A Machine Learning Approach to Automatic Closed Captioning

Lionel Quintanilla, Andrew De La Rosa, Pouya Tavakoli, Brian Tran, Matthew Zaldana

California State University, Long Beach

ENGR 361: Scientific Research Communication

Dr. Maryam Qudrat

12 December 2021

Abstract

Per federal regulations, closed captioning is required for much Internet content. This is sometimes down with automatic captioning, where algorithms are used to create captions. However, due to technological limitations, these methods are often inaccurate and not useful. To solve this issue, we want to create an automatic captioning app that is both accurate and intuitive to use. We plan to do this by devising a captioning method that uses machine learning algorithms for better accuracy, and an application that can operate with web-based video players. To test our application, we will run evaluations on several potential algorithms using industry standard datasets, and conduct focus groups to determine the usability and benefits of the application. Ultimately, we expect the chosen algorithm to exceed current average scores for captioning and for the application to receive generally favorable responses. In terms of logistics, we expect the project to take about a month in total at a cost of several thousand dollars for equipment and facility rentals.

Research Question

The purpose of this research proposal is to develop and test an application that can automatically generate closed captions for Internet-based video players using machine learning techniques that can produce captions within a reasonable range of accuracy.

Aim 1: Devise a machine learning algorithm that can create captions with a reasonable degree of accuracy.

Our first aim is to devise and implement a machine learning algorithm that can automatically create closed captions with a reasonable degree of accuracy based on metrics for measuring the natural language accuracy of machine-generated text. Rather than develop our own algorithm, we plan to adapt several existing methods for caption generation and provide each one with test data to determine the most accurate method for use.

Aim 2: Create an application that can take in videos and display captions on top of the videos.

Our second primary aim is to develop an Internet browser extension that is interoperable with web-based video players. Users should be able to use this application to generate captions for video content, which will then be overlayed on the source video. This application is meant to solve the issue of videos that either do not contain captions or use methods for automatically generating captions that are not sufficiently accurate or readable.

Background/Significance

Closed captions are the “visual display of the audio portion of video programming” and are meant to convey dialogue and other information to deaf and hard-of-hearing individuals (FCC, 2021). Due to the prevalence of hearing disabilities (NIDCD, 2021), the government mandates that certain content includes closed captions (FCC, 2021). To fulfill this requirement, many providers create captions automatically with algorithmic methods. The most common of these is called Automatic Speech Recognition, or ASR. ASR creates captions by splitting audio into smaller samples, which are then processed through a speech recognition algorithm to identify words (Google Cloud, 2021). However, ASR does not produce particularly accurate closed captions and generally reaches only 60% to 70% accuracy (Bond, 2014).

Inaccurate captions can have severe ramifications. For example, having inaccurate captions can reduce the search ranking of the offending web page and may cause the page to be delisted (Bond, 2014). There is also a chance of legal liability, which was seen in 2015 when the National Association of the Deaf sued Harvard and MIT for having nonexistent or inaccurate captions on their educational videos (Rosenblum, 2015). Thus, providing accurate captions is in the best interest of organizations in order to protect themselves from liability and fines.

Many recent advancements in closed captioning have come from a machine learning tool called a neural network. Neural networks are composed of a series of “node layers”, where each node performs a calculation. Each node has an “associated weight and threshold” and will pass its data to other nodes if the calculation reaches its requirements. This allows neural networks to “learn and improve their accuracy” with

repeated calculations but requires a large amount of data to be effective (IBM Cloud Education, 2020). Still, the use of neural networks in captioning has provided promising results. For example, a method called cLSTM-RA uses a neural network to power a functionality called a “residual attention mechanism”, which allows it to pick out finer details in videos (Yang et al., 2020). When tested on the Flickr30k dataset, the cLSTM-RA method scored a 70.5 on the BLEU natural language test, which was higher than all other methods sampled (Yang et al., 2020).

Methods/Research Plan

Design

To determine the optimal machine-learning algorithm for our application, we will be running a multitude of captioning tests using four methods: generative adversarial network, neural machine translation, recurrent neural network, and a triple module encoder-decoder. Each algorithm will be run on about 1000 seconds (about 16 and a half minutes) of video media designed for closed captioning. The resulting captioned dataset will be randomized when presented to a group of sample users for qualitative analysis. Each algorithm will be tested for 1) accuracy of outputted dialogue and descriptive text [Aim 1], 2) quality and accuracy of the text on screen [Aim 2], and 3) Perceived quality of captions to real users.

Sample/Sample Size

Two datasets will be used for training each algorithm before final analysis. Each algorithm will be given the same data in the same order to ensure fair and consistent results. Dataset 1 is called MSCOCO, the industry-standard training set consists of 300,000 images with 5 captions per image (Microsoft, 2021). MSCOCO is good for

training short-descriptive capabilities over a wide range of genres. Dataset 2 is called Flickr30k, consisting of 30,000 images with 158,000 captions spread out among them (Plummer et al., 2015). Flickr30k tests for long descriptive capabilities of complex imagery.

A survey consisting of multiple choice and free response questions will be given to our focus group after they view 5 minutes of randomized captioned footage each. Each participant must watch footage on the iPhone, the laptop, and the TV. The footage will be the same for each device.

Setting

Algorithms will be tested in an air-conditioned computer laboratory where the computers can run uninterrupted. A multimedia room is used to view the final product of the algorithms and analysis of quantitative measures conducted there. Study participants will be invited to the multimedia room after quantitative analysis has been completed. The captioned video will be displayed on an iPhone 11, a 15” laptop monitor, and a 55” TV. Participants watch 5 minutes of video on each device and afterwards, are given time to complete their qualitative assessment survey.

Protocol

The algorithms will run on separate computers. Each computer trains with datasets for one 24-hour cycle. After the algorithm is trained, the automated AT metric tests are run only once per algorithm. Initial results for the test are stored in one file per algorithm contained on the algorithm’s host computer, and then collected on a shared external drive for analysis.

Survey takers are asked to watch a total of 15 minutes of captioned footage before filling out the survey. There is a 15-minute time limit for completion of the survey, but the survey should take no more than 7 minutes to complete. It is to be ensured the area is free of distractions, which includes asking the survey taker to turn off their cellphone.

Analysis Plan

There are two end goals that the analysis portion of the outcomes wishes to achieve. One of these goals is to compare the results of each technique using the four most common machine learning tests to determine the accuracy of the captions. Analysis of caption correctness will be based on accuracy of produced text (AT), visual quality of the text on-screen (VQ), and quality of descriptive language (QL). AT is the derived average of scores from four separate metrics used in testing the accuracy of machine translated text. BLEU, METEOR, ROGUE, and CIDEr are all used as industry standard analysis tools. AT is scored on a scale from 0 to 1.

The second goal of our analysis is to measure the visual appeal of the text generated on screen, as well as the quality of descriptive text for HoH/deaf people. Because there are no standard metrics for measuring visual appeal of on-screen text the survey will contain questions regarding the visual appeal of the text on screen and ask survey-takers multiple questions with a 0 to 1 rating system, with options of 0.1 increments. VQ is the average of scores received from those survey questions. Survey questions pertaining to descriptive language for HoH/Deaf individuals are recorded with the same 0 to 1 rating system. QL is the average of those scores. For all metrics: an average score is around 0.5, a good score is 0.6 to 0.7, an excellent score is 0.8 to 0.9.

The survey also contains open response sections that gauge the overall experience users had with our software, these questions are designed to identify future features or critical errors, otherwise undetectable via metrics.

Outcomes and Expectations

Aim 1

For the first aim, our goal is to provide a technique that should provide immediate improvement in accuracy over existing methods. One way we would be able to decipher the data given from these experiments is using BLEU, which measures the accuracy of machine-generated text at a sentence level (Papineni et al., 2001), METEOR, which measures the accuracy of text at a word level (Lavie et al., 2007); and ROGUE, which measures the accuracy of machine-generated text with substrings, or parts of sentences (Lin et al., 2004). All three metrics compare the text generated by the machine learning algorithm by a set of expected results. This is especially important as Dr. Somang Nam in their article “Modeling Closed Captioning Subjective Quality Assessment by Deaf and Hard of Hearing Viewers” that findings revealed that hard of hearing “viewers would be more likely to detect caption errors if captions were missing more words” (Nam et al., 2020). We must also consider that there exist other factors which “may affect the quality assessment of a viewer, such as the pace or complexity of the visual content that is contained in different genres […] or familiarity with the content topics” (Nam et al., 2020). For these reasons, we expect our technique to score around 60 to 80 marks on all natural-language tests as an average. We desire to achieve an impact where the adapted technique improves not only the accuracy of the words, but also the grammar and the vocabulary to make it easier to understand. This way, the

user does not have to focus on grammatical or unconcise sentences, rather would pay attention to the information presented in the captions themselves.

Aim 2

The expectation for the second aim is that the focus group will answer the questionnaire with at least a 3 on all the questions. This is imperative as computerized examinations can only go so far in measuring the effectiveness of our technique. If our survey group cannot determine whether our method was useful or not, we will not get far with conducting more research. We need an application that is effective and easy to use, where reliability and efficiency are the heart of the extension. Dr. Sheryl Burgstahler wrote in her article “Creating Video and Multimedia Products That Are Accessible to People with Sensory Impairments” that to make attractive and functional captions, the following must be included: Use one or two lines of text, use both uppercase and lowercase letters, use a simple sans-serif font, ensure high contrast” among others (Burgstahler, 2018). This way the captions are not impaired by the visuals on the screen. Articles such as “Methods of Improving and Optimizing React Web-applications” written by Filip Pavic and Ljiljana Brkic support the fact that “data showed the impact loading time has on user experience and users’ subsequent actions, highlighting the need for web application optimization (Pavic et al., 2021). The impact that a good user design would have on our users would not only shift attention to our application for use in general-purpose areas as well as specialized ones, but it would allow us to work with all kinds of videos and images. The MSCOCO data will be especially useful here as it gives insight into how well the data performed in giving accurate descriptions and we feel that given our well-rounded adapted technique allows

us to achieve this. Combining this data with what Dr. Burgstahler commented on caption-captivating methods creates a strong connection between what the users' eyes will see as a sentence on the screen as well as the comprehension that drives their brain in the understanding of the information.

Proposed Timeline

The general timeline of this experiment relies on two different types of development to occur, algorithm development and web extension development. The algorithm requires the act of programming, training, and analyzing the algorithm's results. The programming of the algorithm can vary due to the need for revisions. It is expected that programming will take two weeks to a month. The next part is testing and training the algorithm, which is expected to take one to two weeks. The last item is to analyze the data that was constructed via the training, which is expected to take one to two weeks and possibly extend development by two to three weeks for optimization.

The web extension development will use the captioning algorithm from aim 1 and will take place during the algorithm development to finish both aspects at around the same time. The first part of the web extension development is to create a mechanism where videos can be taken or analyzed by the algorithm. This is expected to take around one to two weeks or longer if there are unexpected errors with different media sources. The next aspect to develop is the ability to display the closed captions from the algorithm, which would take around one to two weeks to develop the methodology to display the captions but could take longer depending on how incomputable the data exported from the algorithm is. Constructing the entire application together with the algorithm is the last part to finish the development of both aims, which is estimated to

take around a week to make sure that the application remains stable before volunteer testing. The last piece of the timeline is to test the entire program with volunteers, which is estimated to take a duration of a week to find possible volunteers for a large testing data set.

Budget

A budget is necessary in order to facilitate the development and testing of this web application. When developing the algorithm for caption creation, it is vital to test six different techniques to understand which one would be best for this caption extension. Based on an article by Brian Benton from Redshift, a general workstation computer that could be used to test these different machine learning techniques costs around $500 to $1,000 (Benton, 2021). The average cost of each computer is about $750 or around $4,500 to purchase a computer per technique. The datasets which are used to test and train each machine learning technique can be obtained for free and do not affect the budget. The development for the second aim will use the computers previously used to develop the caption algorithm. The last portion of the budget is used in order to create a testing environment with volunteers.

Creating a good testing environment for volunteers requires a conference room to conduct tests, a series of computers to host the web application, and money in order to entice volunteers into volunteering. Renting a conference room costs around $70 to $160 per hour, with the average cost being about $115 per day (Peerspace, 2021). The laptops necessary for testing the web application can be rather cheap, in which case Chromebooks can be used with an average price of about $150 per laptop (Pickard, 2021). A good sample size of volunteers should consist of about five to ten people per

test and around three tests would be sufficient to confirm the usability of the web application. Volunteers would need to be compensated with a cash reward for their time, which is expected to be around $25 to $50, or an average of $38 for each volunteer. With an average of seven people per test, the cost to manage each test is around $1,050 for seven lower-end laptops for testing and around $266 for each test's compensation. The budget will total around $6,550 if there is only one day for 3 sets of testing with groups of around seven people.

References

Benton, B. (2021, October 21). *Workstation vs desktop computer: Which do you need?* Redshift by Autodesk. Retrieved December 11, 2021, from https://redshift.autodesk.com/pc-versus-workstation/.

Bond, L. (2019, June 3). Are Automatically Generated Captions and Transcripts Detrimental to Video SEO? 3Play Media. Retrieved November 15, 2021, from https://www.3playmedia.com/blog/are-automatically-generated-captions-transcripts-detrimental-video-seo/.

Burgstahler, S. (2018). *Creating Video and Multimedia Products That Are Accessible to People with Sensory Impairments*. Creating Video and Multimedia Products That Are Accessible to People with Sensory Impairments | DO-IT. Retrieved December 11, 2021, from https://www.washington.edu/doit/creating-video-and-multimedia-products-are-accessible-people-sensory-impairments.

Federal Communications Commission. (2021, January 27). Closed Captioning on Television. Federal Communications Commission. Retrieved November 14, 2021, from https://www.fcc.gov/consumers/guides/closed-captioning-television.

Google Cloud. (2021). *Speech-to-Text Basics*. Google. Retrieved December 11, 2021, from https://cloud.google.com/speech-to-text/docs/basics.

IBM Cloud Education. (2020, August 17). *What Are Neural Networks?* IBM. Retrieved December 11, 2021, from https://www.ibm.com/cloud/learn/neural-networks.

Lavie, A., & Agarwal, A. (2007). Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, 228–231. https://doi.org/10.3115/1626355.1626389

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out, 74–81. https://doi.org/https://aclanthology.org/W04-1013

Microsoft. (2021). *Common Objects in Context*. COCO. Retrieved December 11, 2021, from https://cocodataset.org/#home.

Nam, S., Fels, D. I., & Chignell, M. H. (2020). Modeling closed captioning subjective quality assessment by deaf and hard of hearing viewers. *IEEE Transactions on Computational Social Systems*, *7*(3), 621–631. https://doi.org/10.1109/tcss.2020.2972399

National Institute of Deafness and Other Communication Disorders. (2021, March 25). Quick Statistics About Hearing. Retrieved November 14, 2021, from https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311–318. https://doi.org/10.3115/1073083.1073135

Pavic, F., & Brkic, L. (2021). Methods of improving and optimizing react web-applications. *2021 44th International Convention on Information, Communication*

and Electronic Technology (MIPRO). https://doi.org/10.23919/mipro52101.2021.9596762

Peerspace. (2021, October 23). *How Much Does It Cost to Rent a Hotel Conference Room?* Peerspace. Retrieved December 11, 2021, from https://www.peerspace.com/resources/rent-hotel-conference-room/.

Pickard, J. (2021, November 11). *The Best Cheap Chromebook Prices and deals in December 2021*. TechRadar. Retrieved December 11, 2021, from https://www.techradar.com/news/cheap-chromebook-deals.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1–22. https://doi.org/10.1109/iccv.2015.303

Rosenblum, H. A. (2015, February 12). *NAD Sues Harvard and MIT for Discrimination in Public Online Content*. National Association of the Deaf. Retrieved December 11, 2021, from https://www.nad.org/2015/02/17/nad-sues-harvard-and-mit-for-discrimination-in-public-online-content/.

Yang, L., Hu, H., Xing, S., & Lu, X. (2020). Constrained LSTM and residual attention for image captioning. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(3), 1–18. https://doi.org/10.1145/3386725