International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2244-2254 www.ijaem.net ISSN: 2395-5252

Image Caption Generator Bot Based On Deep Neural Networks

Mr. Mukund Upadhyay, Ms. Shallu Bashambu

Submitted: 05-06-2021 Revised: 18-06-2021 Accepted: 20-06-2021

ABSTRACT:
In this project, we systematically analyze adeepneuralnetworksbasedimagecaptiongeneration method. With an image as the in-put, the method can output an English sen-tence describing the content in the image. Weanalyzethreecomponentsofthemethod:con-volutional neural network (CNN), recurrentneural network (RNN) and sentence genera-tion.By replacing the CNN part with threestate-of-the-art architectures, we find the VG-GNet performs best according to the BLEUscore.Wealsoproposeasimplifiedver-sion the Gated Recurrent Units (GRU) as anew recurrent layer,implementing by bothMATLAB and C++ in Caffe.The simplifiedGRU achieves comparable result when it iscompared with the long short-term memory(LSTM) method.But it has few parameterswhich saves memory and is faster in train-ing.Finally, we generate multiple sentencesusing Beam Search.The experiments showthat the modified method can generate cap-tions comparable to the-state-of-the-art meth-odswithlesstrainingmemory.

I. INTRODUCTION

Automatically describing the content of images us-ing natural languages is a fundamental and challeng-ing task.It has great potential impact.For exam-ple,itcouldhelpvisuallyimpairedpeoplebetterun-derstandthecontentofimagesontheweb.Also,it could provide more accurate and compact infor-mationofimages/videosinscenariossuchasimagesharing in social network or video surveillance sys-tems.Thisprojectaccomplishesthistaskusingdeep

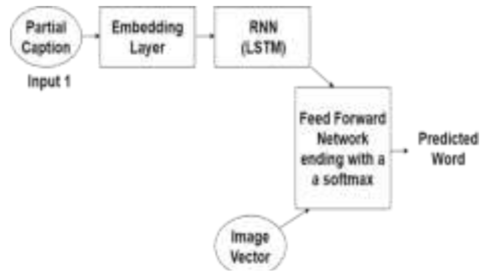


Figure 1: Image caption generation pipeline.The frameworkconsists of a convulitional neural netwok (CNN) followed by arecurrent neural network (RNN). It generates an English sen-tencefromaninputimage.

neural networks.By learning knowledge from im-age and caption pairs, the method can generate im-age captions that are usually semantically descrip-tiveandgrammaticallycorrect.
Humanbeingsusuallydescribeasceneusingnatural languages which are concise and compact.However,machinevisionsystemsdescribest hescene by taking an image which is a two dimensionarrays.Fromthisperspective,Vinyaletal.(Vinyalsetal.,)modelstheimagecaptioningproblemasa language translation problem in their Neural Im-age Caption (NIC) generator system.The idea ismapping the image and captions to the same spaceand learning a mapping from the image to the sen-tences.Donahue et al. (Donahue et al.,) proposedamoregeneralLong-termRecurrentConvolutionalNetwork (LRCN) method.The LRCN method notonly models the one-to-many (words) image cap-tioning,butalsomodelsmany-to-oneactiongenera-tion and many-to-many video description. They alsoprovides publicly available implementation based onCaffe framework (Jia et al., 2014), which furtherboosts the research on image captioning. This workisbasedontheLRCNmethod.

Although all the mappings are learned in an end-to-end framework, we believe the benefits of betterunderstanding of the system by analyzing differentcomponents separately.Fig.1 shows the pipeline.The model has three components. The first compo-nent is a CNN which is used to understand the con-tent of the image. Image understanding answers thetypical questions in computer vision such as "Whatare the objects?", "Where are the objects?"and"How are the objects interactive?".For example,the CNN has to recognize the "teddy bear", "table"and their relative locations in the image.The sec-ond component is a RNN which is used to generatea sentence given the visual feature.For example,the RNN has to generate a sequence of probabili-ties of words given two words "teddy bear, table".The third component is used to generate a sentenceby exploring the combination of the probabilities.This component is less studied in the reference paper(Donahueetal.,).

This project aims at understanding the impact ofdifferent components of the LRCN method (Don-ahueetal.,).Wehavefollowingcontributions:

- understand the LRCN method at the implemen-tationlevel.
- analyze the influence of the CNN componentbyreplacingthreeCNNarchitectures(twofrom author's and one from our implementa-tion).
- analyzetheinfluenceoftheRNNcomponentbyreplacingtwoRNNarchitectures.(onefromauthor's and one from our implementation).analyzetheinfluenceofsentencegenerationmethodbycomparingtwomethods(onefromauthor'sandonefromourimplementation).

II. RELATED WORK

Automaticallydescribingthecontentofanimageisa fundamental problem in artificial intelligence thatconnectscomputervisionandnaturallanguagepro-cessing.Earlier methods first generate annotations(i.e., nouns and adjectives) from images (Sermanetet al., 2013;Russakovsky et al., 2015), then gen-erate a sentence from the annotations (Gupta andMannem,).Donahueetal.(Donahueetal.,)devel-oped a recurrent convolutional architecture suitableforlarge-scalevisuallearning,anddemonstratedthe value of the models on three different tasks: videorecognition, image description and video descrip-tion.In these models, long-term dependencies areincorporated into the network state updates and areend-to-endtrainable.Thelimitationisthedifficultyofunderstandingtheintermediateresult.TheLRCNmethodisfurtherdevelopedtotextgenerationfromvideos(Venugopal anetal.,).

InsteadofonearchitectureforthreetasksinLRCN, Vinyalset al. (Vinyals et al.,) proposed aneural image caption (NIC) model only for the im-age caption generation. Combining the GoogLeNetandsinglelayerofLSTM,thismodelistrainedto maximize the likelihood of the target descrip-tion sentence given the training images.The per-formanceofthemodelisevaluatedqualitativelyandquantitatively. This method was ranked first in theMS COCO Captioning Challenge (2015) in whichthe result was judged by humans. Comparing LRCNwith NIC, we find three differences that may indi-cate the performance differences.First, NIC usesGoogLeNet while LRCN uses VGGNet.Second,NIC inputs visual feature only into the first unit ofLSTM while LRCN inputs the visual feature intoevery LSTM unit.Third, NIC has simpler RNNarchitecture (single layer LSTM) than LRCN (twofactored LSTM layers). We verified that the math-ematical models of LRCN and NIC are exactly thesame for image captioning.The performance dif-ferenceliesintheimplementationandLRCNhastotrade off between simplicity and generality, as it isdesignedforthreedifferenttasks.

Instead of end-to-end learning, Fang et al. (Fanget al.,) presented a visual concepts based method.First, they used multiple instance learning to trainvisual detectors of words that commonly occur incaptionssuchasnouns,verbs,andadjectives.Then,they trained a language model with a set of over400,000 image descriptions to capture the statis-tics of word usage.Finally, they re-ranked cap-tion candidates using sentence-level features and adeep multi-modal similarity model. Their captionshave equal or better quality 34% of the time thanthosewrittenbyhumanbeings.Thelimitationofthemethodisthatithasmorehumancontrolledparam-eterswhichmakethesystemlessre-producible.Webelievethe web application **captionbot**(Microsoft,)isbasedonthismethod.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2244-2254 www.ijaem.net ISSN: 2395-5252

Figure 2: This image shows a group of picture with their captions generated

Karpathyetal.(KarpathyandFei-Fei,)proposeda visual-semantic alignment (VSA) method.Themethod generates descriptions of different regionsof an image in the form of words or sentences (seeFig. 2). Technically, the method replaces the CNNwith Region-based convolutional Networks (RCNN)so that the extracted visual features are aligned toparticular regions of the image.The experimentshows that the generated descriptions significantlyoutperform retrieval baselines on both full imagesandonanewdatasetofregion-levelannotations.

ered from human beings using Amazon's Mechani-cal Turk (AMT). We manually checked some exam-ples by side-by-side comparing the image and cor-responding sentences.We found the captions arevery expressive and diverse. The COCO Captionis the largest image caption corpus at the time ofwriting. There are 413,915 captions for 82,783 im-agesintraining,202,520captionsfor40,504imagesin validation and 379,249 captions for 40,775 im-ages in testing. Each image has at least 5 captions.Thecaptionsfortrainingandvalidationarepubliclyavailable while the captions for testing is reserved bythe authors. In the experiment, we use all the train-ing data in the training process and 1,000 randomlyselectedvalidationdatainthetestingprocess.

III. DESCRIPTION OF PROBLEM

TaskIn this project, we want to build a system thatcan generate an English sentence that describes ob-jects,actionsoreventsinanRGBimage:

$S=f(I)$ (1)

where I is an RGB image and S is a sentence, f isthefunctionthatwewanttolearn.

CorpusWe use the MS COCO Caption (Chen etal.,2015)asthecorpus.Thecaptionsaregath-

where the θ is dropped for convenience, $S_t$is thewordatstept.

The model has two parts. The first part is a CNNwhich maps the image to a fixed-length visual fea-ture. The visual feature is embedded to as the inputvtotheRNN.

$v=W_v(CNN(I))$ (4)

where $W_v$is the visual feature embedding. The vi-sualfeatureisfixedforeachstepoftheRNN.

IntheRNN,eachwordisrepresentedaone-hotvector$S_t$ofdimensionequaltothesizeofthedic-tionary.$S_0$and $S_N$are for special start and stopwords.The word embeddingparameter isW_s:

$x_t=W_tS_t, t\in\{0\cdots N-1\}$ (5)

In this way,the image and words are mapped tothe same space. After the internal processing of theRNN, the features v, $x_t$and internal hidden param-eter$h_t$are decoded into a probability to predict thewordatcurrenttime:

p = LSTM(v,x,h),t∈ {0·t+1N−1}(6)Because ta sentence with higher probability doesnotnecessarymeanthissentenceismoreaccu-ratethanothercandidatesentences,post-processingmethodsuchas**BeamSearch**isusedtogenerate moresentencesandpicktop-Ksentences.

IV. METHOD

Forimagecaptiongeneration,LRCNmaximizestheprobabilityofthedescriptiongivingtheimage:
Thismethodgeneratesmorediverseandaccuratedescriptionsthanthewholeimagemethodsuchas

LRCN and NIC. The limitation is that the methodconsistsoftwoseparatemodels.Thismethodisfur-ther developed to dense captioning (Johnson et al.,2016)andimagebasedquestionandansweringsys-tem(Zhuetal.,2016).

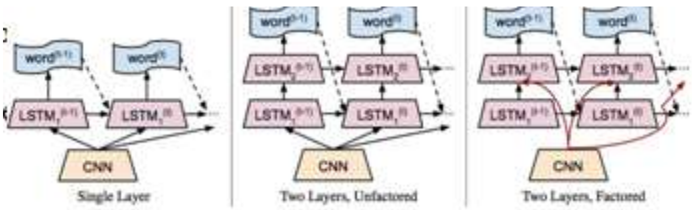
$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta) \quad (2)$$

where θ are the parameters of the model, I is an im-age, and S is a sample sentence. Let the length ofthesentencebeN,themethodappliesthechainrule tomodelthejointprobabilityover$S_0,\cdots,S_N$:

$$\log p(S|\mathbf{I}) = \sum_{t=0}^{N} \log p(S_t|\mathbf{I}, S_0,\cdots,S_{t-1}) \quad (3)$$

4.1 Convolutionalneuralnetwork

Inthisproject,aconvolutionalneuralnetwork(CNN)mapsanRGBimagetoavisualfeaturevec-tor. The CNN has three most-used layers: convolu-tion,poolingandfully-connectedlayers.Also,Rec-



chitecture.Themostrighttwo-layersfactoredLSTMisusedin

Figure 3: Three variations of the LRCN image captioning ar-themethod.Figurefrom(Donahueetal.,).

4.2 Recurrentneuralnetwork

To prevent the gradients vanishing problem, the longshort-term memory (LSTM) method is used as theRNN component.A simplified LSTM updates tified Linear Units (ReLU) f (x)=max(0, x) isused as the non-linear active function.The ReLUisfasterthanthetraditionalf(x)= tanh(x)orf (x)=(1 + $e^{-x})^{-1}$.Dropout layer is used topreventoverfitting.Thedropoutsetstheoutputof

fortimesteptgiveninputsx_t,h_{t-1},and$c_{t-1}$are:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$
$$g_t = \varphi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot \varphi(c_t)$$
(7)

each hidden neuron to zero with a probability (i.e.,0.5). The “dropped out” neurons do not contributeto the forward pass and do not participate in back-propagation.

The AlexNet(Krizhevsky et al., 2012), VGGNet(Simonyan and Zisserman, 2014) and GoogLeNet(Szegedyetal.,2015)arethreewidelyused deepconvolutionalneuralnetworkarchitecture.They share the convolutionpoolingfully-connectionloss function pipeline but with dif-ferent shapes and connections → of layers, especiallytheconvolutionlayer.AlexNetisthefirstdeepcon-volutional neural network used in large scale imageclassification.VGGNet and GoogLeNet achievesthe-start-of-the-artperformanceinImageNetrecog-nitionchallenge2014and2015. x

WhentheCNNcombinestheRNN,therearespe-cific considerations of convergence since both ofthemhasmillionsparameters.Forexample,Vinyalsetal.(Vinyalsetal.,)foundthatitisbettertofixtheparametersoftheconvolutionallayerastheparame-terstrainedfromtheImageNet.Asaresult,onlythenon-convolution layer parameters in CNN and theRNN parameters are actually learned from captionexamples.

where$\sigma(x)=(1+e^{-x})^{-1}and\varphi(x)=2\sigma(2x)$

1.Inadditiontoahiddenunith_t R^N, the LSTMincludesaninputgatei_t R^N, forget gate $f_t R^N$,outputgateo_t R^N,inputmodulationgateg_t R^N,andmemorycellc_t R^N.Thesead-ditionalcellsenabletheLSTMtolearnextremelycomplexandlong-termtemporaldynamics.Addi-tionaldepthcanbeaddedtoLSTMsbystackingthem ontopofeachother.Fig.3showsthree versionofLSTMs.Thetwo-layersfactoredLSTMachievesthebestperformanceandisusedinthemethod.

In this project, we proposed a simplified versionofGRUinsection5.1whichalsoavoidsthevanish-inggradientproblemandcanbeeasilyimplementedin Caffe based on the current Caffe LSTM frame-work.We also provide the MATLAB program intheAppendicesverifyingourderivationofBPTTontheoriginalGRUmodel.

4.3 Sentencegeneration

TheoutputofLSTMistheprobabilityofeachwordin the vocabulary. **Beam search** is used to generatesentences.Beam search is a heuristic search algo-rithm that explores a graph by expanding the mostpromisingnodeinalimitedset.Inadditiontobeam search, we also use k-best search to generate sen-tences.It is very similar to the time synchronousViterbi search. The method iteratively selects the kbestsentencesfromallthecandidatesentencesuptotimet,andkeepsonlytheresultingbestkofthem.

V. IMPLEMENTATION

are:

$$z=\sigma(U_z x_t+W_z s_{t-1}+b_z) r=\sigma(U_r x_t+W_r s_{t-1}+b_r)$$
$$h=\tanh(U_h x_t+W_h(s_{t-1} Ⓢ r)+b_h) s_t=(1-z) Ⓢ h+z Ⓢ s_{t-1}$$
(8)

PreprocessingBecause we want to keep the archi-tecture of the CNN, the input image are randomlycropped to the size of 224224. As a result, onlypart of the images are used in training at particulariteration. Because one image will be cropped mul-tiple times in the training, the CNN can probablyseethewholeimageinthetraining(onceforpartofthe image). However, the method only sees part oftheimageinthetestingexceptthedensecroppingisalso used (our project does not use dense crop). Forthe sentences, the method first creates a vocabularyonly from the training captions and removes lowerfrequencywords(lessthan5).Then,wordsarerep−-resentedbyone-hotvectors. ∈

∈ ∈

5.1 Caffearchitecture

Caffe(Jiaetal.,2014)providesamodifiableframe-workforthestate-of-the-artdeeplearningalgo-rithms.It is implemented using C++ and also pro-videsPythonandMATLABinterfaces.Caffemodel(network)definitionsarewrittenasconfiguration

where z is the update gate, r is the reset gate. sisusedasbothhiddenstatesandcellstates.Withfewerparameters,GRUcanreachacomparableper-formance to LSTM (Jozefowicz et al.,). To imple-ment GRU, we first wrote a MATLAB program tocheck our BPTT[2] gradient derivation. This is dueto the fact that automatic differentiation in Caffeis not supported at layer units level. Followed byour derivation, the calculated gradients only devi-ate from the numerical gradients by around 10^{-5}relatively.However, implementing GRU in Caffeis not straight forward since Caffe is based on acomplicated software architecture trying to provideconvenienceforassembling,notfurtherdeveloping.This is the bottleneck of GRU implementation. Wehave tried a number of implementations based on theoriginal GRU (Equ. 8), with no good results. Finallywe simplified GRU model inspired by the simplifiedSLTM in (Donahue et al.,). We omit the reset gateandaddatransfergatetomakeiteasilyfitintothecurrentCaffeLSTMframeworkas:

$z=\sigma(U_z x_t+W_z s_{t-1}+b_z)$

filesusingtheProtocolBufferLanguage[1]sothatthenetrepresentationandimplementationaresep-

$h=\tanh(U_h x_t +W_h s_{t-1} +b_h)$

(9)

arated.The separation abstracts from memory un-derlying location in CPU or GPU so that switchingbetweenaCPUandGPUimplementationisexactlybyonefunctioncall.However,theseparationmakestheimplementationlessconvenientaswewillshowinthenextparagraph.

5.2 SimplifyandimplementGRUinCaffe

In Caffe, a **layer** is the fundamental unit of com-putation.A**blob**isawrapperovertheactualdataproviding synchronization capability between CPUand GPU. We tried to implement the Gated Recur-rentUnits(GRU)(Choetal.,2014)inCaffe.TheGRU updatesfortimesteptgiveninputsx_t,s_{t-1}

1

$c_t=(1-z)Ⓢh+zⓈc_{t-1}$

$s_t=c_t$

Notethattheomittedresetgatewon'tbringbackthevanishing gradient problem which we see in tradi-tional RNN because we still have the update gate zacting as a weight between the previous state and thecurrent processed input. The added transfer gate, c,seemstobelessuseful,butitisactuallyveryimpor-tant for calculating the gradient in the framework.The parameter gradients in an RNN within a sin-glestep,t,dependsnotonlyon$\partial L_t/\partial s_t$,butalso $\partial L_t/\partial s_{t-i}$wherei=1,2,...,t.InCaffe,$\partial L_t/\partial s_t$ iscalculatedbyouterlayersautomatically,while $\partial L_t/\partial s_{t-i}$need to be calculated by inside layer unit.Toholdandtransferthesetwopartsofgradientsto

https://developers.google.com/

protocol-buffers/docs/proto

[2]Backwardpropagationthroughtime

Method	B-1	B-2	B-3	B-4

CNNs	layer	#param	memory	B-4
AlexNet	8	60	0.9	0.253
VGGNet	16	138	11.6	**0.294**
GoogLeNet	22	12	5.8	0.211

Table 1: Quantitative comparison of CNNs. The number ofparameter (#param) is in the unit of million, and the trainingmemory is in the unit of Gb. In experiment, we found that theBLEU 4 performance is positively related to the number of pa-rameters.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2244-2254 www.ijaem.net ISSN: 2395-5252

thenexttimestep,weuseanotherintermediatevari-

AlexNet +LSTM	0.650	0.467	0.324	0.221
AlexNet +GRU	0.623	0.433	0.292	0.194
VGGNet +LSTM	0.588	0.406	0.264	0.168
VGGNet +GRU	0.583	0.393	0.256	0.168

Table2:AlexNet,VGGNetwithdifferentRNNmodels.OurGRUmodelachievescomparableresultwiththeLSTMmodel,butwithlessparameterandtrainingtime.Thebeamsizeis1.

able, which is the added transfer gate c. This isjust an engineering issue that might not be avoidedwhile developing new models in Caffe. The theoryis always clear and concise (see Appendices for theMATLAB program verifying our BPTT derivationtotheoriginalGRU).

5.3 Trainingmethod

The neural network is trained using the mini-patchstochastic gradient descent (SGD) method. The baselearning rate is 0.01.The learning rate drops 50%in every 20,000 iterations.Because the number oftrainingsamplesismuchsmallerthanthenumberofparameters of the neural network, overfitting is ourbigconcern.Besidesthedropoutlayer,wefixedtheparameters of the convolutional layers as suggestedby(Vinyalsetal.,).Allthenetworkaretrained inaLinuxmachinewithaTeslaK40cgraphiccardwith12Gbmemory.

5.4 Quantitativeresult

Evaluation metrics We use BLEU (Papineni etal., 2002) to measure the similarity of the captionsgenerated by our method and human beings. BLEUisapopularmachinetranslationmetricthatanalyzesthe co-occurrences of n-grams between the candi-date and reference sentences.The unigram scores(B-1) account for the adequacy of the translation,whilelongern-gramscores(B-2,B-3,B-4)accountforthefluency.

DifferentCNNsTable1comparestheperfor-manceofthreeCNNarchitectures(theRNNpartuse LSTM). The VGGNet achieves the best perfor-mance (BLEU 4) and GoogLeNet has the lowestscore.It is out of our expectation at first becauseGoogLeNet achieves the best performance in the Im-ageNet classification task.We discussed this phe-nomenon with our fellows students.One of thempointed out that despite its slightly weaker classi-fication performance, the VGGNet features outper-formthoseofGoogLeNetinmultipletransferlearn-ing tasks (Karpathy, 2015). A downside of the VG-GNet is that it is more expensive to evaluate and ituses a lot more memory (11.6 Gb) and parameters(138 million). It takes more time to train VGGNetand GoogleNet than AlexNet (about 8 hours vs 4hours).

DifferentRNNsTable2comparestheperfor-manceofLSTMandGRU.TheGRUmodelachieves comparable results with less parametersandtrainingtime.

Different sentence generation methods Table 3also analyze the impact of beam size in the **BeamSearch** for different CNN architectures. In general,larger beam size achieves higher BLEU score. ThisphenomenonismuchmoreobviousintheVGGNet than other two CNNs.When the beam size is 1,AlexNetoutperformsVGGNet.Whenthebeamsizeis 10, the VGGNet outperforms AlexNet. The mostprobablereasonisthatAlexNetisgoodatdetecting a single or few objects in an image while VGGNetisgoodatdetectingmultipleobjectsinthesameim-age. When the beam size becomes larger, the VG-GNetbasedmethodcangeneratemoreaccuratesen-tences.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2244-2254 www.ijaem.net ISSN: 2395-5252

#beam	B-1	B-2	B-3	B-4
	AlexNet			
1	**0.650**	0.467	0.324	0.221
5	**0.650**	0.467	0.343	0.247
10	0.644	**0.474**	**0.347**	**0.253**
	VGGNet			
1	0.588	0.406	0.264	0.168
5	0.632	0.450	0.310	0.212
10	**0.681**	**0.513**	**0.390**	**0.294**
	GoogLeNet			
1	0.533	0.353	0.222	0.139
5	0.568	0.385	0.262	0.180
10	**0.584**	**0.410**	**0.292**	**0.211**

Table 3:AlexNet, VGGNet and GoogleNet with differentbeam sizes. Using AlexNet, the impact of the number of beamsize is not significant. Using the VGG net, the impact is signif-icant. Using the GoogLeNet net, the impact is moderate. Thebestscoresarehighlighted.

Method	B-1	B-2	B-3	B-4
LRCN	0.669	0.489	0.349	0.249
NIC	N/A	N/A	N/A	0.277
VSA	0.584	0.410	0.292	0.211
Thisproject	0.681	0.513	0.390	0.294

Table 4: Evaluation of image caption of different methods.LRCN is tested on the validation set (5,000 images).NIC istested on the validation set (4,000 images).VSA is tested onthe test set (40,775 images). This project is tested on the vali-dation set (1,000 images for B-1, B-2, B-3, and 100 images forB-4).

(VGGNet)Amanandwomansittingatatablewithapizza.
(GoogLeNet) A group of people sitting at a din-nertable.
When beam size is 5, the captions are as follows,(AlexNet) A group of people sitting at a table.(VGGNet)Amanandwomansittingatatablewith food.
(GoogLeNet) A group of people sitting at a din-nertable.
Whenbeamsizeis10,thecaptionsareasfollows,(Alex Net) A group of people sitting at a table.(VGGNet)Amanandwomansittingatatable.(GoogLeNet)Agroupofpeoplesittingatadin-nertable.
Fromtheresultlistedabove,wecanseethatwhenthe beam size is fixed, VGGNet can generate cap-tions with more details.When the beam size in-creases, the captions become short and detailed in-formationdisappears.
Although the sentence generated by our methodhas the highest probability, we don't know if thereare other sentences that can describe the image bet-ter.So we use 3-best search to explore the top 3captions.For Fig.4, the captions generated byGoogLeNetwithbeamsize5using3-bestsearcharelistedasfollows,
Agroupofpeoplesittingatadinnertable.
Agroupofpeoplesittingaroundadinnertable.Agroupofpeoplesittingatadinnertablewithplatesoffood.
The above captions are listed in probability de-scending order. We can see that the third sentence isactuallythebestone,althoughitdoesnothavethe

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2244-2254 www.ijaem.net ISSN: 2395-5252

Comparison with other systems Table 4 comparesBLEUscoresoftheresultsfromLRCN,NIC,VSA and this project.The BLEU score of the re-sultofthisprojectiscomparableorbetterthanthosefrom othersystemsalthoughourprojectistestedonlessdatas et(1,000images).

5.5 Qualitativeresult

TakingFig. 4 as an example, we analyzethe captions generated by AlexNet, VGGNet andGoogLeNet.

When beam size is 1, the captions are as follows,(AlexNet) A group of people sitting at a tablewithapizza.

highest probability.This is because when the sentence is long, it is more probable to make mistakes.So, sentences with high probability sometimes tendtobeshort,whichmaymisssomedetailedinforma-tion.However, it does not mean that the sentencewith the highest probability is bad.In most casesweobserved,sentenceswiththehighestprobabilit yare good enough to describe an image while longsentences often include redundant information andoftenmakegrammaticalmistakes.

Fig. 5 shows the good examples of the sentencesgeneratedbythisproject.Mostofthemsucces sfullydescribethemainobjectsandeventsinimages.Fi g.6showsfailedexamplesofthesystem.Theerrors

Task	Wenqiang	Minchen	Jianhui
CNN			100%
GRU		70%	30%
BeamSearch	100%		
Writing	40%	20%	40%

Table 5: Division of work. These only measure the implemen-tation and experimenting workload.All the analyses and dis-cussionsareconductedbyallofus.

Figure4: Sampleimageforqualitativeanalysis.

are mainly from object mis-detections such as anairplane is mis-detected as a kite (row 3 column 1),cellphones are detected as laptop (row 4 column 2).The generated sentences are also has minor grammarerror. For example, “A motorcycle with a motorcy-cle”(row4column3)ishardtounderstand.

VI. LESSONS LEARNED AND FUTURE WORK

This project provides a valuable learning experience.First,theLRCNmethodhasasophisticated pipelinesothatmodifyingpartofthepipelineiscomplicatedthan we expected.We learned how to use one ofthe most popular deep learning frameworks Caffethroughtheproject.

Second, mathematics and the knowledge of particular software architecture are equally importantfor the success of the project. Although we imple-mented the MATLAB version of GRU very earlybefore the deadline of the project, we spent a largeamount of time on implementing the GRU layer inCaffe. The benefit is that we learned valuable first-handexperienceonthedevelopinglevelofCaffein-

steadofpurelyusingexistinglayersinCaffe.
Third, working in a team, we could discuss andrefine a lot of initial ideas.We could also antici-pate problems that could become critical of the caseswe were working alone. Table 5 roughly shows theworkdivisionamongteammembers.

VII. EVALUATION

The project is successful. We have finished allthe goals before the deadline. The system cangenerate sentences that are semantically correctaccording to the image.We also proposed asimplifiedversionofGRUthathaslessparam-etersandachievescomparableresultwiththe

LSTMmethod.
Thestrengthofthemethodisonitsend-to-endlearning framework.The weakness is that itrequires large number of human labeled datawhich is very expensive in practice. Also, thecurrent method still has considerable errors inbothobjectdetectionandsentencegeneration.

VIII. CONCLUSION AND FUTURE WORK

Weanalyzedandmodifiedanimagecaptioningmethod LRCN. To understand the method deeply,wedecomposedthemethodtoCNN,RNN,andsen-tence generation. For each part, we modified or re-placed the component to see the influence on thefinal result.The modified method is evaluated ontheCOCOcaptioncorpus.Experimentresultsshowthat:firsttheVGGNetoutperformstheAlexNetandGoogLeNet in BLEU score measurement; second,the simplified GRU model achieves comparable re-sultswithmorecomplicatedLSTMmodel;third,in-creasing the beam size increase the BLEU score ingeneralbutdoesnotnecessarilyincreasethequalityofthedescriptionwhichisjudgedbyhumans.
Future work In the future, we would like to ex-plore methods to generate multiple sentences withdifferent content.One possible way is to combineinteresting region detection and image captioning.

REFERENCES

[1]. [Chen et al.2015] Xinlei Chen, Hao Fang, Tsung-Yi Lin,RamakrishnaVedantam,SaurabhGupta,PiotrDollar,and C Lawrence Zitnick.2015.Microsoft coco cap-tions:Data collection and evaluation server.arXivpreprintarXiv:1504.00325.
[2]. [Cho et al.2014] Kyunghyun Cho, Bart Van Merrienboer,CaglarGulcehre, DzmitryBahdanau, FethiBougares,Holger Schwenk, and YoshuaBengio.2014.Learn-ing phrase representations using rnn encoder-decoderforstatisticalmachinetranslation.arXivpreprintarXiv:1406.1078.
[3]. [Donahue et al.] Jeffrey Donahue, Lisa Anne Hendricks,SergioGuadarrama,MarcusRohrbach,SubhashiniVenugopalan,KateSaenko,and TrevorDarrell.Long-termrecurrentconvolutionalnetworksforvisualrecognitionanddescription.InIEEECVPR.
[4]. [Fang et al.] Hao Fang, Saurabh Gupta, Forrest Iandola,Rupesh K Srivastava, Li Deng, Piotr Dollar, JianfengGao, Xiaodong He, Margaret Mitchell, John C Platt,et al.From captions to visual concepts and back.InIEEECVPR.
[5]. [Gupta and Mannem] Ankush Gupta and Prashanth Man-nem. From image annotation to image description. InNeuralinformationprocessing.
[6]. [Jia et al.2014] Yangqing Jia, Evan Shelhamer, Jeff Don-ahue,SergeyKarayev,JonathanLong,RossGirshick,Sergio Guadarrama, and Trevor Darrell. 2014. Caffe:Convolutionalarchitectureforfastfeatureembedding.In Proceedings of the ACM International ConferenceonMultimedia,pages675–678.
[7]. [Johnson et al.2016] JustinJohnson,AndrejKarpathy,and Li Fei-Fei. 2016. Densecap: Fully convolutionallocalization networks for dense captioning.In IEEECVPR.
[8]. [Jozefowicz et al.] RafalJozefowicz, Wojciech Zaremba,andIlyaSutskever.Anempiricalexplorationofrecur-rentnetworkarchitectures.InICML.
[9]. [Karpathy and Fei-Fei] Andrej Karpathy and Li Fei-Fei.Deepvisual-semanticalignmentsforgeneratingimagedescriptions.InIEEECVPR.
[10]. [Karpathy2015] Andrej Karpathy.2015.Cs231n: Con-volutionalneuralnetworksforvisualrecognition.http://cs231n.github.io/.[Online;ac-cessed11-April-2015].
[11]. [Krizhevsky et al.2012] Alex Krizhevsky, Ilya Sutskever,and Geoffrey E Hinton. 2012. Imagenet classificationwithdeepconvolutionalneuralnetworks.InAdvancesinneuralinformationprocessingsystems.

[12]. [Microsoft] Microsoft.captionbot.https://www.captionbot.ai/.[Online;accessed22-April-2015].

[13]. [Papineni et al.2002] KishorePapineni,SalimRoukos,Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a methodfor automatic evaluation of machine translation.InProceedingsofthe40thannualmeetingonassociationforcomputationallinguistics,pages311–318.

[14]. [Russakovsky et al.2015] OlgaRussakovsky,JiaDeng,Hao Su,Jonathan Krause,Sanjeev Satheesh,SeanMa,ZhihengHuang,AndrejKarpathy,AdityaKhosla,Michael Bernstein, et al.2015.Imagenet large scalevisualrecognitionchallenge.IJCV,115(3):211–252.

[15]. [Sermanet et al.2013] Pierre Sermanet, David Eigen, Xi-angZhang,MichaëlMathieu,RobFergus,andYannLeCun.2013.Overfeat:Integrated recognition, lo-calizationanddetectionusingconvolutionalnetworks.arXivpreprintarXiv:1312.6229.

[16]. [Simonyan and Zisserman2014] KarenSimonyanandAndrew Zisserman.2014.Very deep convolutionalnetworks for large-scale image recognition.arXivpreprintarXiv:1409.1556.

[17]. [Szegedy et al.2015] ChristianSzegedy,WeiLiu,Yangqing Jia, Pierre Sermanet, Scott Reed, DragomirAnguelov, Dumitru Erhan, Vincent Vanhoucke, andAndrewRabinovich.2015.Goingdeeperwithconvolutions.InIEEECVPR.

[18]. [Venugopalan et al.] SubhashiniVenugopalan,MarcusRohrbach,JeffreyDonahue,RaymondMooney,TrevorDarrell,andKateSaenko.Sequencetosequence-videototext.InIEEEICCV.

[19]. [Vinyals et al.] Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan.Show and tell: A neu-ralimagecaptiongenerator.InIEEECVPR.

[20]. [Zhu et al.2016] Yuke Zhu, Oliver Groth, Michael Bern-stein, and Li Fei-Fei.2016.Visual7W: GroundedQuestionAnsweringinImages.InI