I (GoogLeNet)	**0.446**	**0.297**
LSTM-YT (AlexNet)	0.333	0.291
S2VT (RGB + Optical Flow)	-	0.298
SA (GoogLeNet)	0.403	0.290
LSTM-E (VGG)	0.402	0.295
h-RNN (VGG)	0.443	**0.311**
SA (GoogLeNet + C3D)	0.419	0.296
LSTM-E (VGG + C3D)	0.453	0.310
h-RNN (VGG + C3D)	0.499	0.326

Human-in-the-Loop Learning

Motivation:

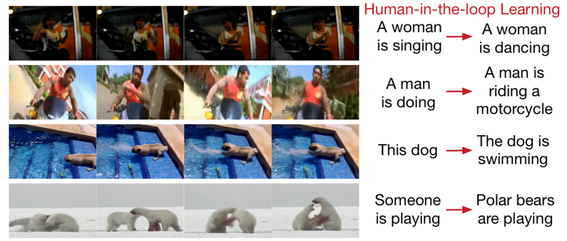
- Get human users involved in the learning process;
- Embed human knowledge into the model;
- Correct model's mistakes.

Procedure:

- For an inaccurate prediction, users provide **a missing topic** t;
- **Activation enhancement:**
  - Retrieve a set of neurons associated with t;
  - Calculate the average activations of these neurons of the training videos with t;
  - Enhance the activations by adding the average activations from v to v^*;
- **Correction propagation:** generalize the error to future unseen data and diagnose potential problems by finetuning the model:

$$L_{human} = \|v' - v^*\|_2^2 + \mu\|\theta' - \theta\|_2^2$$

Demonstration:



Conclusions

- A novel technique to improving interpretability during training;
- We simultaneously improve the interpretability of the learned features in deep neural networks and achieve better performance;
- Propose a prediction difference maximization algorithm;
- Introduce a human-in-the-loop learning procedure.