Mean Box Pooling: A Rich Image Representation and Output Embedding for the Visual Madlibs Task

Ashkan Mokarian
ashkan@mpi-inf.mpg.de
Mateusz Malinowski
mmalinow@mpi-inf.mpg.de
Mario Fritz
mfritz@mpi-inf.mpg.de

Scalable Learning and Perception
Max Planck Institute for Informatics
Saarbrucken, Germany

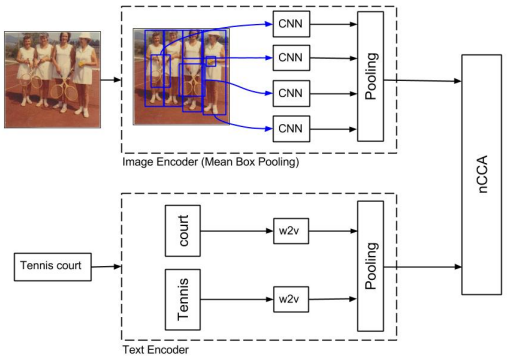

Figure 1: Overview of our full model, i.e. our proposed image representation using Mean Box Pooling, text encoding using average of Word2Vec representations, and normalized CCA for learning the joint space.

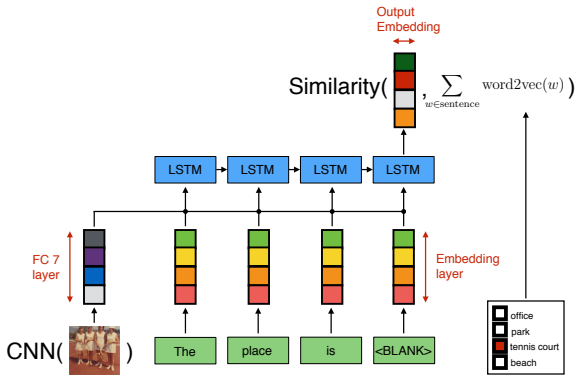

Figure 2: CNN+LSTM architecture that learns to choose the right answer directly in the embedding space. The output embedding is jointly trained with the whole architecture via backpropagation.

Question answering about real-world images is a relatively new research direction that requires a chain of machine visual perception, natural language understanding, and deductive capabilities to successfully come up with an answer on a question about visual content. In contrast to many classical Computer Vision problems such as recognition or detection, this task does not evaluate any internal representation of a method. In our experiments we consider the multi-choice Visual Madlibs dataset [2] as the ambiguities in the output space are rather minimal for this task.

In our paper, we present two novel architectures for question answering about real-world images. First, we argue for a rich image representation in the form of pooled CNN representations of highly overlapping object proposals, which we call Mean Box Pooling. Such a representation, allows for a more fine grained, multi-scale and multi-parts object analysis compared to global CNN representations while increasing the chances of finding relevant objects compared to object detectors. This representation, together with Normalized Canonical Correlation Analysis (nCCA) improves over the state-of-the-art on this dataset. The approach is depicted in Figure 1. Second, motivated by the popularity of deep architectures for visual question answering, which combine a global CNN image representation with an LSTM question representation, as well as the leading performance of nCCA on the multi-choice Visual Madlibs task, we propose a novel extension of the CNN+LSTM architecture, which we call Embedded CNN+LSTM, that chooses a prompt completion out of four candidates by doing comparisons directly in the embedding space at test time. This contrasts with the prior approach of [2] that uses a post-hoc comparison between the discrete output of the CNN+LSTM method and all four candidates. To achieve this, we jointly train LSTM together with the cosine similarity metric between the output embedding of the network and language representation of the ground truth completion. Such an approach integrates more tightly with the multi-choice filling the blanks task, and significantly outperforms the prior CNN+LSTM methods [1, 2]. This approach is depicted in Figure 2.

We evaluate our methods on the multiple choice task of the aforementioned Visual Madlibs dataset. In this scenario, a textual prompt together with 4 candidate completions is given. The purpose is to fill the blank symbol by a machine. The dataset is split into different category types such as scenes, affordances, emotions, etc. Interestingly, against our intuitions, a large number of highly overlapping proposals significantly helps for the task outperforming even MSCOCO ground truth bounding boxes. The results with a large number of proposals (100 proposals) and high threshold for suppressing non-maxima (NMS=0.75) are shown in Table 1. Although nCCA tops the leaderboard on the Visual Madlibs task, the largest body of work on the question answering about images combines a CNN with an LSTM. We hypothesis that the comparison for the multiple choice task should be directly done in the output embedding space. Our results, shown in Table 2, confirm our hypothesis.

[1] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015.

[2] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. Visual madlibs: Fill in the blank image generation and question answering. In *ICCV*, 2015.

	Easy Task		Hard Task	
	nCCA (ours)	nCCA [2]	nCCA (ours)	nCCA [2]
Scenes	86.2	**86.8**	69.0	**70.1**
Emotion	**52.5**	49.2	**39.4**	37.2
Past	**80.8**	77.5	**54.6**	52.8
Future	**81.1**	78.0	**56.1**	54.3
Interesting	**78.2**	76.5	**54.2**	53.7
Obj. attr.	**62.4**	47.5	**45.7**	43.6
Obj. aff.	**83.3**	73.0	**63.6**	63.5
Obj. pos.	**77.5**	65.9	**56.3**	55.7
Per. attr.	**56.0**	48.0	**44.2**	38.6
Per. act.	**83.0**	80.7	**65.5**	65.4
Per. loc.	**84.3**	82.7	**65.2**	63.3
Pair's rel.	**75.3**	63.0	**55.7**	54.3
Average	**75.0**	69.1	**55.8**	54.4

Table 1: nCCA (ours) uses the representation with object proposals. nCCA uses the whole image representation. Results in %.

	Easy Task		Hard Task	
	Embedded CNN+LSTM (ours)	CNN+LSTM [2]	Embedded CNN+LSTM (ours)	CNN+LSTM [2]
Scenes	**74.7**	71.1	**62.1**	60.5
Emotion	**36.2**	34.0	**34.3**	32.7
Past	**46.8**	35.8	**42.5**	32.0
Future	**48.1**	40.0	**41.4**	34.3
Interesting	**49.9**	39.8	**40.1**	33.3
Obj. attr.	**46.5**	45.4	**40.6**	40.3
Obj. aff.	**68.5**	-	**86.4**	-
Obj. pos.	**53.3**	50.9	**45.0**	44.9
Per. attr.	**40.7**	37.3	**40.0**	35.1
Per. act.	**64.1**	63.7	**53.7**	53.6
Per. loc.	**61.5**	59.2	**51.4**	49.3
Pair's rel.	**66.2**	-	**54.5**	-
Average	**54.7**	47.7	**49.3**	41.7

Table 2: Embedded CNN+LSTM (ours) measures a similarity between candidate answers and the architecture's output. Results in %.