Automated Audio Captioning with Epochal Difficult Captions for curriculum learning

Andrew Koh*, Soham Tiwari †, Chng Eng Siong*
* Nanyang Technological University, Singapore
† Manipal Institute of Technology

Abstract—In this paper, we propose an algorithm, Epochal Difficult Captions, to supplement the training of any model for the Automated Audio Captioning task. Epochal Difficult Captions is an elegant evolution to the keyword estimation task that previous work have used to train the encoder of the AAC model. Epochal Difficult Captions modifies the target captions based on a curriculum and a difficulty level determined as a function of current epoch. Epochal Difficult Captions can be used with any model architecture and is a lightweight function that does not increase training time. We test our results on three systems and show that using Epochal Difficult Captions consistently improves performance by up to 0.013 SPIDER score.

I. INTRODUCTION

The Automated Audio Captioning task is a cross-modal task where the model is trained to generate a descriptive sentence for an input audio. This task has many practical real world use cases, such as closed captioning for people who are deaf or hard of hearing. Automated Audio Captioning describes the sound events that are happening in both the foreground and background of the audio, as opposed to Automated Speech Recognition, which predicts the low time frame phonemes and utterances.

The Automated Audio Captioning task requires the model to be able to capture both high and low level acoustic events that are occurring throughout the audio [1]. For instance, the model would be required to caption both low time frame events such as explosion sounds happening only in a few seconds and high time frame events such as sirens which often occur throughout an extended period of the audio clip. As such, previous work have often used pretrained models such as PANNs that have already been trained on audio tagging in order to aid the model in generating factually correct captions. In a similar vein, several authors [2], [3] have used keyword estimation, a supplementary and intermediate objective where the model has to predict one or a few keywords from the encoder to help the decoder to generate better captions. The keyword estimation task has shown potential in helping the model perform better. However, the keyword estimation task requires heavy preprocessing to create the audio-keywords pairs for training. Furthermore, while the different authors have the same motivation behind keyword estimation task, their implementations and audio-keywords pairs are often different and subject to their own interpretation [4], [3]. This makes it hard to compare results and methods across different papers. We find that there is an elegant solution to this. Epochal Difficult Captions modify the training targets directly during the training loop. Based on a curriculum, the training captions start off with a higher keywords to sentence ratio during the earlier epochs in the training run. This is done by removing stopwords directly from the batched training targets. As the training progresses, the keywords to sentence ratio will be gradually reduced until the sentence reverts to its original unmodified sentence.

We test Epochal Difficult Captions using two different publicly available systems [1], [5] from previous work. Both systems are of the typical encoder-decoder structure. The encoder consists of either a single [5] pretrained CNN or both [1] a pretrained CNN and transformer encoder. The decoder used is a 2 layer transformer decoder. The extracted audio features from the encoder is passed to the decoder for cross attention. Both systems are trained using cross entropy, but [5] has the additional option of using reinforcement learning to optimize the metrics directly. We experiment using both cross entropy, and reinforcement learning where available. We find that using Epochal Difficult Captions consistently improves performance across these settings with no noticeable extra compute time and power needed. Epochal Difficult Captions is used only during training and hence does not affect any part of the inference process.

The rest of this paper follows this structure. In section II, we introduce related work pertaining to Automated Audio Captioning datasets, popular model architectures and training objectives. We also discuss the keyword estimation task, a popular intermediate task in Automated Audio Captioning. In Section III, we explain our approach, Epochal Difficult Captions for Curriculum Learning. This section shows how we derive our difficulty curve and levels, and how stopwords are removed as an function of the difficulty level. In Section IV, we specify our experimental settings, details, and process to evaluate the effectiveness of Epochal Difficult Captions. In Section V, we examine the performance of the models trained with Epochal Difficult Captions for Curriculum Learning. Finally, we give our concluding remarks in Section VI.

II. RELATED WORK

A. Audio Captioning Datasets

The Clotho Dataset v2.1 [6] consists of 6974 audio samples of duration from 15 to 30 seconds. Each audio clip has 5 corresponding captions annotated by humans. Each caption
has a length of 8 to 20 words, and the captions describe the events in that audio. The Clotho Dataset v1 was augmented by including more training samples and a validation split to form Clotho v2.1. The Clotho Dataset was created to address the critiques of the larger Audiocaps dataset [7], which was found to contain biases in the annotation process. There are also other niche datasets like the Hospital Scene and Car Scene dataset [8] which are Mandarin annotated.

In this work, we focus only on the Clotho Dataset v2.1, henceforth referred to as the Clotho Dataset.

B. Model Architectures and Objectives

Most work use an encoder-decoder model architecture. The encoder is used to encode audio information to produce audio features, which is then passed to the decoder for cross attention and to generate the caption. Most architectures converge to a similar theme - using a pretrained convolutional neural network such as PANNs for encoding, and a transformer decoder for decoding. However, the training process often differs from the pretraining stage and an intermediate or supplementary objective.

For instance, the AT-CNN [9], and the Audio Captioning Transformer [10] pretrains their model encoder on audio tagging as a pretraining task before finetuning their model directly on Automated Audio Captioning. Other authors also use self supervised supplementary objectives [1] or caption retrieval [11] to guide the model training. Self-Critical Sequence Training [12], a reinforcement learning tactic used in Image Captioning, has also been used to optimize the model for Audio Captioning. Keyword Estimation, which we will go into detail in Section II-C, has also been a popular choice [4], [3] of an intermediate objective.

In this work, we build on the work of [1] and [5] by using their model architectures that has been proven to achieve competitive results. Henceforth, we refer [5] and [1] as System 1 and System 2 respectively. Figure 1 delineates their system architectures.

C. Keyword Estimation

Several authors have applied the Keyword Estimation task as a intermediate task to aid the training of the model for Automated Audio Captioning. This is done by decomposing the Automated Audio Captioning task into two smaller tasks, Keyword Estimation and Caption Generation. The model encoder then performs Keyword Estimation to predict a set of keywords, which is then passed to the decoder for generation. The Keyword Estimation task helps to mitigate the problem of indeterminacy in word selection [4] hence reducing the solution space for the decoder to generate better captions.

Since the Clotho Dataset does not come with labels for keywords, authors often use their own interpretation to create labels for the keyword Estimation task. [4] uses a rule based keyword extraction method to produce keywords labels for training. Frequent word lemmas of nouns, verbs, adjectives, and adverbs as regarded as keywords. On the other hand, [3] extracts their keywords by using the Natural Language Toolkit\footnote{https://github.com/nltk/nltk} to discard ‘useless words’ such as make, go, etc. Remaining verbs are used in their original forms, and the nouns are not changed.

Firstly, given keyword estimation requires some form of pre-processing subject to the interpretation of the researcher, it is difficult to compare the effectiveness of the keyword estimation task across different pieces of work. Secondly, decomposing the Automated Audio Captioning task into subtasks keyword estimation and caption generation requires the decoder to rely on the prediction of the keywords from the encoder. This might result in error propagation if the keywords are inaccurate. With these considerations, we propose the Epochal Difficult Captions for curriculum learning, which has the same motivation behind keyword estimation, but does not require the decomposition of Automated Audio Captioning into subtasks.

III. EPOCHAL DIFFICULT CAPTIONS FOR CURRICULUM LEARNING

Epochal Difficult Captions for Curriculum Learning is an algorithm to modify the target captions relative to the current stage of the training setup. The difficulty level, \( D \), is a function of the current epoch. \( D \) indicates how the curriculum modifies the target captions. In this paper, we use two different curriculum. The first is the stopwords curriculum which modifies the target captions based on a predefined list of stopwords. The second curriculum is the frequency-based curriculum, which modifies the target caption based on how frequent each word occurs in the dataset.

A. Difficulty level as a function of epoch

The difficulty level in curriculum learning dictates the complexity of the captions and it should increase from a lower to a higher magnitude relative to the current epoch. We use an exponential function (Equation 1) followed by a floor and ceiling function (Equation 2) to generate a difficulty value \( D \) for all non-negative epoch values \( D \in (0, 1) \).

\[
D' = 1 - e^{(-\alpha \cdot \text{epoch})} \tag{1}
\]

\[
D = \begin{cases} 
0.05, & \text{if } D < 0.05 \\
D', & \text{otherwise} 
\end{cases} \tag{2}
\]

where the hyperparameter \( \alpha \) in Equation 1 controls the rate of increase of difficulty. At \( D = 1 \), the original captions are not modified. The difficulty level should asymptotically tend to 1 nearing the max epoch. Therefore, \( \alpha \) requires the max epoch hyperparameter to be taken into consideration. In our experiments, we use \( \alpha = 0.20 \) when max epoch is 30, \( \alpha = 0.10 \) when max epoch is 60, and \( \alpha = 0.05 \) when max epoch is 100.

\[1\text{https://github.com/nltk/nltk} \]
Figure 1: Detailed overview of System 1 and System 2. Both System 1 and 2 uses a pretrained CNN10 in the encoder. System 2 has an additional 2 layer transformer appended to the CNN10 in the encoder. For the decoder, both System 1 and 2 use a transformer decoder. However, System 1 uses pretrained word2vec word embeddings, while System 2 trains the word embeddings from scratch.

Table I: Examples of target captions at different epochs when Epochal Difficult Captions (EDC) stopwords curriculum is applied. In this table, we show Epochal Difficult Captions for 100 epochs with $z\alpha$ value 0.05. At low epoch values (epoch 0), almost all or all the stopwords are removed from the target caption. As training progresses (epoch 50), more stopwords are gradually returned to the target captions. At high epoch values (epoch 100), the target captions are almost identical to the reference captions, but there is still a low chance of some stopwords being removed.

| Target Caption | Epoch 0 | Epoch 50 | Epoch 100 |
|----------------|---------|----------|-----------|
| water flowing at a constant rate then slows down to a drip | water flowing a constant rate slows down to a drip | water flowing at constant rate slows down to a drip | water flowing at constant rate then slows down to a drip |
| a person is arranging objects on a table and pouring water into a glass | person arranging objects table pouring water glass | person is arranging objects on table and pouring water into glass | person is arranging objects on a table and pouring water into a glass |
| a man slides the file across the wood tapping it a few times at the end | man slides the file across the wood tapping times end | a man slides the file across wood tapping few times at end | man slides the file across the wood tapping a few times at the end |

B. Stopwords Curriculum

In the stopwords curriculum for Epochal Difficult Captions, the target captions for each audio are modified by removing stopwords. Stopwords are common words that carry little useful information and can be obtained from the public nltk library. Some examples of stopwords are ‘for’, ‘do’, ‘its’, ‘yours’, ‘such’, ‘into’. Removal of these stopwords mimic the keyword estimation task described in Section II-C by only including unique and less prevalent words in the caption.

Algorithm 1 describes the stopwords curriculum. Each stopword in the caption is an candidate for removal depending on the current difficulty level of the current epoch. The greater the difficulty level $D$ is, the lower the probability of removing a stopword from the caption.

In Figure 2, we plot the relationship between $D$, the total number of words used per epoch in training, and the average number of unique words per batch during training. At low difficulty levels, the total number of tokens per epoch increases rapidly with each epoch. At higher difficulty levels above 0.9, the total number of tokens fluctuates around the same range. Using such a token distribution based on a fixed curriculum allows us to gradually expose the model to progressively more unique words. In Table I, we also show a
The number of training tokens per epoch. Even at high difficulty levels, the number of tokens per epoch fluctuates.

Figure 2: Scatter plot of $D$ (x-axis) against total tokens per epoch (left y-axis) and average unique tokens per batch (right y-axis) for different epoch values 100 (a), 60(b), 30(c). The plots demonstrate the relationship between the difficulty, $D$ and the number of training tokens per epoch. Even at high difficulty levels, the number of tokens per epoch fluctuates.

**Algorithm 1** Stopwords Curriculum

**Input:** $y$ - the original target caption  
**Input:** $D$ - difficulty value (Equation 1, 2)  
**Input:** $SW$ - list of stopwords.  
**Output:** $y'$ - modified caption.  
1: $P(\text{remove}) \leftarrow (1 - D)$  
2: $y' \leftarrow []$  
3: for word $\in y$ do  
4: if word $\in \text{stop_words}$ then  
5: $s \leftarrow \text{sample uniform}(0,1)$  
6: if $s < P(\text{remove})$ then  
7: $y' \leftarrow y'$  
8: else  
9: $y' \leftarrow y' + \text{word}$  
10: end if  
11: end for  
12: return $y'$

**Algorithm 2** Frequency Curriculum

**Input:** $y$ - the original target caption  
**Input:** $D$ - difficulty value (Equation 1, 2)  
**Input:** $F$ - list of words ranked in descending order according to its frequency.  
**Output:** $y'$ - modified caption.  
1: $R_{\text{keep}} \leftarrow 0.5(D + 1)$  
2: $F' \leftarrow F[: \text{len}(F) * R_{\text{keep}}]$  
3: $y' \leftarrow []$  
4: for word $\in y$ do  
5: if word $\in F'$ then  
6: $y' \leftarrow y' + \text{word}$  
7: else  
8: $y' \leftarrow y'$  
9: return $y'$

few examples of how the target captions changes according to the epoch value.

### C. Frequency Curriculum

The frequency curriculum for Epochal Difficult Captions entails removing words from captions with a low frequency of occurrence in the entire text corpus. Only high-frequency words are retained in the captions when the difficulty is low. In the initial epochs, the target captions contain words which occur more frequently in the training dataset. As the difficulty level increases with each epoch (Equation 1), the rarer words are added to the target captions. As a result, the captions in the latter epochs are more 'complex,' containing more words. This curriculum is described in detail in Algorithm 2.

We use this frequency curriculum to provide a contrast to the stopwords curriculum (Section III-B).

### IV. Experimental Details

#### A. Data

We use Clotho dataset v2.1 for all our experiments as mentioned in Section II-A. Since we are building on the previous systems of two authors [1], [5], we use the same settings to preprocess the waveforms into the spectrograms as inputs into the models. For both systems, we use 64 Mel-bands, sampling rate of 44100, FFT window length of 1024, and a hop size of 512.

#### B. Training and Evaluation

For our experiments on both System 1 and System 2, we try to replicate the original work without using Epochal Difficult Captions to act as a control experiment, then rerun the same experiments with Epochal Difficult Captions. Where possible, we use the original hyperparameters.

For System 1 [5], we use a batch size of 32 with no gradient accumulation for 30 epochs and 60 epochs with no early stopping. System 1 has two different training setups. The first setup first optimizes on a cross entropy loss for 30
epochs, then uses a reinforcement learning technique, Self Critical Sequence Training (SCST) [12], to train the model for 60 epochs. Learning rate is set to $1 \times 10^{-4}$ for the cross entropy loss and $5 \times 10^{-5}$ for the reinforcement training. SpecAugmentation [13] is applied to all log mel-spectrogram inputs as a data augmentation tactic. We do apply label smoothing. For inference, we perform beam search with beam size of 3 for decoding.

For System 2 [1], we use a batch size of 64 with a gradient accumulation steps of 4 for 100 epochs. Learning rate is set to $3 \times 10^{-4}$ and SpecAugmentation [13] is applied to all log mel-spectrogram inputs as a data augmentation tactic. We do not apply label smoothing. For inference, we perform beam search with beam size of 4 for decoding.

The COCO image captioning evaluation process [14] is used to evaluate the generated caption. There are a few metrics used. The BLEU$_k$ scores [15] measures of n-gram overlap between the generated and reference caption, while ROUGE$_k$ [16] bases its score off the longest common sequence. METEOR [17] uses a harmonic mean of unigram precision and recall to score the generated sequence. CIDEr [18] uses the average cosine similarity between the candidate sentence and the reference sentences to produce a score. Next, SPICE [19] uses semantic propositional information to score the generated captions. Finally, the SPIDEr metric is a combination of both SPICE and CIDEr. We use all the aforementioned metrics for evaluation. The SPIDEr metric is however the metric that researchers try to beat.

V. EXPERIMENTAL RESULTS AND ANALYSIS

We compare the effectiveness of using Epochal Difficult Captions for Curriculum Learning in 3 different settings. The results are shown in Table II. From the table, we observe that the most meaningful metric, the SPIDEr score, increases across all 3 systems. For System 2 particularly, the effect of using Epochal Difficult Captions leads to an improvement of all metrics.

As described in Section III, Epochal Difficult Captions allows the model to gradually transit from training on prominent keywords to the original unmodified caption. We believe this allows the model to focus on more important words such as nouns and action verbs in the earlier stage of the training, thus providing the model with a better ability to recognize objects and actions from the log mel spectrogram. As the training proceeds, filler words and syntactically important terms are consecutively added epoch-wise to enable the model to generate grammatically sound sentences.

A. Comparison with the Frequency Curriculum

The frequency curriculum described in Algorithm 2 can be regarded as being the inverse of the stopwords Curriculum. As opposed to the stopwords curriculum which starts which less frequent words early in the training and introduces more common words as the training proceeds, the frequency curriculum starts with frequent words inclusive of the stopwords, and introduces less common words as the training progresses. As such, the frequency curriculum allows us to provide a contrast to the stopwords curriculum and determine if the improvements in performance for the stopwords curriculum is a result of the stopwords curriculum being a proper extension of the keyword estimation problem.

Table III shows the comparison of the frequency curriculum and the stopwords curriculum on System 2. The frequency curriculum does slightly improve the performance across all metrics. This shows that the frequency curriculum is an effective method to aid the model training. However, the effectiveness of the frequency curriculum pales in comparison to the stopwords curriculum. Therefore, we can infer that the stopwords curriculum is a closer approximation to the keyword estimation problem compared to the frequency curriculum.
VI. CONCLUSION

This work introduces the use of Epochal Difficult Captions for Curriculum Learning for Automated Audio Captioning. Epochal Difficult Captions is an elegant evolution to the keyword estimation problem. We have examined the difficulty curve for Curriculum Learning and described the stopwords curriculum. We have shown that Epochal Difficult Captions can be easily added into any Automated Audio Captioning system during the training stage. We have compared the stopwords curriculum and the frequency curriculum and have shown that the stopwords curriculum is more effective and is a better approximation to the keyword estimation problem compared to the frequency curriculum. Using Epochal Difficult Captions for Curriculum Learning during model training results in an improvement across multiple metrics and the SPIDER score in 3 different settings in 2 systems for Automated Audio Captioning.

ACKNOWLEDGMENT

This research is supported by ST Engineering Mission Software & Services Pte. Ltd under collaboration programme (Research Collaboration No: REQ0149132). The computational work for this research was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

REFERENCES

[1] Andrew Koh, Fuzhao Xue, and Eng Siong Chng, “Automated audio captioning using transfer learning and reconstruction latent space similarity regularization,” 2021.
[2] Yuma Koizumi, Ryo Masumura, Kyosuke Nishida, Masahiro Yasuda, and Shiochirou Saito, “A Transformer-based Audio Captioning Model with Keyword Estimation,” arXiv:2007.00222 [cs, eess, stat], Aug. 2020, arXiv: 2007.00222.
[3] Zhongjie Ye, Helin Wang, Dongchao Yang, and Yuexian Zou, “Improving the performance of automated audio captioning via integrating the acoustic and semantic information,” 2021.
[4] Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda, “Audio Captioning using Pre-Trained Large-Scale Language Model Guided by Audio-based Similar Caption Retrieval,” arXiv:2012.07331 [cs, eess], Dec. 2020, arXiv: 2012.07331.
[5] Xinshao Mei, Qiushi Huang, Xubo Liu, Gengyue Wu, Yucong Wu, Jinzheng ZHAO, Shengchen Li, Tom Ko, H. Tang, Xi Shao, Mark D. Plumbley, and Wenwu Wang, “An encoder-decoder based audio captioning system with transfer and reinforcement learning,” in Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021), Barcelona, Spain, November 2021, pp. 206–210.
[6] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, “Clotho: An Audio Captioning Dataset,” arXiv:1910.09387 [cs, eess], Oct. 2019, arXiv: 1910.09387.
[7] Chris Donggug Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, “AudioCaps: Generating Captions for Audios in The Wild,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, pp. 119–132, Association for Computational Linguistics.
[8] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu, “Audio Caption in a Car Setting with a Sentence-Level Loss,” arXiv:1905.13448 [cs, eess], Oct. 2020, arXiv: 1905.13448.