Journal Homepage: -www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI:10.21474/IJAR01/1328
DOI URL: http://dx.doi.org/10.21474/IJAR01/1328
ISSN NO. 2320-5407

RESEARCH ARTICLE

IMPLEMENTATION OF ARTIFICIAL NEURAL NETWORKS IN MULTIMODAL IMAGE RECOGNITION AND TEXT.

[*]Dr.Geetha S[1] and Ashish Sharma[2].

1. Professor in School of Computer Science and Engineering, VIT University. Chennai, Tamil Nadu- India.
2. M. Tech student in School of Computer Science and Engineering, VIT University. Chennai, Tamil Nadu-India.

Manuscript Info

Manuscript History

Received: 12 June 2016
Final Accepted: 26 July 2016
Published: August 2016

Key words:-
Convolutional Neural Network, Deep learning, Image recognition,Recurrent Neural Network.

Abstract

Deep networks are already applied successfully to unsupervised feature learning for single modalities (e.g., text, images or audio). At First, rapid progress carried out in object detection has identified models that efficiently identify and label multiple regions of an image. Secondly, recent advances in image captioning have expanded the complexity of the label space from a permanent set of categories to sequence of words able to express significantly richer concepts. Here, we propose a unique application of deep networks to learn features over multiple modalities (image to text) i.e. image captioning. In this model of image captioning, Convolutional Neural Networks is applied to multiple regions of images followed by bidirectional Recurrent Neural Networks applied to sentences. Thus, the two sub-networks, CNN and RNN interact with each other in a multimodal layer to form the entire m-RNN model.This model is capable of learning long-term interactions. This arises from using a repeated visual memory that learns to reconstruct the visual features as new words are read or generated.

Introduction:-

Practically, image captions describe not only the objects but also image and their relationships. For humans even a glance at a picture is enough to indicate and describe large amount of details regarding visual scene[1]. However, this unique capability of humans has proved to be difficult to recognize our visual recognition models. The majority of previous work in visual recognition deals with labelling pictures with a set of visual classes and vast progress has been achieved in these areas[2,3].While vocabularies of visual ideas represent a suitable modeling assumption, they are extremely restrictive compared to the huge amount of descriptions that humans will compose.

Some brilliant approaches that address the question of generating image representation have been developed [4,5]. However,these models mostly depend on visual ideasand sentence templates, which urge limits on their selection. The backbone of these works has been on reducing complex visual scenes into one sentence, which comes under dense captioning.

In this work, we attempt towards the goal of generating dense descriptions of pictures (Figure 1). The main challenge regarding this goal is in the design of a model that is sufficient enough to reason about the contents of images and their description in natural language concurrently. Additionally, the model should be independent of

Corresponding Author:-Dr.Geetha S
Address:-Professor in School of Computer Science and Engineering, VIT University. Chennai, Tamil Nadu- India.

assumptions concerning specific hard-coded templates, rules or categories and in lieu depends on learning from the instructions. The second, real challenge is that datasets of image captions isaccessible in vast quantities on the net[6,7,8] but these descriptions multiplex mentions of manyentities whose locations in the picturesareunknown.

Our core insight is that we will grasp these enormous image-sentence datasets by treating the sentences as weak labels, in which adjacent segments of words correspond to somespecific, but unknown location in the image. Our approach is to conclude these alignments and use them to learn a related model of descriptions. Thus, our contributions are two fold:

Deep neural network model usually conclude the hidden alignment between segments of sentences and the region of the image that they relate. The two modalities are related through a typical, multimodal embedding space and a structured objective. We certify the effectiveness of this approach on image-sentence renewal experiments in that we tend to be superior the progressive approaches.

Multimodal Recurrent Neural Network model is introduced that takes input image and outputs its description in text. In this way the generated sentences significantly exceed retrieval based mostly baselines, and produce smart qualitative predictions.Thus train the model on the inferred results and appraise its performance on a new dataset of region-level annotations.

Model Architecture:-

Overview The ultimate goal of our model is to generate image captions. During training, the input to our model is a set of images and outputs their corresponding sentence descriptions. (Fig 2) This model aligns sentence snippets to the visual regions in such a way that they describe through a multimodal embedding. We then treat these correspondences as trainingknowledge for a second, multimodal Recurrent Neural Network model is a model that learns to generate the brief extract.

Learning to align visual and language data:-

The alignment model assumes an input dataset of pictures and their corresponding sentence descriptions. It means that sentences written by people build frequent references to somespecific, but unknown location in the image. We would prefer to infer these latent correspondences, with the eventual goal of later learning to generate these captions from image regions. The approach of Karpathy et al [9],is basically to ground dependency tree relations to image regions with a ranking objective. Our contribution is within the use of bidirectional recurrent neural network to compute word representations in the sentence, and get rid of the need to compute dependency trees and permitting unbounded interactions of words and their context within the sentence. We simplify their objective and show that each modifications improve ranking performance.

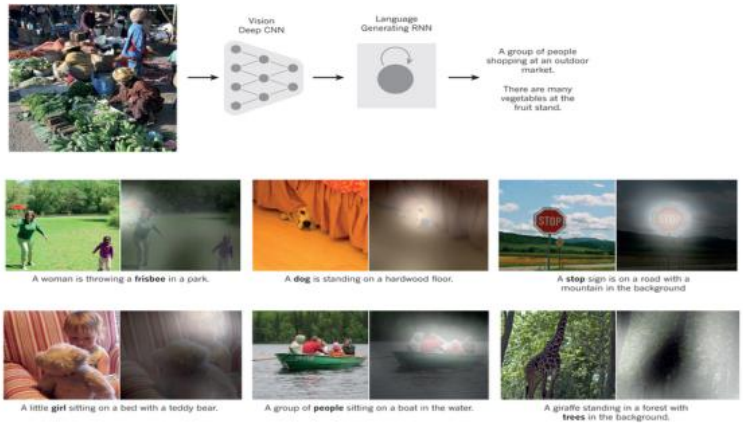

Fig. 1:- Image Captioning architecture flow

At first we describe neural networks that map words and image regions into a common, multimodal embedding. Then we introduce our unique objective, which learns the embedding representations therefore semantically similar ideas across the two modalities occupy near regions of the area.

Representing images:-

From the prior work [4,9] we observe that sentence descriptions build frequent references to objects and their attributes. Thus, we follow the methodology of Girshick et al. [10]to detect objects in each picture with a Region Convolutional Neural Network (RCNN). The CNN is pre-instructand fine-tuned on the 200 categories of the ImageNet Detection Challenge [2]CNN operates over volumes. All of these layers are going to take volumes of activation and they are going to produce volumes of activations. Volumes include spatial dimensions of width, height, and depth that will maintain through the computation. Here depth is not to be confused with depth of a network. This is just depth of the 3 dimension of a volume. Convolution layer is the building block of a convolution network. The convolution works by receiving some input volume[11] that goes into the layer and then we have all these filters in a convolution network. We have the filters to convolute it over this input volume. Thus we slide filter spatially through all spatial locations of this input volume and are going to compute dot products along the way

In (W^TX+b)suppose W is the filter and we are sliding them through spatially input volume and along the way as we are sliding this filter through we are competing W^TX +b. Here, X is a small piece of input volume.Now as we are sliding this filter through this volume spatially,we will end up with carving out an entire activation map of activations of responses of that filter in every single spatial position.Here stride refers how much we shift our filter at a time.

Output size:(N-F)/stride +1

Following Karpathy et al. [9] we use the prime 19 detected locations additionally to the total image and determine the representations supported the pixels $I_b$with in every bounding box as follows:

$$v = W_m[CNN_{\theta c} (I_b)]+ b_m(1)$$

where $CNN(I_b)$ transforms the pixels within bounding box I_b into 4096-dimensional activations of the absolutely connected layer just before the classifier. The CNN parameters $\theta_c$contain approximately 60 million parameters. The matrixW_m has size h×4096, where h is the size of the multimodal embedding area (h ranges from 1000-1600 in our experiments). Every image is so diagrammatic as a set of h-dimensional vectors $\{v_i \mid i = 1\ldots\ldots20\}$.

When we perform convolution neural network operations we won't shrink the volume size spatially, so we are preserving the spatial size. The reduction of the spatial size will we handled by pooling layer. Pooling layers take your input volume and they just squish and compress it spatially by doing a down sampling operation. This down sampling operation happens on every single activation map independently. We can perform max pooling or average pooling to shrink the size of activation map. In pooling layer ,we accept a volume of activation $W_1\times H_1\times D_1$.We produce a volume of activations $W_2\times H_2\times D_2$.Here we need to known the filter size and which stride to go as well. Here we are not changing the depth of the volume.

Representing sentences:-

The intermodal relationships can be established by representing the words within the sentence in the same h-dimensional embedding space as the image regions has taken. The simplest approach may be to project every individual word directly into this h-dimensional embedding. However, this approach does not think about any ordering and word context data in the sentence. An extension to this plan is to use word bigrams, or dependency tree relations as proposed [9]. However, this still imposes an unpredictable max size of the context window and needs the use of Dependency Tree Parsers which may be prepare on unrelated text corpora. Toaddress these relevant issues, we propose to use abidirectional Recurrent Neural Network (BRNN) [12] to calculate the word description. The BRNN takes a sequence of N words (encoded in a 1-of-k representation) and transforms all into an h-dimensional vector. However, the representation of every word is improved by a variably-sized context surrounding that word. Using the index t = 1 …N to denote the position of a word in a sentence, the precise form of the BRNN is as follows:

$$x_t = W_w \Pi_t(2)$$

$$e_t = f(W_e x_t + b_e)(3)$$

$$h_t^f = f(e_t + W_f h_{t-1}^f + b_f)(4)$$

$$h^b_t = f(e_t + W_b h^b_{t+1} + b_b)(5)$$

$s_t = f(W_d(h^f_t + h^b_t) + b_d)(6)$

Here, Π_t is an indicator column vector that has a single one at the index of the t-th word in a very word stock. The weights W_w specify a word embedding matrix which can be initialized with 300-dimensional word2vec [13] weights and keep stable due to overfittingconsiderations. However, in practice we notice very little amendment in final presentation once these vectors are instruct, even from random initialization. Note that the BRNN consists of two free streams of process, one moving left to right (h_t^f) and the other right to left (h_t^b). The final h-dimensional representation $s_t$for the t-th word could be a task of both the word at that location and near its close condition within the sentence. Thus, every s_t is a function of all words within the entire sentence, but our actual finding is that the final word representations (s_t) align most powerfully to the visual concept of the word at that location (Π_t). We experience the parameters W_f,W_b,W_d and the respective biases b_e,b_f,b_b,b_d. The actual size of the unrevealed representation in these experiments ranges between 300-600 dimensions. Thus we place the activation function f to the corrected linear unit(ReLU), which computes $f : x \rightarrow max(0,x)$.

Alignment objective:-

We have representthe transformations that map each image and sentence into a set of vectors during aunusal h-dimensional space. Since the direction is at the level of entire pictures and sentences, our strategy is to prepare an image-sentence score as a function of the independent region word scores. Basically , a sentence-image pair ought to have a high matching score if its words have a positive support in the image. The model of Karpathy et al. [9] interprets the dot product $v_i^T s_t$between the i-th section and t-th word as a function of similarity and use it to define the score between image k and sentence *l* as:

$S_{kl}=\Sigma_{t\epsilon gl}\Sigma_{t\epsilon gk} max(0,v_i^T s_t)(7)$

Here,$g_k$is the set of image extract in image *k* and $g_l$is the set of sentence extract in sentence *l*. The indices *k,l*range over the pictures and sentences within the instructions set.

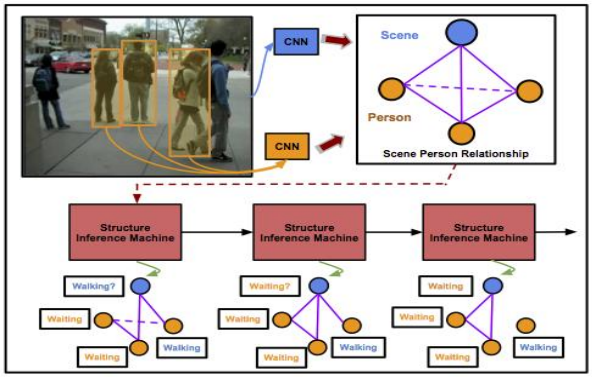

Fig2:- Convolution Neural Networks for Image Detection

Together with their supplementary Multiple Instance Learning objective, this score convey the presumption that a sentence fragment orient to a subset of the image section whenever the dot product is positive. We interpret that the following reformulation simplifies the model and increases the requirement for extra objectives and their hyper parameters.

$S_{kl}=\Sigma_{t\epsilon gl} max_{t\epsilon gk}(0,v_i^T s_t)(8)$

This objective stimulate oriented image-sentences pairs to have a higher score than misoriented pairs, by a margin.
It is very common to pad the input sometime with zero. In some cases if we pad the border as 1, then we get size output as input. The significance of zero padding is that if we take filters and we just start convolving on top of the image, the size just shrinks over time. But this is not nice property to have because by this we end up in a quick decrease of spatial size. We don't want to rapidly decrease the size of our representation because there are fewer numbers that are representing the original image and so to keep fixed size representation for convenience reasons and for representation reasons.

Multimodal Recurrent Neural Network forgenerating descriptions:-

In this section we presume input set of pictures and their textual descriptions. These could be full pictures and their sentence descriptions, or regions and text brief extract, as discribe in the previous section. The key problem is in the design of a model that may speculate a variable-sized sequence of outputs given in a picture. In previously developed language models based mostly on Recurrent Neural Networks (RNNs) [14,15,16] this is achieved by defining a likelihood distribution of next word in a very sequence given the present word and context from previous time steps. We search a straightforward however effective extension that in addition conditions the generative method on the content of input image. More formally, during practice our Multimodal RNN takes the image smallest element *I* and a sequence of input vectors *(x1…….xT*). It then assess a sequence of hidden states *(h1……ht*) and a sequence of outputs *(y1……yt)* by practicing the following recurrence relation for t = one to

$Tb_v = W_{hi}[CNN_{\theta c}(I)]$(9)

$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + 1(t = 1)\ \Theta\ b_v)$(10)

$y_t = \text{softmax}(W_{oh}h_t + b_o)$(11)

In the equations above,W_{hi},W_{hx},W_{hh},W_{ho}, $x_i$andb_h, b_o are learnable parameters, and $CNN_{\theta c}(I)$ is the last layer of a CNN. The output vector y_t holds the (unnormalized) log probabilities of words in the word stock and one further dimension for a special END representation. Note that we offer the image context vectorb_v to the RNN solely at the primary repetition of a process, which we have a tendency to found to work higher than at each time step. Basically we found that it will facilitate to pass each *bv; (Wh_xx_t)* through the execution function. A typical size of the masked layer of the RNN is 512 neurons.

RNN and Image Captioning:-

Recurrent Neural Networks is really taking place on language models and it is predicting next word in the sentence conditioned on both the previous words and image information. Recurrent networks are flexible models. It allows us to work with sequences at both inputs, output or both. Even if your input is fixed size,then it can still process it sequentially. It can also process fixed output sequentially. In image captioning,we are using recurrent neural network based language model in which we predict the next word in the sequence. We put the words in term of vectors that represents every word and then the target always are the next word and these are just targets and these are softmax classifiers which are back propagated and train over time. Training this on a lot of sentences would give us a language model and a way to imagine P(next word/previous words).

RNN can train all kinds of characters level or word level sequences. So, we will basically predicting a distribution over the next word given the sequence of previous words. P(next word/previous words).We are going to write down a set of equations that express a probability of next word given previous words. We are going to condition on these words and we are going to ask the network to tune its weights so that its predicting the correct next word. Characters at bottom and targets at top. Every single box here is a vector and every single arrow here indicates dependence. So, for example vector $y_o$is a function of vector h_0 and by 'a function of'' means this is fully connected layer. So we have a bunch of neurons connected fully to all the neurons in the previous volume. So these are all vectors and we are doing matrix vector multiplication to get from vector to vector. For example we are going to associate every single word with a 300 dimensional vector that represents the word. Every single y has 10001numbers (assuming 10000 word vocabulary+ 1). 1 since we want to have special end token that is basically a dot at the end of sentence. Probability of word and every single number tells us the likelihood of that particular word following the previous context. So the hidden representation mediates the context in this network. At every step of the way we are doing matrix multiplies times the input data.

We take the image to be tested and pipe it into the CNN.We got 4096 dimensional vector that represents the image on high level and after that we starts sampling words but that representation are feasting to the first hidden layer.It conditioned the generated process and then we continue to sample next word and so on until the network sample the end token.

From here CNN understand what it is of and then influence the generated process of RNN.We train this all end to end on data so that the sentences that are generated in that RNN are consistent with what is in the training data. CNNs are made up of neurons that have learnable weights and biases .We use the VGG-16 architecture for its state-

of-the-art performance. It consists of 13 layers of 3 × 3 convolutions interspersed with 5 layers of 2×2 max pooling. We eliminate the final pooling layer, so an input image of shape 3×W ×H gives rise to a variable of features of shape C×W0×H0 where C=512,W0 =[W/16],andH0 =[H /16]. The output of this network encodes the appearance of the image at a set of uniformly sampled image locations, and forms the input to the localization layer.

Object detection:-

In this model the visual processing module is a Convolutional Neural Network (CNN),which is a powerful model for visual recognition tasks. Instead of just classifying an image as some category labels, we also want to draw bounding box in the image to say where that class occurs. Another important topic is detection.So here there is again some fixed numbers of object categories but we actually want to find all instances of those categories inside the image and dropbox around them. The first convolution neural network consists of a filter bank of filters that we slide over the image which gives us the raw filter weight. The first application of these models to dense prediction tasks was introduced in R-CNN,where each region of interest was processed independently. Further work has focused on processing all regions with only single forward pass of the CNN,and on eliminating explicit region proposal methods by directly predicting the bounding boxes either in the image coordinate system, or in a fully convolutional and hence position invariant settings. Most related to our approach who develop a region proposal network (RPN) that regresses from anchors to regions of interest. Also,we replace their pooling mechanism with a differentiable, spatial soft attention mechanism. This change allows us to back propagate through region proposal network and train the whole model jointly.

Several recent approaches to Image Captioning rely on a combination of RNN language model conditioned on image information. A recent related approach is the use as of t attention mechanism over regions of the input image with every generated word. The approach to spatial attention is more general in that the network can process arbitrary regions in the image instead of only discrete grid positions. During generation the model where the visual information is only passed to the language model once on the first time step. At last, the metrics developed for the dense captioning task are inspired by metrics developed for image captioning.

Recognition Network:-

The recognition network is a fully-connected neural network that processes region features from the localization layer. The features from each region are flattened into a vector and passed through two full-connected layers, where each using rectified linear units. For each region this produces a code of dimension D = 4096 that compactly encodes its visual aspect. The codes for all positive regions are collected into a matrix of shape B×D and passed to the RNN language model.Also, we allow the recognition network one more chance to refine the confidence and position of each proposal region. It outputs a final scalar confidence of each proposed region and four scalars encoding a final spatial offset to be applied to the region programme. These two outputs are computed as a linear transform from the D-dimensional code for each region.

Results:-

Datasets:-

The datasets used are Flickr30K and MSCOCO datasets in our experiments. These datasets contain 31,000 and 123,000 images respectively and each image is explained with 5 sentences using Amazon Mechanical Turk. For Flickr30K, we use 1,000 images for validation, 1,000 for testing and for MSCOCO we use 5,000 images for both validation and testing.

Data Preprocessing:-

Weconvert all sentences to lowercase and discard other non-alphanumeric characters. Then filter words to those that occur at least 5 times in the training set. This results in 7414, and 8791 words for Flickr30K, and MSCOCO datasets respectively.

Learned region and word vector magnitudes:-

This model learns to modulate the magnitude of the region and word embedding. Due to their inner product interaction, representations of visually discriminative words have embedding vectors with higher magnitudes, which in turn translates to a higher influence on the image-sentence score.

Table:- I Comparison to other work by evaluation of full image predictions on 1,000 test images.B-n is BLEU score that uses up to n-grams. High is good in all columns. For future comparisons, our METEOR/CIDEr Flickr8K scores are 16.7/31.8 and the Flickr30K scores are 15.3/24.7.

	Flickr8K				**Flickr30K**				**MSCOCO 2014**					
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	METEOR	CIDEr
Nearest Neighbor	-	-	-	-	-	-	-	-	48.0	28.1	16.6	10	15.7	38.3
Mao et al. (2014)[17]	57	27	22-		542319		-		-	-	-	-	-	-
Google NIC (2014)[18]	6442		28	-	67.3	42.3	27.7	18.3	67.6	46.1	32.9	24.6	-	-
LRCN (2014)[19]	-	-	-	-	58.8	39.1	25.1	16.5	62.8	44.2	30.4	-	-	-
MS Research(2014)[20]	-	-	-	-	-	-	-	-	-	-	-	21.1	20.7	-
Chen and Zitnick et al(2014)[21]	-	-	-	14.1	-	-	-	12.6	-	-	-	19.0	20.4	-
Our model	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	19.5	66.0

References:-

1. L.Fei-Fei, A. Iyer, C. Koch, and P. Perona. 2007. What do we perceive in a glance of a real-world scene?*Journal of vision* 7(1):10.
2. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh,S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg ,and L. Fei-Fei. 2014 Imagenet large scale visual recognition challenge, 1, 2,3
3. M. Everingham, L. Van Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The pascal visual object classes (voc) challenge.June 2010International Journal of Computer Vision, 88(2):303– 338,.
4. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. 2011Baby talk: Understanding and generating simple image descriptions. In CVPR,. 1, 2, 3
5. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. 2010.Every picture tells a story: Generating sentences from images. In ECCV,1, 2
6. M. Hodosh, P. Young, and J. Hockenmaier.2013 Framing image description as a ranking task: data, models and evaluation metrics. Journal ofArtificial Intelligence Research. 1, 2, 5.
7. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL,. 1, 5
8. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick2014. Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312,. 1, 5
9. A. Karpathy, A. Joulin, and L. Fei-Fei. 2014Deep fragment embeddings for bidirectional image sentence mapping. arXiv preprint arXiv:1406.5679,. 2, 3, 4, 5, 6
10. R. Girshick, J. Donahue, T. Darrell, and J. Malik2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR,. 3
11. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei- Fei. 2009Imagenet: A large-scale hierarchical image database. In CVPR,. 3
12. M. Schuster and K. K. Paliwal. 1997Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on,.3
13. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In NIPS, 2, 3
14. T. Mikolov, M. Karafi´at, L. Burget, J. Cernock`y, and S. Khudanpur. 2010. Recurrent neural network based language model. In INTERSPEECH, 2, 4
15. I. Sutskever, J. Martens, and G. E. Hinton. 2011. Generating text with recurrent neural networks. In ICML, 2, 4, 8
16. J. L. Elman.1990. Finding structure in time. Cognitive science, 14(2):179–211,.4.
17. J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. 2014.Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090,. 2,6, 7
18. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan2014.Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555, 2, 5, 6, 7
19. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell.2014.Long-term recurrent convolutional networks for visual recognition and description. arXiv preprint arXiv:1411.4389, 2, 6, 7
20. H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Doll´ar, J. Gao, X. He, M. Mitchell, J. Platt, et al.2014.From captions to visual concepts and back. arXiv preprint arXiv:1411.4952, 2, 7
21. X. Chen and C. L. Zitnick. 2014. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654,. 2, 7.