EVENT RECOGNITION BASED ON CLASSIFICATION OF GENERATED IMAGE CAPTIONS

Funded by

SAMSUNG

Authors: A.V. Savchenko, E.V. Miasnikov

Presenter: Andrey V. Savchenko
Dr. of Sci., PhD,
- Samsung-PDMI Joint AI Center
- Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics, Nizhny Novgorod, Russia

E-mail: avsavchenko@hse.ru

Outline of the talk

1. Event recognition in still images
2. Proposed Method based on Image captioning
3. Experimental results
4. Conclusion and future work

Event in image recognition

"An event captures the complex behavior of a group of people, interacting with multiple objects, and taking place in a specific environment. Images from the same event category may vary even more in visual appearance and structure" (Wang et al, IJCV 2018)

Children's birthday

Easter

Christmas

Halloween

Hiking

Road Trip

Skiing

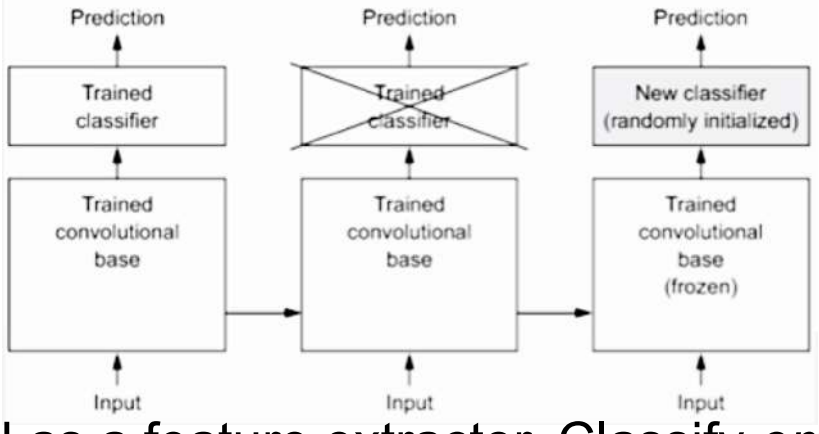
Problem formulation

Event recognition in single images: it is required to assign an input photo X from a gallery to one of C > 1 event categories (classes). Training set contains *N* reference images (examples) $\{X_n\}$, $n \in \{1, \ldots N\}$, with known class label $c_n \in \{1, \ldots C\}$

Key idea: Despite traditional usage of a CNN as a discriminative model in a classifier design, we propose to borrow **generative models to represent an input image in the other domain**.

Conventional approach: discriminative models

1) Fine-tune convolutional neural network (CNN) pre-trained on ImageNet, Places, etc.



2) pre-trained CNN as a feature extractor. Classify *embeddings (features)* from one of the last CNN's layers: *D*-dimensional feature vector $\mathbf{x}=[x_1,\ldots, x_D]$. Training set is associated with embeddings $\{\mathbf{x}_n\}$, $\mathbf{x}_n=[x_{n,1},\ldots, x_{n;D}]$.

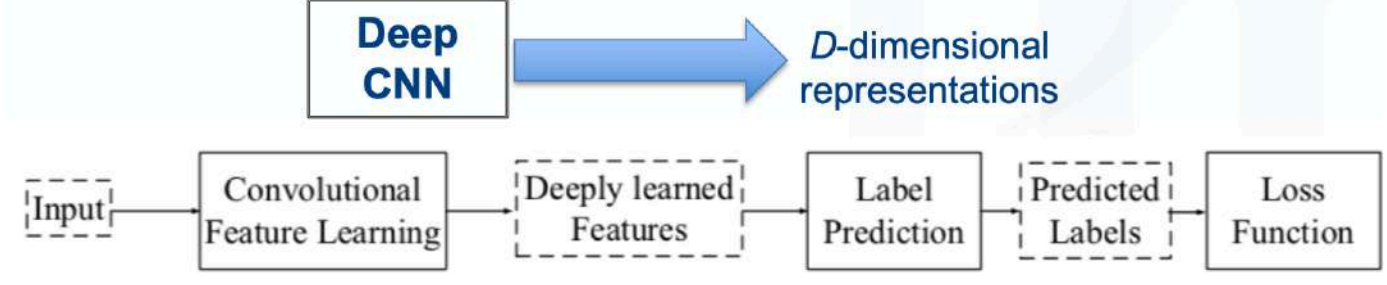


Image captioning: generate textual descriptions of images

by Joi Ito

the trail climbs steadily uphill most of the way.

by Danail Nachev

the stars in the night sky.

by Justin Higuchi

musical artist performs on stage during festival.

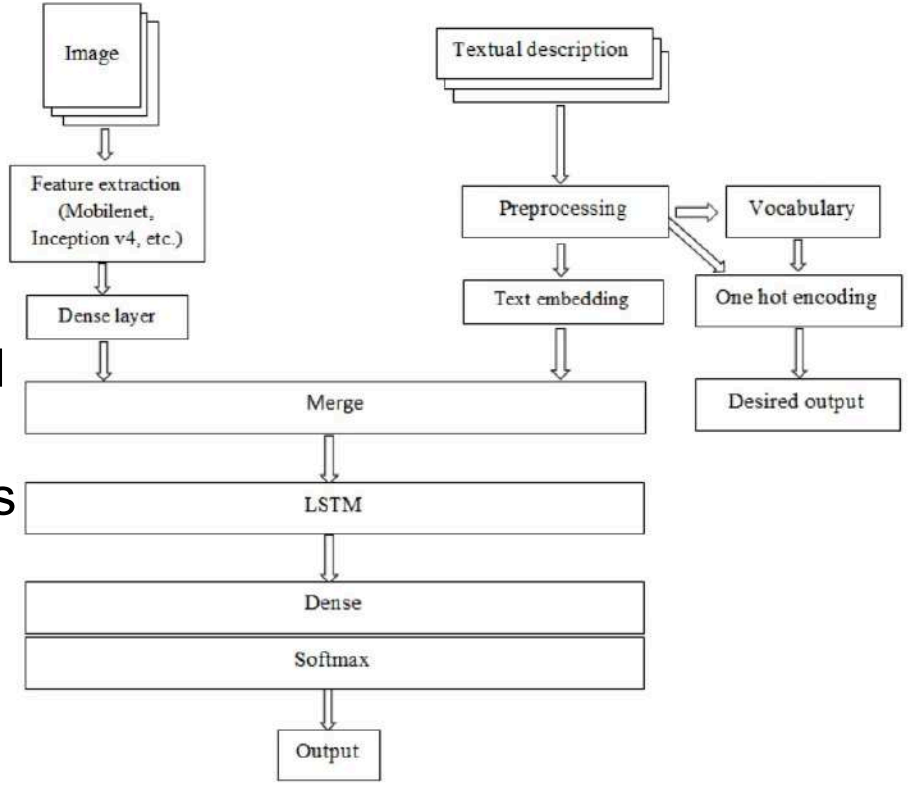
by Viaggio Routard

popular food market showing the traditional foods from the country.

Image captioning models

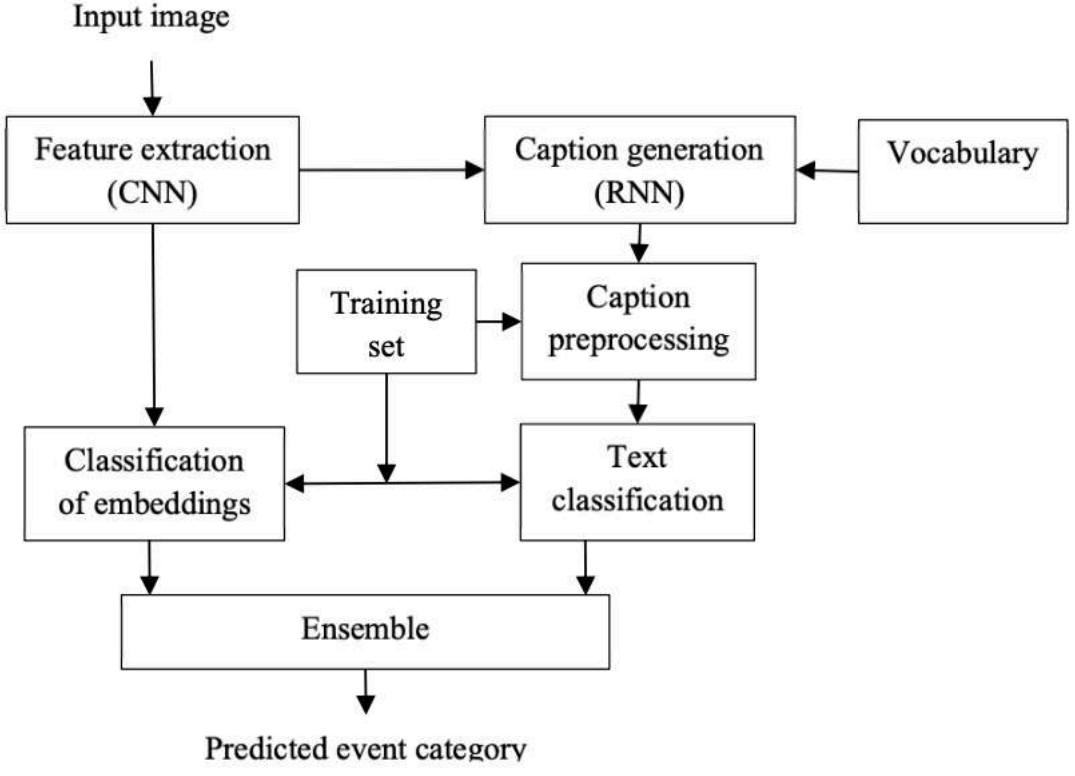
- Show and Tell/Show, Attend and Tell
- Neural Baby Talk
- Multimodal RNN
- **ARNet (Auto-Reconstructor Network)**

Embeddings **x** and previously generated words are fed to a special RNN-based neural network (generator) that produces the caption, which describes the input image. It is generated sequentially, word-by-word starting from t_0 =< START > token until a special t_{L+1} =< END > word is produced using the maximal output of Softmax layer.



Proposed pipeline

The textual descriptions of images generated by image captioning models can be fed to the input of a classifier in an ensemble in order to improve the event recognition accuracy of traditional methods.



Proposed pipeline

Caption is represented as a sequence of $L > 0$ tokens $t = \{t_0, t_1 ..., t_{L+1}\}$ from the vocabulary ($t_l \in V$, $l \in \{0,...,L\}$).
As event recognition task has nothing serial or temporal, obtained captions are one-hot encoded and summarized into a sparse feature vector suitable for the learning of an arbitrary classifier.

Input sequence
{1, 5,10,5,8,2}

Vectorized sequence
{1,1,0,0,1,0,0,1,0,1,0,…,0}

Motivation: the possibility to combine conventional CNNs with a completely different approach in an ensemble with high diversity.

$$\mathbf{p}_{ensemble} = [p_1, ..., p_C] = w \cdot \mathbf{p}_{emb} + (1 - w)\mathbf{p}_{txt}. \quad (1)$$

$$c^* = \underset{c \in \{1,...,C\}}{\operatorname{argmax}} p_c. \quad (2)$$

Qualitative results

a woman is doing a handstand at a local fair
PersonalSports (texts)
ReligiousActivity (embeddings)
PersonalArtActivity (ensemble)

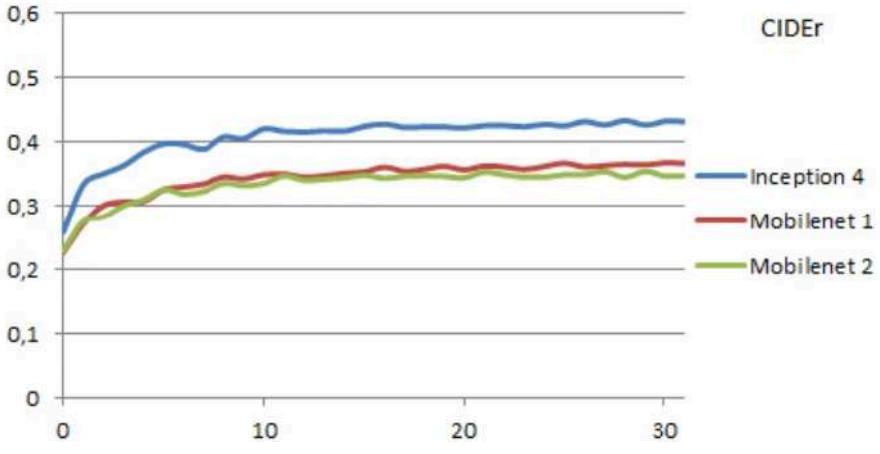
person , a painting by person
Museum (texts)
UrbanTrip (embeddings)
PersonalArtActivity (ensemble)

the statue of liberty and the moon
ThemePark (texts)
Christmas (embeddings)
ThemePark (ensemble)

the tower of the city
ThemePark (texts)
Architecture (embeddings)
ThemePark (ensemble)

Image captioning results

Training curve of CIDEr (consensus-based image description evaluation) – the word overlap between generated and reference captions



TensorFlow

Keras

Cross-dataset CIDEr for ARNet

Dataset used for training	Dataset used for estimation		
	Coco	Conceptual Captions	Flickr
Coco	1,086	0.166	0.33
Conceptual Captions	0.301	0.476	0.205
Flickr	0.193	0.073	0.446 (0.503)

Event classification results. Datasets

1. WIDER (Web Image Dataset for Event Recognition) with 50,574 images and $C = 61$ events (parade, dancing, meeting, press conference, etc.).

2. ML-CUFED (Multi-Label Curation of Flickr Events Dataset) contains $C = 23$ common event types.

Event recognition accuracy (%)

WIDER

Classifier	Features	Lightweight models	Deep models
SVM	Embeddings	48.31	50.48
	Objects	19.91	28.66
	Texts	26.38	31.89
	Proposed ensemble (1), (2)	48.91	51.59
Fine-tuned CNN	Embeddings	49.11	50.97
	Objects	12.91	21.27
	Texts	25.93	30.91
	Proposed ensemble (1), (2)	49.80	51.84
Baseline CNN [6]			39.7
Deep channel fusion [6]			42.4
Initialization-based transfer learning [4]			50.8
Transfer learning of data and knowledge [4]			53.0

ML-CUFED

Classifier	Features	Lightweight models	Deep models
SVM	Embeddings	53.54	57.27
	Objects	34.21	40.94
	Texts	37.24	41.52
	Proposed ensemble (1), (2)	55.26	58.86
Fine-tuned CNN	Embeddings	56.01	57.12
	Objects	32.05	40.12
	Texts	36.74	41.35
	Proposed ensemble (1), (2)	57.94	60.01

And summarizing our results we have the following conclusions

- In contrast to conventional fine-tuning of CNNs, we proposed to use image captioning, i.e., a generative model that converts images to textual descriptions.
- It is experimentally shown that the image captions can be classified more accurately than the features from an object detector.
- An ensemble of CNN and our approach provides state-of-the-art results for several event datasets.

Future work

impact on recognition accuracy arising from erroneous captions being generated should be examined

Thank you for your attention

Any Questions?