Medical Visual Question Answering

09/12/2018

Avi Turner - 039991708

[avi.turner111@gmail.com](mailto:avi.turner111@gmail.com)

055-6699069

Table of content

[Introduction 3](#_Toc532554264)

[Glossary 3](#_Toc532554265)

[Background 4](#_Toc532554266)

[Deep learning 4](#_Toc532554267)

[VQA - Visual Question Answering 5](#_Toc532554268)

[The Challenge 6](#_Toc532554269)

[VQA-Med 6](#_Toc532554270)

[Offered solution 7](#_Toc532554271)

[Preliminary results 8](#_Toc532554272)

[Metrics 8](#_Toc532554273)

[WBSS 8](#_Toc532554274)

[BLEU 8](#_Toc532554275)

[Results from ImageCLEF VQA-Med 2018: 9](#_Toc532554276)

[Comparison to literature 9](#_Toc532554277)

[**Bibliography** 11](#_Toc532554278)

[Additional resources 11](#_Toc532554279)

# Introduction

Visual Question Answering, or VQA, is a research domain of building software systems that can intelligently answer natural language questions regarding a given image. i.e. given a coupled image and question (input), it will output an appropriate answer, also phrased in natural language.

Such a software system, that can comprehend the content of an image and describe it with a natural language, and more so in the clinical context, is a new and exciting domain that is lately getting increasing interest in the NLP and computer vision communities.

This field relies heavily on foundations that were established in the Deep Learning, NLP and computer vision

The VQA domain is rather new (The earliest paper I know of was from 2014) and hence has much room for research and development.

In this project we will focus on implementing such a system, in the context of the medical industry and in specific medical diagnosis that are described in a natural language based on images from medical imaging devices and question about those images.

We expect to gain better results then the winning project in the ImageCLEF 2018 contest. ImageCLEF 2018

## Glossary

1. NLP – Natural language processing
2. RNN - Recurrent neural network - A flavor of neural networks. In this network type, the output of every layer acts also as the layers input in the next iteration. This architecture allows the networks to "remember" the state accumulated so far, and this enables describing sequences. Used extensively in the field of NLP.
3. LSTM – Long Short-Term Memory – A flavor of RNN. Addresses the challenge of vanishing/exploding gradient.
4. VGG, ResNet - Deep networks (19 & 152 layers respectively) (figure 1). Some of the Image recognitions contest winners (2014/2015 respectively). Both are widely used in the field of computer vision.
5. FC – Fully Connected - A neuron network layer where all units are connected to all the units in the next layer.

# Background

## Deep learning

Deep Learning is basically a neural network with hidden layers, that is – non-input/output layers. In practice it usually refers to numerous hidden layers. The definition of "numerous" has somewhat changed along the years along with the technology evolution, starting at 8 (**2012**) and beyond 150 (**2015**) [figure 2]

The foundations of classical neural networks were established by 1943, when a paper was published (McCulloch & Pitts, 1943) that suggested a simple model for the neuron unit that was a base for the emerging research field of neural networks. By **1957** there were well known implementations of neural networks and its basic building block- the perceptron.

In **1969**, a book was published and attacked the ideas behind neural networks. The book presented an alleged defect in the basic neuron. Arguments presented in the book, and in particular the one that claimed that regardless of the weights we use as input, the perceptron cannot function as logical xor gate, delayed the development of neural networks for many years.

In **1980** a learning algorithm was "re discovered" after originally published in **1974**. This algorithm allowed neural networks to learn nonlinear models as well as it gave the whole domain a leap forward in the popularity of neural networks.

Another leap forward was established in **1989**, when a proof (Lewicki & Marino, 2004) showed that neural networks can approximate any function. By **1990** there were successful production implementations of neural networks in the fields of Medicine, Marketing, Risk management and more.

In the early **2000**s, along with a significant improvement in hardware performance, the "age of the Internet" and "information explosion", neural networks began to yield good results in the field of machine learning. As hardware progressed, and especially the usage of graphic processors, neural networks models with more layers were now feasible and converged in practical time while yielding good results. This also allowed neural networks to solve more complex problems. These networks were known as **deep learning networks**.

Since **2012**, all the winners of large computer vision object identification competitions are based on deep learning networks. And since **2015** - with better performance than a human being (Figure 2)

Figure 1

The percentage of error of deep learning networks and the number of layers in the network

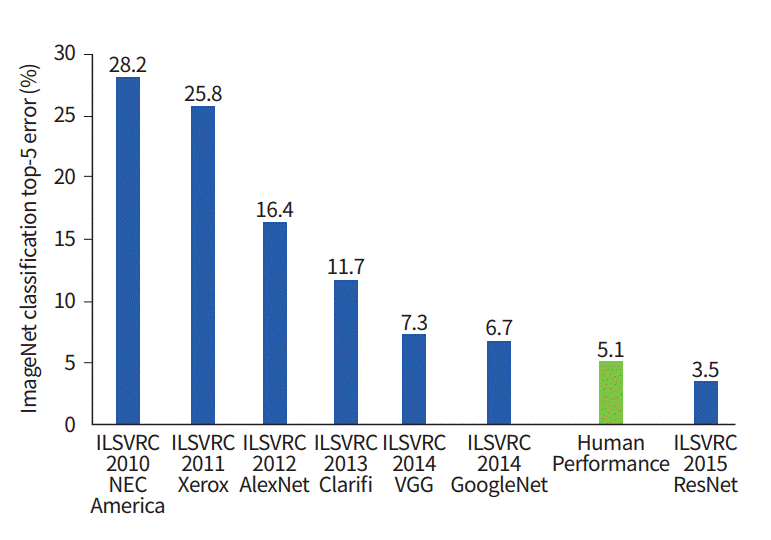
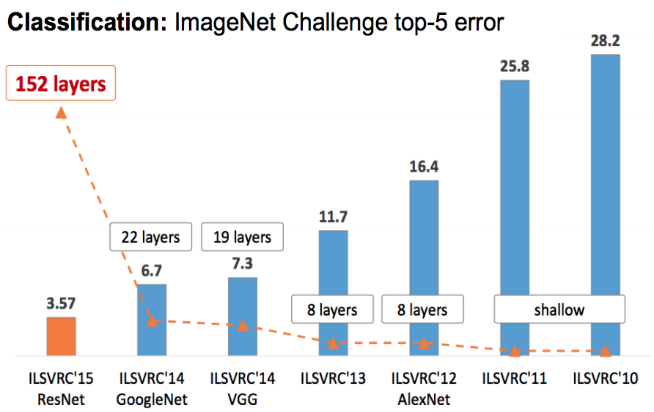


Figure 2

The percentage of error in learning networks is profound in relation to a person's performance

## VQA - Visual Question Answering

The evolvement of Deep Learning, technology – and GPUs in specific and data availability allowed learning of more raw and complex data such as natural languages and raw images. That is – that the image itself is the input and features extracted from the image.

VQA is a natural evolution of those two fields and combining the with synergy into a single system.

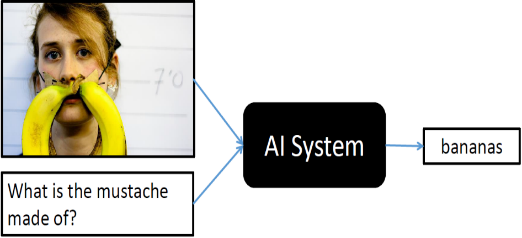
This field of study deals with the construction of software systems capable of answering questions, formulated in natural language, for a given picture.

The field is relatively new (in 2014, the first dataset and article on the subject were published), and it still has much room for research and progress.

As for today, the leading algorithms reach 70% accuracy (Fukui et al., n.d.) compared to 83% accuracy in humans - it should be noted that these results were trained on a dataset of hundreds of thousands of images.

For Illustrating:

Given an Image and an answer, the VQA system should supply an answer in the context of the question and the answer.



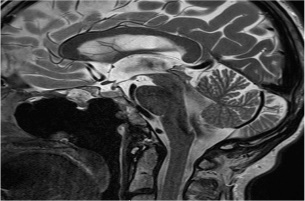
VQA

In the dataset, an image can have multiple questions / answer.



# The Challenge

## VQA-Med



what does t2 weighted sagittal magnetic resonance scan show?

VQA

total excision tumor with packg material seen posterior part clivus extradural plane

A software system that can understand the content of an image and describe it in a formulated natural language manner is a new exciting field of research that is getting increasing interest in the NLP and computer vision communities.

Such a system can be highly beneficial and solve real world problems in more than one way:

1. As part of their work, many experts and researchers write reports on microscopic findings in medical images, and an automatic system that describes what is in the picture can significantly reduce the burden and release bottlenecks throughout the medical service process.
2. These days, in medical centers, there is a great deal of strain on doctors, a condition that may cause insufficient attention to medical diagnosis - and as a result, a risk to human life. Such a system can serve as a control system and "raise a flag" for any diagnosis that differs from the system's output so that an expert can have a second look on the diagnostics that differed
3. In many places, even in the Western world, advanced medicine is out of reach for the underprivileged. A matured VQA-Med system will provide quick and inexpensive diagnosis for those who do not have available medicine services. Even in the absence of such maturity, it will be able to provide a second opinion that is cheap and almost automatic for those who cannot afford it.

As mentioned above, the field is still in its infancy and it is reasonable to assume that in the scope of this project we will not produce a mature enough system to fully address these problems, but there is an expectation to lay another milestone on the way to these goals.

ImageCleff is a voluntary organization that aims to provide an evaluation forum for the cross–language annotation and retrieval of images. Motivated by the need to support multilingual users from a global community accessing the ever-growing body of visual information, the main goal of ImageCLEF is to support the advancement of the field of visual media analysis, indexing, classification, and retrieval, by developing the necessary infrastructure for the evaluation of visual information retrieval systems operating in both monolingual, cross–language and language-independent contexts. ImageCLEF aims at providing reusable resources for such benchmarking purposes[[1]](#footnote-1).

In 2018, ImageCLEF started an evaluation campaign to focus on visual question answering in the medical domain (Hasan et al., 2018).

As part of this project, we expect to achieve progress in terms of accuracy of results compared to similar work done in the ImageCLEF 2018 competition

# Offered solution

In this project we propose to build a VQA system as follows:

Inputs:

* Images
* Question/Answers pairs on the given images

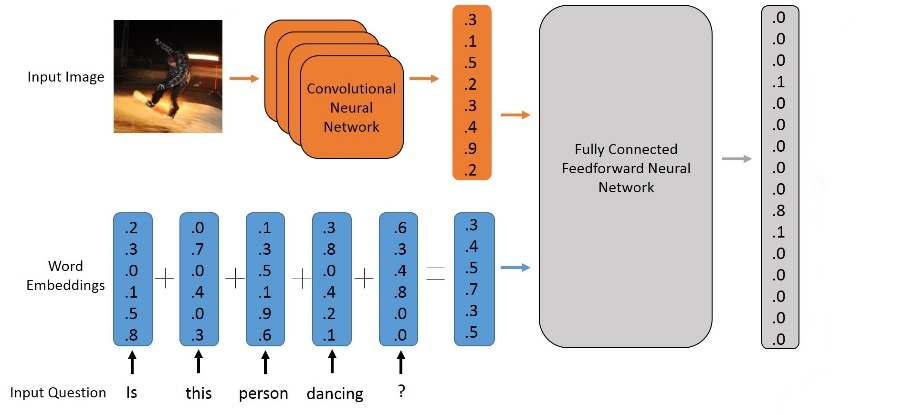
In a similar manner to other models developed in recent years, we will use the following general structure:

1. A network that receives input through two branches (Fig. 3) (Atg, 2018)
   1. A visual information branch. (a)
      1. Input: Image
      2. Output: Vector that represents the image properties.
   2. An NLP branch. (b)
      1. Input: The text of the question
      2. Output: a vector that represents the properties of the text

The image branch uses a trained Deep Learning network (e.g. VGG) to extract features from the image that are encoded in the form of a vector.

The NLP branch uses an LSTM network to take NLP encoding of the question and the internal state of LSTM as features and extracts a vector that represents those features.

The two branches are then combined by an FC layer and a classical neuron network whose output is the probability of each of the words in the corpus to being found in answer to the given question and picture.



(b)

(a)

figure 3

A classic structure of VQA network

The input to the VQA system is the pictures and questions / answers received within from the ImageCLEF contest. In the meantime, we used images from the 2018 contest that include 2280 images with 5300 questions and answers. For the system evaluation we received a similar set of 325 pictures and 500 questions / answers.

The source of the corpus content is a previous challenge of ImageCLEF (ImageCLEF 2017 caption prediction), in which the task was to predict what the caption attached to a medical picture would be in PubMed Central articles (an open archive of medical articles and articles).

The images + captions were then used as the input for a question / answer generator algorithm. In the first stage, the algorithm simplifies complex sentences into few simple sentences, then replaces the subject of the sentence with a question word and sets the subject the answer to the newly generated question. Finally, the questions / answers were also reviewed by a pair of experts who cleaned the results into a reasonable outcome.

Q: what does 3d CT view show?

A: rotational movements measured

Q: what shows air space consolidation in left lower lobe?

A: CT



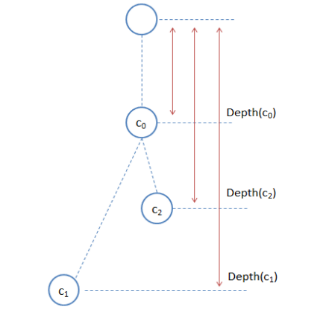
# Preliminary results

## Metrics

### WBSS

The WBSS metric estimates the machine generated output based on semantic similarity in the biomedical field.

WBSS calculates the similarities between the answer (the output) and the "ground truth" that is provided as a parameter.



*Figure 4*

*Visualization of*

*depth calculation (function d)*

The score is based on the semantic similarity at the word level.

The words in the corpus, in our case a medical natured one, are arranged hierarchically as a graph. The similarity between two words is represented by the length of the shortest path between these two words in the graph. If the graph was constructed correctly, words of similar meaning would have a shorter path than words of other meaning. for example, [tumor, cancer] / [tumor, fracture] we would expect that the distance between tumor and cancer would be shorter.

The score is obtained by:

Where c1 and c2 are the words that we want to calculate their similarities, and c0 is the most specific word (graph hierarchly wise) in the path for both words, e.g. the common ancestor closest to both words (Buscaldi, Tournier, Aussenac-gilles, & Mothe, 2012).

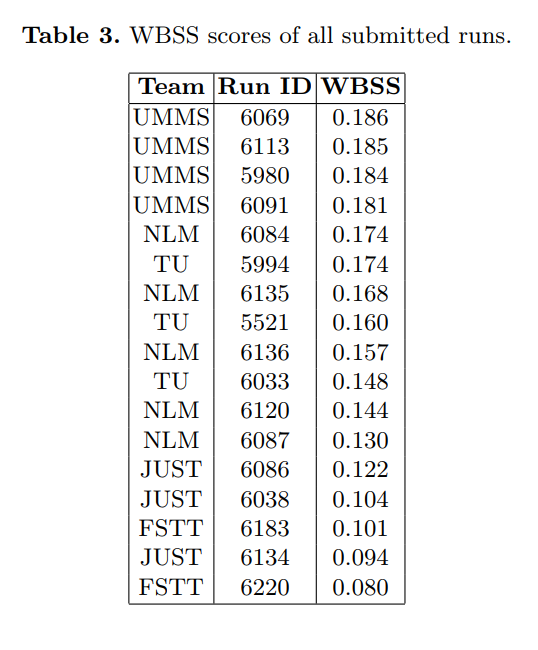
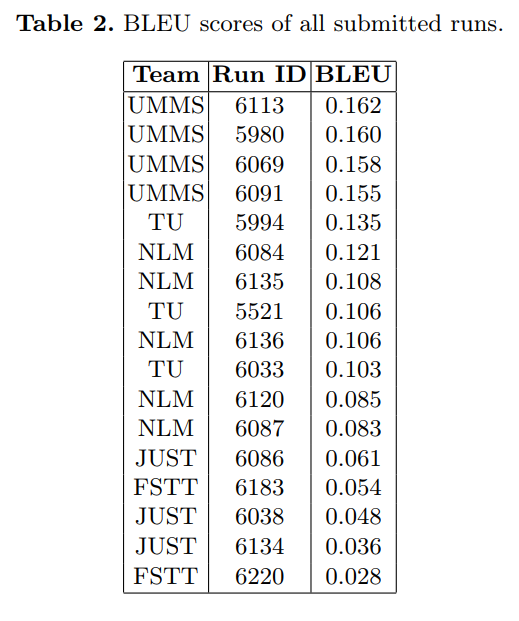
### BLEU

An algorithm for estimating the quality of text translated by machine in comparison to a "ground truth":

* The words are tokenized - from sentences to a corpus of words:
  + All letters are converted to lower case
  + Punctuation is removed.
* Common words (stop words) are removed.
* Stemming - representation of different forms of a word single normalized representation *(fishing*, *fished*, *fisher*   *fish*)
* The score is a based on the number of n-grams that appear in the answer and in ground truth, there is no importance to the correctness or structure of the sentence.

## Results from ImageCLEF VQA-Med 2018:

(Hasan et al., 2018)



Graded results we got for our submission to the 2018 contest (after it was closed):

|  |  |  |
| --- | --- | --- |
|  | BLEU | WBSS |
| 19/10/2018 | 0.157 | 0.155 |
| 19/10/2018 | 0.146 | 0.143 |

# Comparison to literature

Since the field is still in its infancy, the only source I found that deals with VQA in general and in the medical field was the ImageCLEF contest overview

The dataset of image-questions-answers we used came from the 2018 competition, where the winner accomplished the following results:

|  |  |  |
| --- | --- | --- |
|  | BLEU | WBSS |
| During contest | 0.158 | 0.186 |
| After Contest | 0.188 | 0.209 |

The competition overview showed that most contestants used some form of deep learnings system that included some or all the following characteristics

1. Encoding the question using libraries / NLP algorithms
2. Deep Convolutional networks for image encoding (with or without pre-trained networks such as VGG / ResNet)
3. Recurrent networks (RNN) to encode the question (with or without pre-trained encoders from another Corpus)

Some contestants formulated the task as a multi-label multi-class classification problem[[2]](#footnote-2), while others considered it as a generation task[[3]](#footnote-3), and tried to generate a suitable answer.

Another feature that some of the contestants used was the implementation of Attention to link the question to the image in the network structure[[4]](#footnote-4).

There have also been attempts to use more advanced tools such as (Hasan et al., 2018):

1. Multi modal compact bilinear (MCB) pooling
2. Multimodal factorized bilinear (MFB) pooling
3. Embedding based topic modeling (ETM) [[5]](#footnote-5)

In our implementation we are using:

1. Encoding the question using an RNN network, in particular - LSTM.
2. Pre-trained Deep convolutional networks for image features extraction.
3. Encoding question words using an NLP library that was pre-trained Wikipedia
4. Planning to implement: Usage of Attention.

It is worth noting that in language-to-language competitions, the BLEU usually evaluates results as high as 0.33. Scores here were significantly lower. And one of the lessons learned from the competition is that there is a need to increase the size of the train dataset in order to obtain more meaningful results.

# **Bibliography**

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (n.d.). VQA : Visual Question Answering, 1–23.

Atg, U. (2018). Read This Month, (January).

Buscaldi, D., Tournier, R., Aussenac-gilles, N., & Mothe, J. (2012). IRIT : Textual Similarity Combining Conceptual Similarity with an N-Gram Comparison Method. *First Joint Conference on Lexical and Computational Semantics (\*SEM 2012)*, 552–556.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (n.d.). for Visual Question Answering and Visual Grounding.

Goyal, Y., Khot, T., Dhruv, D. S., & Devi, B. (n.d.). Making the V in VQA Matter : Elevating the Role of Image Understanding in Visual Question Answering.

Gupta, A. K. (n.d.). Survey of Visual Question Answering: Datasets and Techniques.

Hasan, S. A., Ling, Y., Farri, O., Liu, J., Henning, M., & Lungren, M. (2018). Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task.

Lewicki, G., & Marino, G. (2004). Approximation of functions of finite variation by superpositions of a sigmoidal function. *Applied Mathematics Letters*, *17*(10), 1147–1152. https://doi.org/10.1016/j.aml.2003.11.006

Lin, X., & Parikh, D. (n.d.). Active Learning for Visual Question Answering : An Empirical Study arXiv : 1711 . 01732v1 [ cs . CV ] 6 Nov 2017, 1–23.

McCulloch, W. S., & Pitts, W. H. (1943). originally published in: Bulletin of Mathematical Biophysics, Vol. 5, 1943, p. 115-133. *Bulletin of Mathematical Biophysics*, *5*, 115–133. https://doi.org/10.1016/j.joca.2006.06.009

Teney, D., & Centre, A. (2017). Tips and Tricks for Visual Question Answering : Learnings from the 2017 Challenge.

Zhang, P., Goyal, Y., Summers-stay, D., Batra, D., Tech, V., & Parikh, D. (n.d.). Yin and Yang: Balancing and Answering Binary Visual Questions.

## Additional resources

Deep Learning and Visual Question Answering

<https://towardsdatascience.com/deep-learning-and-visual-question-answering-c8c8093941bc>

Deep Learning for Visual Question Answering

<https://avisingh599.github.io/deeplearning/visual-qa/>

1. https://www.imageclef.org/ [↑](#footnote-ref-1)
2. A technique that the winning group used [↑](#footnote-ref-2)
3. A technique the group with the worst results used [↑](#footnote-ref-3)
4. A technique that the winning group used [↑](#footnote-ref-4)
5. A technique that the winning group used [↑](#footnote-ref-5)