VIBIKNet: Visual Bidirectional Kernelized Network for the VQA Challenge

Marc Bolaños[1,2], Álvaro Peris[3], Francisco Casacuberta[3] and Petia Radeva[1,2]

(1) Universitat de Barcelona, (2) Computer Vision Center, (3) Universitat Politècnica de València

IEEE 2016 Conference on Computer Vision and Pattern Recognition

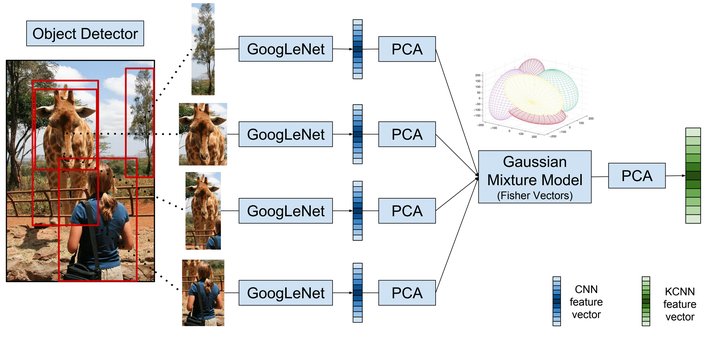
CVPR 2016

VQA Visual Question Answering

Introduction

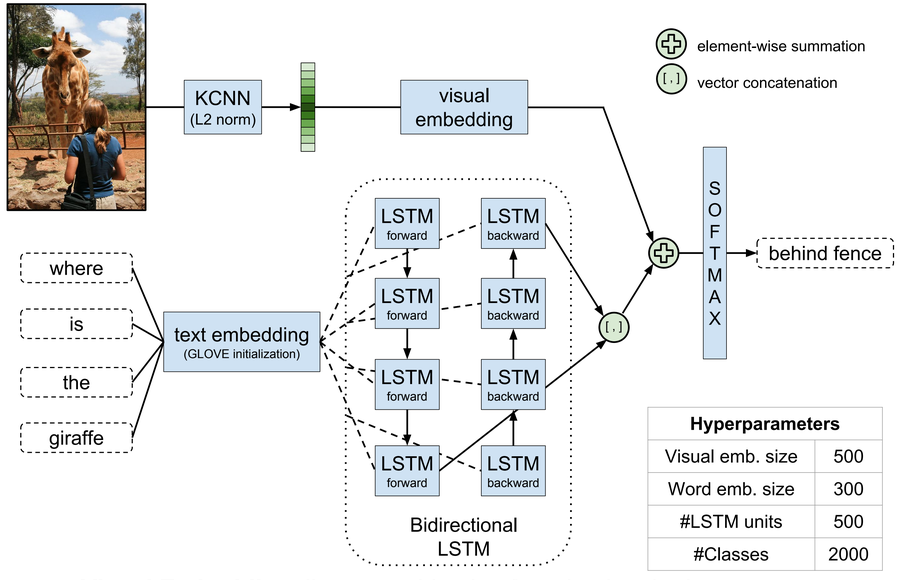
- VQA as a classification task.
- Image processed by a Kernelized CNN [2].
- Question processed by a Bidirectional LSTM.
- Multimodal combination using a classifier.

Kernelized CNN



Kernelized CNN [2] (KCNN): uses a set of object proposals and their rotations and learns a GMM for obtaining a final rotation-invariant Fisher Vector representation of the whole image.

VIBIKNet



Hyperparameters	
Visual emb. size	500
Word emb. size	300
#LSTM units	500
#Classes	2000

- **Visual Embedding**: linear combination for adapting the image representation to the dataset at hand.
- **Text Embedding with GLOVE initialization**: embedding matrix pre-trained using GLOVE [4] that is adapted during training.
- **BLSTM**: bidirectional representation (past-to-future and future-to-past) of the input question for a more robust representation.

Our architecture was inspired by the work in [1] and [3].

Results

Results								
Model	**Accuracy on dev 2014**				**Accuracy on test 2015**			
	Yes/No	**Number**	**Other**	**Overall**	**Yes/No**	**Number**	**Other**	**Overall**
LSTM	79.00	38.16	33.68	52.88	-	-	-	-
BLSTM	79.13	38.26	33.52	52.96	78.30	38.88	38.97	54.86
BLSTM train+dev	-	-	-	-	78.88	36.33	40.27	**55.77**

Examples

Question: What is the person holding?

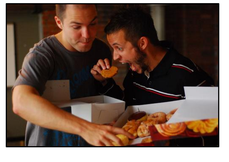
VIBIKNet: umbrella

Question: What character is on the backpack?

VIBIKNet: monkey

Question: What are the dogs doing?

VIBIKNet: playing frisbee



Question: What is the man holding near his mouth?

VIBIKNet: donut

Conclusions

- Robust image analysis.
- Full question context taken into account.
- Future work:
  - Attention mechanisms.
  - LSTM decoder for complex answers.

Download VIBIKNet

Bibliography

[1] M. Malinowski, M. Rohrbach, and M. Fritz. "Ask your neurons: A neural-based approach to answering questions about images," *Proc. IEEE ICCV*, 2015.
[2] Z. Liu, "Kernelized Deep Convolutional Neural Network for Describing Complex Images," *arXiv:1509.04581*. 2015.
[3] Á. Peris, M. Bolaños, P. Radeva, and F. Casacuberta. "Video Description using Bidirectional Recurrent Neural Networks," *arXiv:1604.03390.* 2016.
[4] J. Pennington, R. Socher and Ch.D. Manning. "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conf. on EMNLP,* 2014.

Contacts: Marc Bolaños: marc.bolanos@ub.edu, Álvaro Peris: lvapeab@prhlt.upv.es, Francisco Casacuberta: fcn@prhlt.upv.es, Petia Radeva: petia.ivanova@ub.edu

Acknowledgments: Partially supported by TIN2015-66951-C2-1-R, SGR 1219, PrometeoII/2014/030 and R-MIPRCV; P. Radeva by ICREA Academia. We acknowledge NVIDIA for a GPU donation.