Caption Generator for Images Powered by Watson Visual Recognition

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | **INTRODUCTION** | | | |  | | |  |  | |
|  | * 1. **Overview**   Image caption generator is a process that involves computer vision and natural language processing concepts to recognize the objects or context of an image and describe them in a natural language like English. Image capturing plays an important role in automatic caption generation, which refers to the process of capturing images and generating textual description from an image based on the objects and actions in the image. The ability to recognize image features and generate accurate, syntactically reasonable text descriptions is important for many tasks in computer vision. Some of the important applications of this project are self-driving cars and aid to the physically challenged people. | | | | | | | | | |
|  | * 1. **Purpose**   The purpose of this project is to generate caption for the given images  Automatic driving is one of the biggest challenges and if we can properly caption the scene around the car, it can give a boost to the self-driving system. Also, this project will be useful to create a product for the physically challenged people, which will guide them travelling on the roads without the support of anyone else. | | | | | | | | | |
| **2** | **LITERATURE SURVEY** | | | |  | | |  |  | |
|  | **2.1 Existing problem**  Image captioning [1-3] has been proposed that seek to ground the words in the predicted caption to regions in the image. As the visual attention is often derived from higher convolutional layers of a CNN, the spatial localization is limited and often not semantically meaningful. Researchers also handle this job by obtaining image features with an object detector, and generating captions through an attention LSTM.  Yao et al. [4] introduced two graph convolutional networks to find the semantic relationship graph, and a spatial relationship graph that classifies the relationship between two boxes into 11 classes, such as “inside”, “cover”, or “overlap”.  The self-critical reinforcement learning for sentence generation [5] has also proven to be important for state-of-the-art captioning approaches, such as those above. Liu et al. in [6] proposed an alternative reinforcement learning approach over a visual policy that, in effect, acts as an attention mechanism to combine features from the image regions provided by an object detector. The visual policy, however, does not utilize spatial information about these image regions.  Recent developments in NLP, namely the Transformer architecture [7] have led to significant performance improvements for various tasks such as translation, text generation, and language understanding. | | | | | | | | | |
|  | **2.2 Proposed solution**  This project aims to build an application which takes input as image, analyses it and generate the captions in the form of speech. To achieve this, this project uses IBM Services like node-red service to build a web UI where user uploads a picture. This picture is analyzed by visual recognition service and the analyzed description is then converted in to text and speech by text to speech service. | | | | | | | | | |
| **3** | **THEORITICAL ANALYSIS** | | | |  | | |  |  | |
|  | **3.1 Block diagram**  Generating a description of an image is called image captioning. Image captioning requires recognizing the important objects, their attributes, and their relationships in an image. It also needs to generate syntactically and semantically correct sentences. Fig. 1 shows the system design of image captioning.  https://lh4.googleusercontent.com/Vw2z1Tm-ujyHlVCphHwHr_msZrWbwdrQgR51rzFQNxnFMu8nQtNzrJfUU3jAbtYaMnXAqwsoNmqBBWo-IbyMUbjMc_vU2V4nBo2ytUsgAqR232CEBy9PMuEbv_zXqlwHXhHkcKY  **Fig. 1 System design of image captioning** | | | | | | | | | |
|  | **3.2 Software designing**  This project uses the following services: i) IBM Watson Visual Recognition Service The IBM Watson Visual Recognition service uses learning algorithms to analyze images for content such as objects, scenes, and faces. This service is used to create a visual recognition modeler which is used to automatically train a model to classify images for scenes, objects, or your custom content.  ii) Text to speech services  The Text to Speech service converts text to natural-sounding speech. This service streams the synthesized audio back with minimal delay. The audio uses appropriate cadence and intonation for its language and dialect to provide voices that are smooth and natural. The service can be used in applications such as voice-automated chatbots, as well as a variety of voice-driven and screenless applications, such as tools for the disabled or visually impaired, video narration and voice over, and educational and home-automation solutions.  iii) Node-red service  The node-red service provides a pre-configured Node-RED application, including a Cloudant service to store the application flow configuration. This is used to add services, generate and download the code, use the IBM Cloud Developer Tools CLI to run and debug locally, then deploy to Cloud Foundry or a DevOps Pipeline. | | | | | | | | | |
| **4** | **EXPERIMENTAL INVESTIGATIONS**  This project works the following two steps: (i) converting the scene into text and (ii) converting the text to voice. Both are now famous applications of Deep Learning. Initially, the visual recognition model generates class keywords that describe the image. Then, the text to speech service converts written text to natural-sounding speech. | | | | | | | | | |
| **5** | **FLOWCHART**  The Fig. 2 shows the node-red flow of the image caption generation.    **Fig. 2 Node-red flow of image captioning** | | | | | | | | | |
| **6** | **RESULT**  The Fig. 3 and Fig. 4 show the input and output screens.    **Fig. 3 Input screen**    **Fig. 4 Output screen** | | | | | | | | | |
| **7** | **ADVANTAGES & DISADVANTAGES**  The visual recognition model identifies and generates class keywords that describe the image. Then, the text to speech service converts generated text to natural-sounding speech. But, it is required to improve the performance of the system. | | | | | | | | | |
| **8** | **APPLICATIONS**   * Self-driving cars — Automatic driving is one of the biggest challenges, and if it can properly caption the scene around the car, it will give a boost to the self-driving system. * Aid to the blind — This motivated to create a product for the blind which will guide them travelling on the roads without the support of anyone else.   The above two applications can be done by first converting the scene into text and then the text to voice. Both are now famous applications of Deep Learning. | | | | | | | | | |
| **9** | **CONCLUSION**  The project caption generation for images powered by Watson visual recognition converts the scene into the text using visual recognition service, and the generated text to speech by Watson text to speech service. | | | | | | | | | |
| **10** | **FUTURE SCOPE**  CCTV cameras are everywhere today, but along with viewing the world, if we can also generate relevant captions, then we can raise alarms as soon as there is some malicious activity going on somewhere. This could probably help reduce some crime and/or accidents. | | | | | | | | | |
| **11** | **BIBILOGRAPHY**   1. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057, 2015. 2. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4651–4659, 2016. 3. Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov. Review networks for caption generation. In Advances in Neural Information Processing Systems, pages 2361–2369, 2016. 4. T. Yao, Y. Pan, Y. Li, and T. Mei. Exploring visual relationship for image captioning. In European Conference on Computer Vision, pages 684–699, 2018. 5. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7008–7024, 2017. 6. D. Liu, Z.-J. Zha, H. Zhang, Y. Zhang, and F. Wu. Context-aware visual policy network for sequence-level image captioning. In Proceedings of the 26th ACM International Conference on Multimedia, MM ’18, pages 1416–1424. ACM, 2018. 7. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017. | | | | | | | | | |
|  | **APPENDIX** | | | |  | | |  |  | |
|  | | Source Code: | | | | | | | |
|  | |  |  |  | |  |  | | |

|  |
| --- |
| [{"id":"aa14204.f27c3e","type":"tab","label":"TJBot Camera","disabled":false,"info":"" }, {"id":"b8c01f09.8bb37","type":"debug","z":"aa14204.f27c3e","name":"Print msg.result.images","active":true,"tosidebar":true,"console":false,"complete":"result.images","statusVal":"","statusType":"auto","x":590,"y":260,"wires":[]},{"id":"ce52d36.e13123","type":"comment","z":"aa14204.f27c3e","name":"Reload picture into Node-RED Dashboard","info":"<script>\n window.location.reload(false); \n</script>","x":1200,"y":40,"wires":[]},{"id":"a5c23424.28f5d8","type":"debug","z":"aa14204.f27c3e","name":"What did Watson find?","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"template","targetType":"msg","statusVal":"","statusType":"auto","x":900,"y":360,"wires":[]},{"id":"335b3735.c599f8","type":"ui\_template","z":"aa14204.f27c3e","group":"2637150a.7d26da","name":"","order":5,"width":"9","height":"12","format":"<div ng-bind-html=\"msg.template\"></div>","storeOutMessages":true,"fwdInMessages":true,"resendOnRefresh":false,"templateScope":"local","x":860,"y":280,"wires":[[]]},{"id":"6d921df6.a91924","type":"visual-recognition-v3","z":"aa14204.f27c3e","name":"","vr-service-endpoint":"https://api.us-south.visual-recognition.watson.cloud.ibm.com/instances/bf92b316-5e86-408b-9a4e-b600fee29361","image-feature":"classifyImage","lang":"en","x":330,"y":300,"wires":[["62828679.3484f8","b8c01f09.8bb37"]]},{"id":"62828679.3484f8","type":"function","z":"aa14204.f27c3e","name":"Process Results","func":"if (typeof msg.result == 'undefined') {\n return null;\n}\n\n// Text Extraction\nif (typeof msg.result.images[0].text != 'undefined') {\n var image\_text = msg.result.images[0].text;\n msg.payload = image\_text;\n msg.template = image\_text;\n if( image\_text.length >0 ) {\n msg.template= \"Watson found the words: \"+image\_text;\n }\n return msg;\n}\n\nvar bestcolor = -1;\nvar colorscore = 0;\nvar c\_id = 0;\nvar say = \"\";\nvar item;\n\nfor ( c\_id=0; c\_id < (msg.result.images[0].classifiers.length); c\_id++ ){\n // find the best color, if any\n for( i =0; i<(msg.result.images[0].classifiers[c\_id].classes.length); i++ ){\n var object = msg.result.images[0].classifiers[c\_id].classes[i].class;\n if ( object.includes(\"color\") ) {\n if( msg.result.images[0].classifiers[c\_id].classes[i].score > colorscore){\n bestcolor = i;\n colorscore = msg.result.images[0].classifiers[c\_id].classes[i].score;\n }\n }\n }\n \n var bestItem = 0;\n var itemScore = 0;\n for( i =0; i<(msg.result.images[0].classifiers[c\_id].classes.length); i++ ){\n var object = msg.result.images[0].classifiers[c\_id].classes[i].class;\n if ( !object.includes(\"color\") ) {\n if( msg.result.images[0].classifiers[c\_id].classes[i].score > itemScore){\n bestItem = i;\n itemScore = msg.result.images[0].classifiers[c\_id].classes[i].score;\n }\n }\n }\n \n if( bestcolor != \"-1\") {\n // found a color\n item = msg.result.images[0].classifiers[c\_id].classes[bestcolor].class + \" \" + msg.result.images[0].classifiers[c\_id].classes[bestItem].class;\n bestcolor = -1;\n } else {\n item = msg.result.images[0].classifiers[c\_id].classes[bestItem].class;\n }\n// say = say + \" Watson's \" + msg.result.images[0].classifiers[c\_id].name + \" classifier thinks this picture contains a \" + item +\".\";\n say = say + \" Watson thinks this picture contains a \" + item +\".\";\n}\nmsg.payload = say;\n\nvar picInfo = msg.result.images[0].classifiers[0].classes;\nvar arrayLength = picInfo.length;\n\nmsg.template=\"<style>h4 { text-align: center; margin: 10px; }\";\nmsg.template=msg.template+\"table { width: 440px; margin-top: 10px; }\";\nmsg.template=msg.template+\"th, td { padding: 8px; text-align: left; border-bottom: 1px solid #ddd; background-color: #FFFFFF; width: 50%;}\";\nmsg.template=msg.template+\".classifier {background-color: rgb(85,150,230);text-align: center;}\";\nmsg.template=msg.template+\".title { background-color:LightGrey;}</style>\";\n\nmsg.template=msg.template+\"<h2>\"+say+\"</h2><table span=100%><tr><th>Class</th><th>Confidence</th></tr>\";\nfor (var i = 0; i < arrayLength; i++) {\n msg.template = msg.template + \"<tr><td>\" + picInfo[i].class + \"</td><td>\" + picInfo[i].score + \"</td></tr>\";\n}\nmsg.template = msg.template + \"</table>\"\n\n\nreturn msg;","outputs":1,"noerr":0,"initialize":"","finalize":"","x":560,"y":340,"wires":[["335b3735.c599f8","a5c23424.28f5d8","ffbd7138.638a6","fba47e02.208bc"]]},{"id":"98289d2e.d7d9c","type":"comment","z":"aa14204.f27c3e","name":"Take Photos from a Raspberry Pi Camera","info":"If you add a Raspberry Pi Camera to your TJBot,\nhe can \"see\" and take Photos through his \n\"eye\" cutout.\n\nThis Node-RED flow requires the\nnode-red-contrib-camerapi npm package.\n\nDon't forget to enable the Camera in the \nRaspberry Pi Configuration utility.","x":200,"y":38,"wires":[]},{"id":"fba47e02.208bc","type":"watson-text-to-speech","z":"aa14204.f27c3e","name":"","lang":"en-US","langhidden":"en-US","langcustom":"NoCustomisationSetting","langcustomhidden":"","voice":"en-US\_AllisonVoice","voicehidden":"en-US\_AllisonVoice","format":"audio/wav","password":"Mahesh18","apikey":"EGeT\_Bs96j3xZpSQfn5HeQe-WFBvVviGEnEXgyBqTOy9","payload-response":true,"service-endpoint":"https://api.eu-gb.text-to-speech.watson.cloud.ibm.com/instances/f17de8c5-fc70-4b6f-b715-2619bc233deb","x":880,"y":480,"wires":[["8415ec9d.561ec"]]},{"id":"ffbd7138.638a6","type":"debug","z":"aa14204.f27c3e","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"payload","targetType":"msg","statusVal":"","statusType":"auto","x":870,"y":420,"wires":[]},{"id":"54671aa1.a7c834","type":"ui\_template","z":"aa14204.f27c3e","group":"12b04aca.de3335","name":"","order":1,"width":"9","height":"8","format":"<!DOCTYPE html>\n<html lang=\"en\">\n<head>\n <title>Visual Recognition</title>\n <meta charset=\"utf-8\">\n <meta name=\"viewport\" content=\"width=device-width, initial-scale=1\">\n <link rel=\"stylesheet\" href=\"https://maxcdn.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css\">\n <script src=\"https://ajax.googleapis.com/ajax/libs/jquery/3.4.1/jquery.min.js\"></script>\n <script src=\"https://cdnjs.cloudflare.com/ajax/libs/popper.js/1.14.7/umd/popper.min.js\"></script>\n <script src=\"https://maxcdn.bootstrapcdn.com/bootstrap/4.3.1/js/bootstrap.min.js\"></script>\n <style>\n .bg-light {\n \n background-color: #A42156!important;\n\t}\n\tdashboard-template h4 {\n color: #111111 !important;\n background-color: #A42156 !important;\n }\n\t\n\t.custom-file {\n\t\tmargin-bottom: 14px;\n\t\t}\n\t\t\n\t.table .thead-dark th {\n\t\tcolor: #fff;\n\t\tbackground-color: #A42156;\n\t\tborder-color: #A42156;\n\t}\n\th5{\n\t text-align: center;\n\t\t color: #EB9E30;\n\t\t}\n\t\t.text-center {\n\t\t text-align: center;\n\t\t}\n\t.imgdiv\n\t{\n\t\talign:center;\n\t}\n\t</style>\n</head>\n<body>\n\t<nav class=\"navbar navbar-expand-sm bg-light\">\n\t\t<div class=\"justify-content-center\">\n\t\t\t<h4 class=\"text-center\">Caption Generator Board </h4>\n\t\t</div>\n\t</nav>\n\t<br><br>\n <div class=\"container\">\n\t\t<div class=\"row\">\n\t\t\t<div class=\"col-sm-2\">\n\t\t\t</div>\n\t\t\t<div class=\"col-sm-8\">\n\t\t\t\t\n\t\t\t\t\t<div class=\"custom-file\">\n\t\t\t\t\t\n\t\t\t\t\t\t<input type=\"file\" name=\"pic\" accept=\"image/\*\" onchange=\"readURL(this);\" class=\"custom-file-input\" id=\"customFile\">\n\t\t\t\t\t\t\n\t\t\t\t\t\t<label class=\"custom-file-label\" for=\"customFile\">Choose file</label>\n\t\t\t\t\t</div>\n\t\t\t\t\t<br>\n\t\t\t\t\t<div class=\"imgdiv\">\n\t\t\t\t\t\t<img src=\"#\" id=\"blah\" class=\"rounded\" alt=\"Selected Image\">\n\t\t\t\t\t</div>\n\t\t\t\t\t<!--<md-button ng-click=\"send({payload:action()})\">\n\t\t\t\t\t\tPredict\n\t\t\t\t\t</md-button>-->\n\t\t\t\t\t<button type=\"submit\" ng-click=\"send({payload:action()})\" class=\"btn btn-success\">Submit</button>\n\t\t\t\t\n\t\t\t\t\n\t\t\t</div>\n\t\t\n</body>\n</html>\n<script>\nvar x=\"\";\n function readURL(input) {\n if (input.files && input.files[0]) {\n var reader = new FileReader();\n\n reader.onload = function (e) {\n $('#blah')\n .attr('src', e.target.result)\n .width(150)\n .height(200);\n };\n\n reader.readAsDataURL(input.files[0]);\n x= input.files[0]\n\n \n \n\n }\n }\n function getdata(data)\n\t{\n\t\t var html = '';\n if(data != 0)\n\t\t\t{\n\t\t\t\t$.each(data, function(i){\n\t\t\t\tvar row = data[i];\n\t\t\t\tconsole.log(row);\n\t\t\t\thtml += '<tr>';\n\t\t\t\thtml += '<td>';\n\t\t\t\thtml += row.class;\n\t\t\t\thtml += '</td>';\n\t\t\t\thtml += '<td>';\n\t\t\t\thtml += row.score;\n\t\t\t\thtml += '</td>';\n\t\t\t\thtml += '</tr>';\n\t\t\t});\n\t\t\t}\n\t\t\telse\n\t\t\t\thtml+=\"<div>No Data</div>\";\n\t\t\t$('#scoretable').html(html);\n\t}\n (function(scope) {\n scope.$watch('msg.payload', function(data) {\n console.log('Position 2');\n console.dir(data);\n getdata(data);\n\n });\n })(scope);\n \n \n// or overwrite value in your callback function ...\nthis.scope.action = function() { return x; }\n\n</script>","storeOutMessages":true,"fwdInMessages":true,"resendOnRefresh":false,"templateScope":"local","x":160,"y":300,"wires":[["6d921df6.a91924"]]},{"id":"8415ec9d.561ec","type":"play audio","z":"aa14204.f27c3e","name":"","voice":"0","x":890,"y":560,"wires":[]},{"id":"2637150a.7d26da","type":"ui\_group","z":"","name":"Results","tab":"2ed27235.7b9cee","order":2,"disp":true,"width":"9"},{"id":"12b04aca.de3335","type":"ui\_group","z":"","name":"Watson Visual Recognition","tab":"2ed27235.7b9cee","order":1,"disp":true,"width":"12","collapse":false},{"id":"2ed27235.7b9cee","type":"ui\_tab","z":"","name":"IBM","icon":"dashboard","order":3,"disabled":false,"hidden":false}] |

Ada