Image Captioning: Generation from Visual Input

Term Project CMPS185, Winter 2017
By: Diana Gonzalez Santillan

UNIVERSITY OF CALIFORNIA SANTA CRUZ

What is Image Captioning?

- Process of automatically assigning natural language descriptions to an image through machine learning and neural networks.
- Useful applications include creating tools for the visually impaired.
- Two approaches: Retrieval of pre-existing captions from database vs. natural language generation (NLG) from visual input.
- This report explores models for the NLG approach.

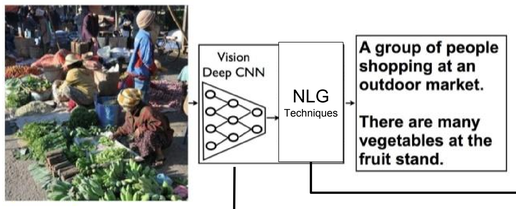


Fig. 1
Main outline of image caption generation from visual input. Vinyals et al., 2015.

Convolutional Neural Networks

- Query image is represented as a *width* x *height* x *RGB* volume and input into a series of convolutional layers.

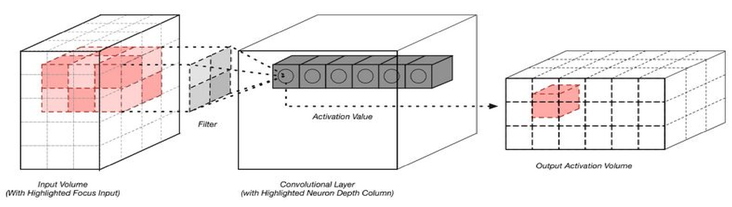


Fig. 2 Convolutional Layer: Input, filter, output. Gibson and Patterson (G&P), 2016.

- Each layer has filters that slide along the width & height of the input volume, computing a 2D activation map.
- Activation maps are stacked to create the output volume, which is input to another convolutional layer, or to a pooling layer that down-samples volumes.

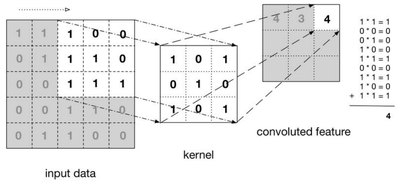


Fig. 3 Filter close up. G&P, 2016.

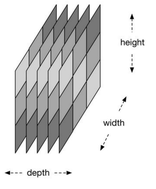


Fig. 4 Activation maps (output). G&P, (2016).

- Finally a fully connected layer is used to assign semantic class scores to the image, that is, determine which objects are most likely to be in the picture.

Conditional Random Fields (2011)

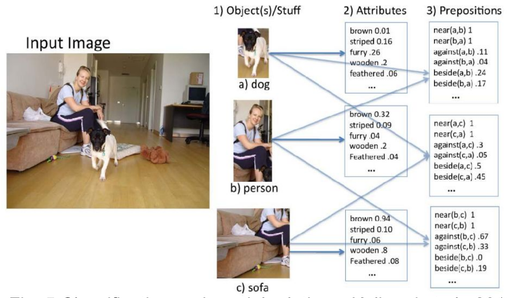
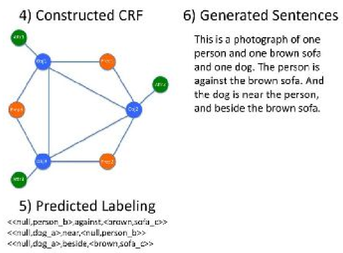


Fig. 5 Classification and spatial relations. Kulkarni et al., 2013.

- Objects and attributes are identified using convolutional neural networks.
- Spatial relations are found using prepositional relationship functions.



- A conditional random field (CRF) is created, it has attribute, object, and relationship nodes.
- *N*-gram probabilistic language models are used to predict best labeling.

Fig. 6 Generating the sentence. Kulkarni et al., 2013.

Visual Dependency Representations (2013)

- Objects are detected with convolutional neural networks and spatial relationships are found using a Visual Dependency Grammar (VDG).
- VDG quantifies how much relationship there is between two objects and uses this information to create the Visual Dependency Representation (VDR) tree.
- VDR is depth-first traversed to fill in slots of sentence templates which can be designed manually or with machine learning.

AUX	is, are	T_1	DT O_i AUX REL DT O_j. T_4?
DT	a, an	T_2	DT O_i AUX REL DT O_j REL DT O_k. T_4?
PRP	she, he, it, they	T_3	REL DT O_j.
REL	relationship	T_4	PRP AUX $\{\text{REL DT } O_i\}_{i=1}^{\lvert dependents \rvert}$.
O_i	object i in VDR		

Table 1 Sentence templates for VDR. Elliot and Keller, 2013.

Fig. 8 VDR based models. Elliot and Keller., 2013.

- STRUCTURE model: access to detected objects and VDR during training:

"A man is beside a woman above a horse, a horse is beside a woman beside a beach."

- PARALLEL model: in addition it has access to semantic trees of sample captions during training- verbs.

"A man is riding a horse above a beach, a horse is beside a beach beside a woman."

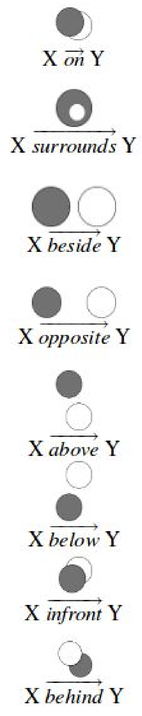


Fig. 7 VDG. Elliot and Keller, 2013.

Scene Graphs (2015)

In living room, there are two gray sofas next to each other and a table in front of them. There is a huge window in the back wall.

Fig. 9 Scene graph generated description. Lin et al., 2015.

Graph $G = (O, E)$ where
Vertices $O = \{o_1, o_2, \ldots, o_i\}$ are tuples
$o_i = (c_i, A_i)$ where
c_i = object class and
A_i = set of attributes of object i.
Edges $E = \{e_{ij}\}$ where
e_{ij} = relationship between o_i and o_j.

- Relationships are found with semantic grammars similar to VDR.
- Given a scene graph, a set of semantic trees is created with respect to the weights of each object. Example:

```
on-top-of(indet(color(box, red)), indet(table))
```

- Weights are calculated as a function of how likely the object is to be in the description, and the confidence that it is the "main actor" in the scene.
- Then each tree is traversed to fill in template slots in a similar process to the VDR models.

Conclusions

- The task of generating completely new natural language captions poses linguistical issues that can be avoided by choosing from a predefined set of captions.
- Retrieval methods find images in the database that are similar to the query images, rank, and sometimes combine their captions.
- This allows for a wider vocabulary and avoids linguistical issues.
- However, retrieval approaches need much larger training datasets to produce relevant output.

References

- Oriol Vinyals et al. "Show and tell: A neural image caption generator". http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html, 2015.
- Rafaella Bernardi et al. "Automatic description generation from images: A survey of models, datasets and evaluation measures". https://www.jair.org/media/4900/live-4900-9139-jair.pdf, 2016.
- Gibson, Adam, and Josh Patterson. "4. Major Architectures of Deep Networks." *Deep Learning*. N.p.: O'Reilly Media, 2016. N. pag. Safari Books Online.
- Girish Kulkarni et al. Babytalk: "Understanding and generating simple image descriptions". http://www.tamaraberg.com/papers/babytalk_pami.pdf, 2013.
- Desmond Elliot, and Frank Keller. "Image description using visual dependency representations". http://www.aclweb.org/anthology/D13-1128, 2013.
- Dahua Lin et al. "Generating multi-sentence natural language descriptions for indoor scenes". http://www.cs.toronto.edu/~urtasun/publications/lin_etal_bmvc15.pdf, 2015.