On Memorability and Style of Audio Features in Multimedia Evaluation

Yutong Zhao

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-94
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-94.html

May 14, 2021

Acknowledgement

I would like to acknowledge the support of my research advisor Professor Gerald Friedland, who made it possible for me pursue academic research and a graduate degree. I would like to thank Irving Fang, Jeffrey Kim, Benny Sihao Chen, and Anusha Dandamudi for their assistance with my research. Finally, I thank my family and Teresa Pho for providing me the endless support I required while undertaking this experience.

On Memorability and Style of Audio Features in Multimedia Evaluation

by

Tony Zhao

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gerald Friedland, Chair
Professor Gopala Krishna Anumanchipalli, Second Reader

Spring 2021

The thesis of Tony Zhao, titled On Memorability and Style of Audio Features in Multimedia Evaluation, is approved:

Chair Ass. Adj. Prof. [signature] Date May 14, 2021

Prof. Gopala Anumanchipalli, EECS [signature] Date 05/14/2021

Date

University of California, Berkeley

On Memorability and Style of Audio Features in Multimedia Evaluation

Abstract

On Memorability and Style of Audio Features in Multimedia Evaluation

by

Tony Zhao

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Gerald Friedland, Chair

The absolute quantity of content on multimedia platforms such as social networks has grown exponentially in the past few years. As such, algorithms for automatic evaluation for digital content become increasingly important in enhancing the relevancy of retrieval, especially the ability to capture abstractions of more abstract aspects such as beauty, artistry, interestingness. Particularly, we see a need for models that can evaluate abstract notions of digital content such as style and memorability.

While there has been development and achievement in using artificial neural networks for predicting image memorability as well as for image style transfer, such success have not completely been replicated in different media modalities such as video or audio. Given the broad topic of multimedia, this report seeks to focus on the background and context of audio features in a framework of multimedia evaluation. We also present our experiments for establishing use of audio features in both predicting video memorability as well as in audio style transfer. Furthermore, this work outlines both current challenges as well as the path forward for systematic augmentation and optimization of multimedia attributes such as memorability and style.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to acknowledge the support of my research advisor Professor Gerald Friedland, who made it possible for me pursue academic research and a graduate degree. I would like to thank Irving Fang, Jeffrey Kim, Benny Sihao Chen, and Anusha Dandamudi for their assistance with my research. Finally, I thank my family and Teresa Pho for providing me the endless support I required while undertaking this experience.

Chapter 1

Introduction

Multimedia platforms have grown exponentially in the amount of data over the past decade, arising from social networks or advertiser content. Therefore algorithms for enhancing the relevancy of retrieval of digital content ever increasingly important.

Human cognition has a massive capacity for understanding and remembering media. Despite the huge variance in the nature of digital content, humans tend to remember and forget the same videos. Humans also can agree on a cohesive thematic style of a given group of similar video content as well as make distinctions between different artistic styles. These phenomena suggests that despite varying personal experiences and priors, people share some innate characteristics in encoding, understanding, and discarding the same types of information. A simple example is the fact that humans can easily identify and recognize images and videos of people, salient activities, distinct events, or landmarks. However, images and videos of scenes that lack distinctiveness such natural landscapes are far less memorable. This indicates that memorable and forgettable images differ in their intrinsic visual features that make retaining information more or less difficult.

Among important metrics such as aesthetics or novelty, memorability is an important aspect to consider when evaluating multimedia and subsequent retrieval. This is especially relevant in the domain of advertising or educational content, highlighting the need for the capability of predicting the memorability of a particular piece of video content.

We took inspiration in modeling media memorability through multi-modal ensemble methods. That is, we used multiple features of different modalities such as image and text embeddings to predict media memorability. Our novel contribution is the demonstration that audio embeddings can be highly effective as features for predicting video memorability.

Beyond predicting memorability with audio embeddings, we want to identify algorithms that can optimize memorability without altering the underlying content. These algorithms would directly involve the ability to separate the non-content aspects, or *style*, from digital content. Such style embeddings can then be independently synthesized with other content, also known as style transfer. Audio style transfer is still an open research area, and we introduce our approach to evaluating the ability of style transfer algorithms to preserve information for multimedia evaluation.

Chapter 2

Background and Related Work

2.1 Measuring memorability: The Memory Game

Computational understanding of media memorability follows on from the task of image memorability, established by the seminal work of Isola et al. [11]. The work defines an image's memorability as the probability that an observer will detect a repetition of the same image a few minutes after the first exposure, when presented as a streaming sequence of image. Such a method of measuring memorability is generally referred to as the "Memory Game". This definition serves to simplify the measurement of memorability for data mining purposes, laying the groundwork for training memorability predictors.

For participants in the Memory Game, they are presented with a sequence of digital content elements, either images or videos, which are comprised of "target" elements as well as "filler" elements, which are usually random sampling of mutually exclusive subsets of a large reference dataset. The participants are not made aware of the distinction between target elements and filler elements. The first role of filler elements is as spacing and content in between the first appearance and second appearance of a target element. The other role filler elements play are to be "vigilance" tasks, where filler elements are repeated to continuously check that a participant is attentive to the task at hand. Each target element is expected to repeat only once, while filler elements are presented once unless it was a vigilance task filler in which case it was repeated twice.

Generally, criteria are used to ensure consistency and performance of participants once they entered the game. Generally these criteria are based on both vigilance tests as well as setting the max amount of times a participant can enter the game.

Finally, after collection of data, each element is assigned a *memorability score*, which is defined as the percentage of correct detections by participants. This memorability score $m^{(i)}$ can also be interpreted as the average hit rate per element and the value that minimizes the l_2 error

$$\sum_j ||x_j^{(i)} - m^{(i)}||_2^2 \Rightarrow \frac{1}{n^(i)} \sum_j x_j^{(i)}$$

Experimentally, we see that memorability follows a log-linear relationship with the time delay between repetitions. However, work has shown that rank memorability stable over time [11]. This indicates that a highly memorable element will remain more memorable than a unmemorable element, even if memorability degrades over time in both cases. Practically, this degree of stability indicates that we can model rank memorability as time-independent.

Khosla et al. [12] expands on memory degradation as a function of time, which allows for data collection with highly variable time delays. We assume that the memorability of element i is $m_T^{(i)}$ where the time interval between repeated displays is T. This allows memorability of element i to be modeled as

$$m_T^{(i)} = \alpha \log(T) + c^{(i)}$$

where $c^{(i)}$ is the *base memorability* for the given element and α is the decay factor of memorability over time. We hypothesize that each participant in the memory game may have slightly different decay factors based on differing abilities of recognition, but the variance is low enough to ignore individual differences between participants.

Since this model of memorability is a function of time, we obtain the relationship

$$m_t^{(i)} - m_T^{(i)} = \alpha \log(t) - \alpha \log(T)$$

$$\Rightarrow m_t^{(i)} = m_T^{(i)} + \alpha \log\left(\frac{t^{(i)}}{T}\right)$$

In the case of a memory game with n observations for element i, we have $x_j^{(i)} \in \{0, 1\}$ equal to 1 if the repeated element was correctly recognized after time $t_j^{(i)}$ and 0 otherwise. Finally, for N images, we can calculate the adjusted memorability score by optimizing the overall l_2 error, E as

$$E(\alpha, m_T^{(i)}) = \sum_{i=1}^{N} \sum_{j=1}^{n^{(i)}} ||x_j^{(i)} - m_{t_j}^{(i)}||_2^2$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n^{(i)}} \left|\left| x_j^{(i)} - \left[m_T^{(i)} + \alpha \log\left(\frac{t_j^{(i)}}{T}\right)\right]\right|\right| \tag{2.1}$$

2.2 Predicting memorability

We have seen models, notably MemNet [12], that have achieved near-human level of performance for the task of predicting image memorability. These techniques apply the recent success of convolutional neural networks (CNNs) in visual recognition tasks.

The metric of choice is Spearman's rank correlation coefficient when evaluating these models. This metric assesses how well the relationship between two variables can be described using a monotonic function. Spearman's rank r_s can be computed as the Pearson correlation coefficient of the rank variables. Thus for a sample of size n, the n raw values X_i, Y_i are

converted to ranks g_{X_i}, g_{Y_i}, and r_s is computed as

$$r_s = \rho_{g_X, g_Y} = \frac{\mathbf{cov}(g_X, g_Y)}{\sigma_{g_X} \sigma_{g_Y}}$$

where $\mathbf{cov}(g_X, g_Y)$ is the covariance of the rank variables, σ_{g_X} and σ_{g_Y} are the standard deviations of the rank variables, and ρ denotes the Pearson correlation coefficient.

While predicting image memorability has proven to work reliably, research on video memorability is still in the early stage, as the nature of video content makes evaluating memorability difficult. Spearman's rank correlation is still the metric of choice for video memorability.

Borrowing from work in image memorability, image-based features extracted from pre-trained convolutional neural networks have been demonstrated as useful in predicting the memorability of videos.

Related work [23] has also highlighted the usefulness of other features such as semantic, saliency, and colour features. Image captioning models have proved effective for predicting memorability scores [5]. Video-based features, such as C3D and I3D, have also recently been considered in the study of video memorability [22]. Finally, the best performing models of the 2019 iteration of the Predicting Media Memorability challenge utilized ensemble models with the above-mentioned features [2].

2.3 Style transfer

The seminal work establishing neural algorithms of artistic style was introduced by Gatys et al. with the use of convolutional neural networks [6]. This approach separates the *style* from the *content* of two different images and recombines them to create a image in the target style, see Figure 2.1. The feature responses in the layers of a convolutional network act as the content representations of an image. Meanwhile the style representation is instead defined as the correlation between the different feature responses. By including the feature correlations of multiple layers, we can obtain a stationary, multi-scale representation of the input image, which captures its texture and style information but not the global arrangement.

Work in audio style transfer is still ongoing, as replicating the success of image style transfer is an open research area. The format of audio data presents the first challenge, as it is 1-dimensional as opposed to the 2-dimensional nature of image data. Generally the approach for neural algorithms for audio is to first transform the audio waveforms into 2-dimensional spectrograms, commonly using Mel-frequency cepstral coefficients. These can then be fed into convolutional neural networks analogous to image neural algorithms. Unfortunately, style transfer algorithms for images do not immediately perform similarly for audio as demonstrated [8].

One of the immediate challenges identified by Grinstein et al. is the lack of a clear and formal definitions of style and content in audio. For images, style generally encompasses the space-invariant intra-patch statistics such as texture, while content refers to the broad

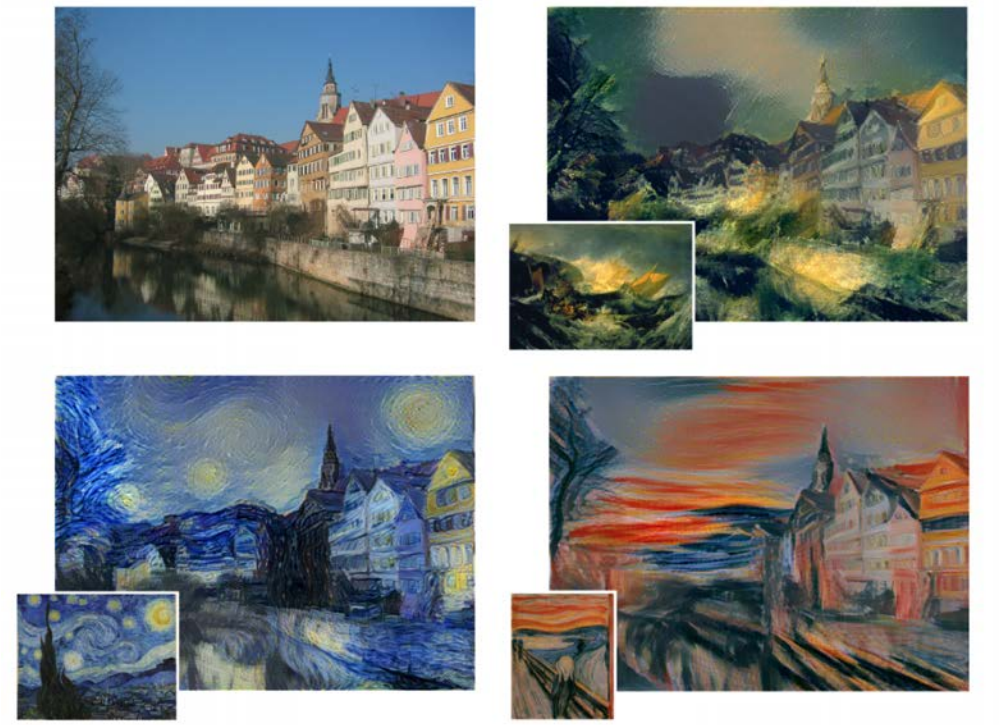

Figure 2.1: Neural style transfer as demonstrated by Gatys et al. [6]

semantic and geometric composition of the scene. In audio, style and content are even more difficult to define and can vary depending on the task at hand. [8] For speech, content may refer to linguistic information such as phonemes and words while style may refer to the particularities of the speaker such as emotion, accent, and intonation. For music, content may be the underlying melody while style may be the timbre and musical genre.

In our style experiments, we focus primarily in speech style transfer, also known as voice conversion, and the ability to preserve information after synthesis. This has been a developing field as multiple approaches have been proposed with some successes in voice conversion. Qian et al. [19] proposed AutoVC as an autoencoder-based approach to voice conversion with a carefully-designed bottleneck. Lin et al. [15] instead introduces FragmentVC which relies on attention mechanisms in their transformer encoder blocks to extract and fuse voice fragments.

2.4 Applications of style transfer

Recently, use of techniques such as style transfer have indicated that that the field has advanced from simply measuring image memorability to using memorability as an evaluation metric. Siarohin et al. [24] introduced an algorithm that extracts *style seeds* from several

target images which are then applied as filters onto source images, before retrieving the styles that maximize the predicted memorability of an image.

Automatic speech recognition (ASR) models have been proven to have performance disparities between different accents [14]. A very promising solution has been to use audio style transfer as accent modification to increase the accuracy of ASR models on non-native speakers [20]. Such methods can help reduce the necessity to collect and train with non-native accents during ASR model training, mitigating the obstacle of data scarcity.

Chapter 3

Memorability

3.1 Setup

The dataset is composed of a subset of short videos selected from TRECVid 2019 Video-to-Text dataset [1]. Each TRECVid video is accompanied by textual captions describing the content of the video.

Each video also has an associated memorability score for both the short-term where scores a measured a few minutes after the memorization process, and the long-term where scores are measured 24-72 hours after the memorization process. Memorability score refer to the probability of a video to be remembered after a time duration and is measured using recognition tests.

In the memory game tests, participants are expected to watch 180 and 120 videos in the short-term and long-term memorization steps respectively. In the first step of the evaluation, target videos are repeated after a few minutes to collect short-term memorability labels. After 24 hours to 72 hours, the same participants are expected to attend the second step for collecting long-term memorability labels. The participant is directed to press a button indicating recognition of a previously seen video. Both the short-term and long-term memorability scores are calculated as the percentage of correct recognition for each video by the participants.

The training set comprises of 590 short videos (1-8 seconds long) with 2-5 human-annotated captions each. Another development set of 410 additional videos was made available later in the project, but was not used as we discovered the quality of annotated memorability scores to be worse than that of the original training set.

Memorability score adjustment

The short-term and long-term memorability scores of each video were adjusted using methods outlined in previous work (See Equation 2.1). This adjustment involves iteratively updating

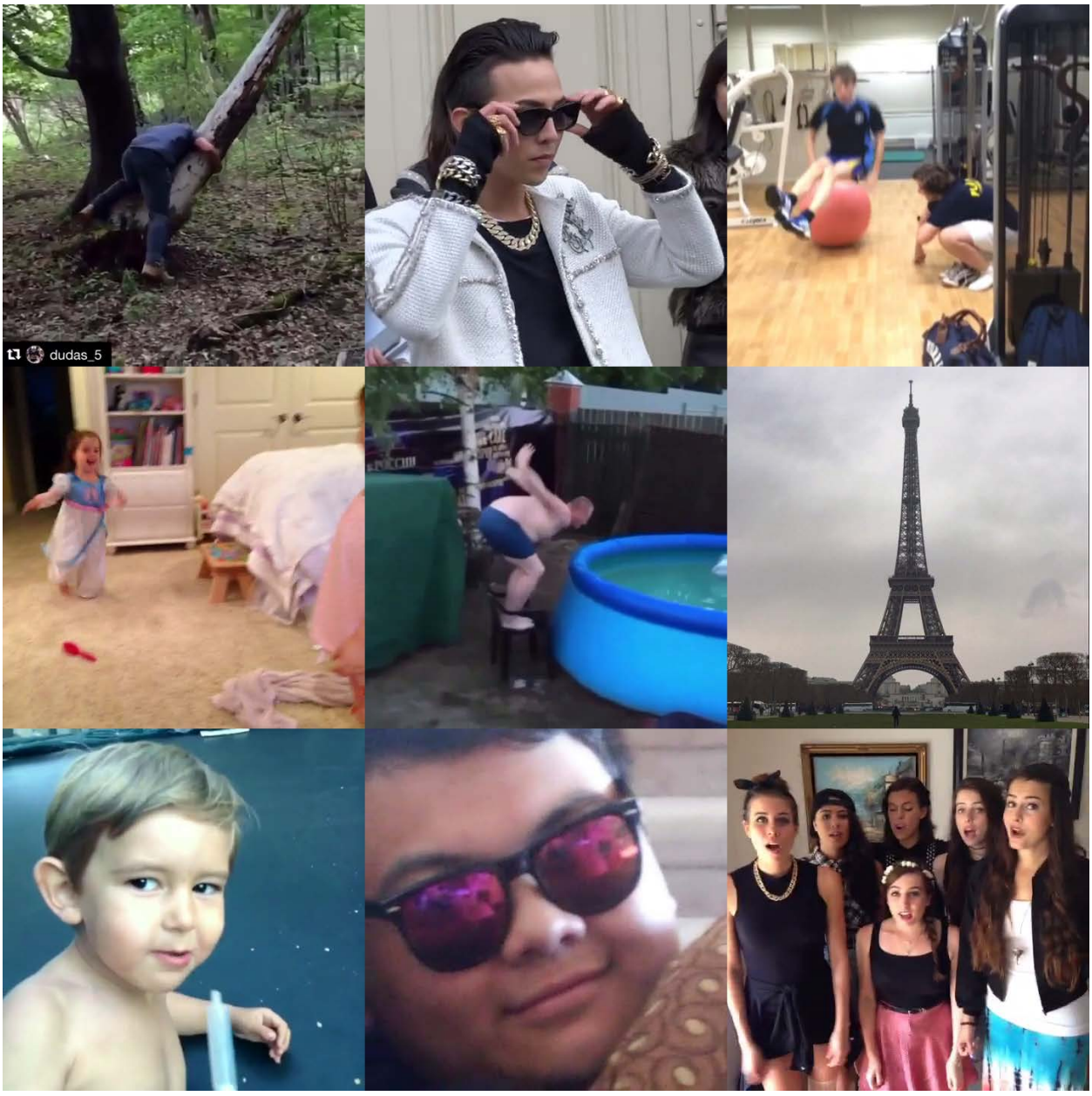

Figure 3.1: Selected frames from videos with high memorability scores

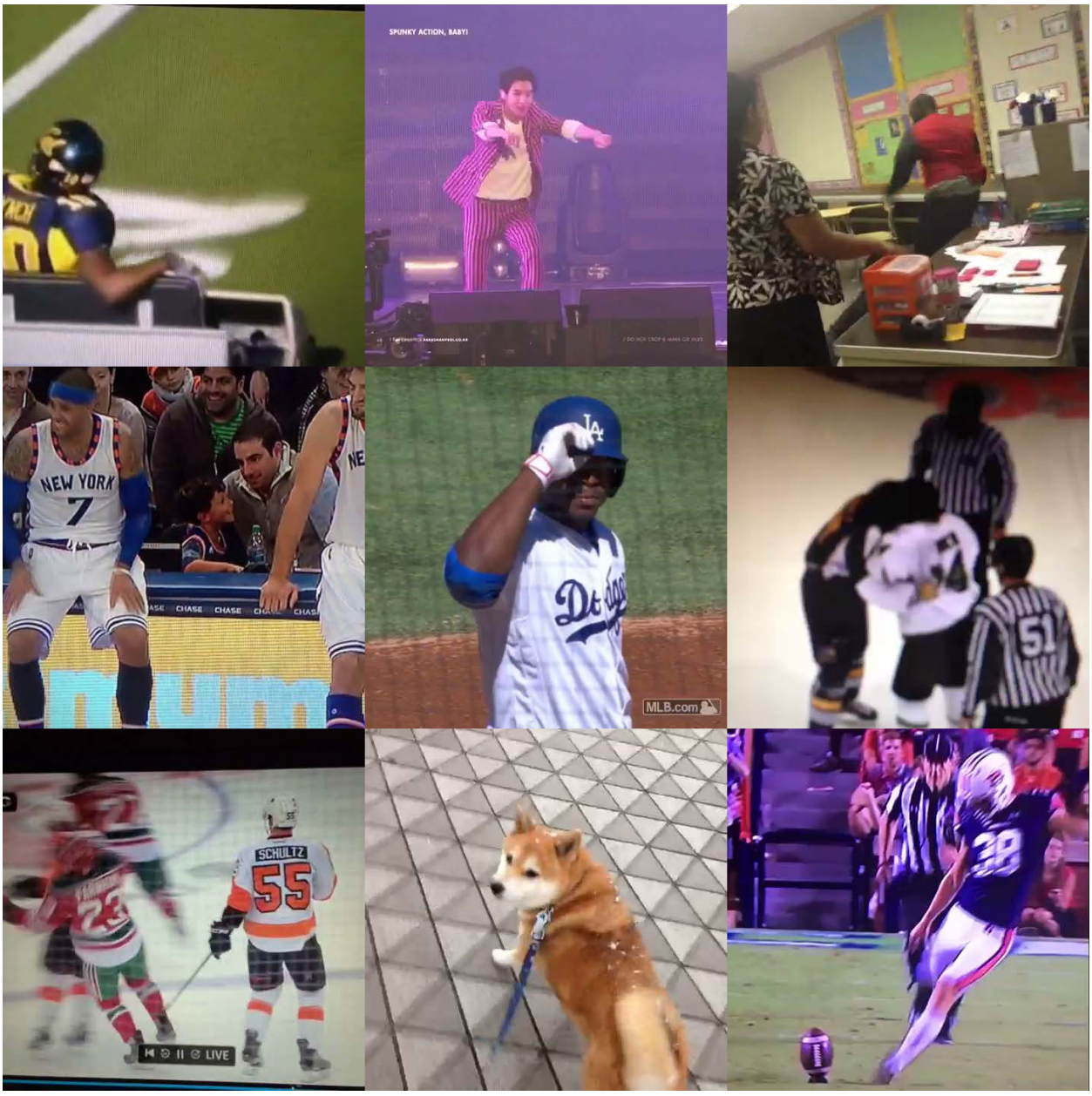

Figure 3.2: Selected frames from videos with low memorability scores

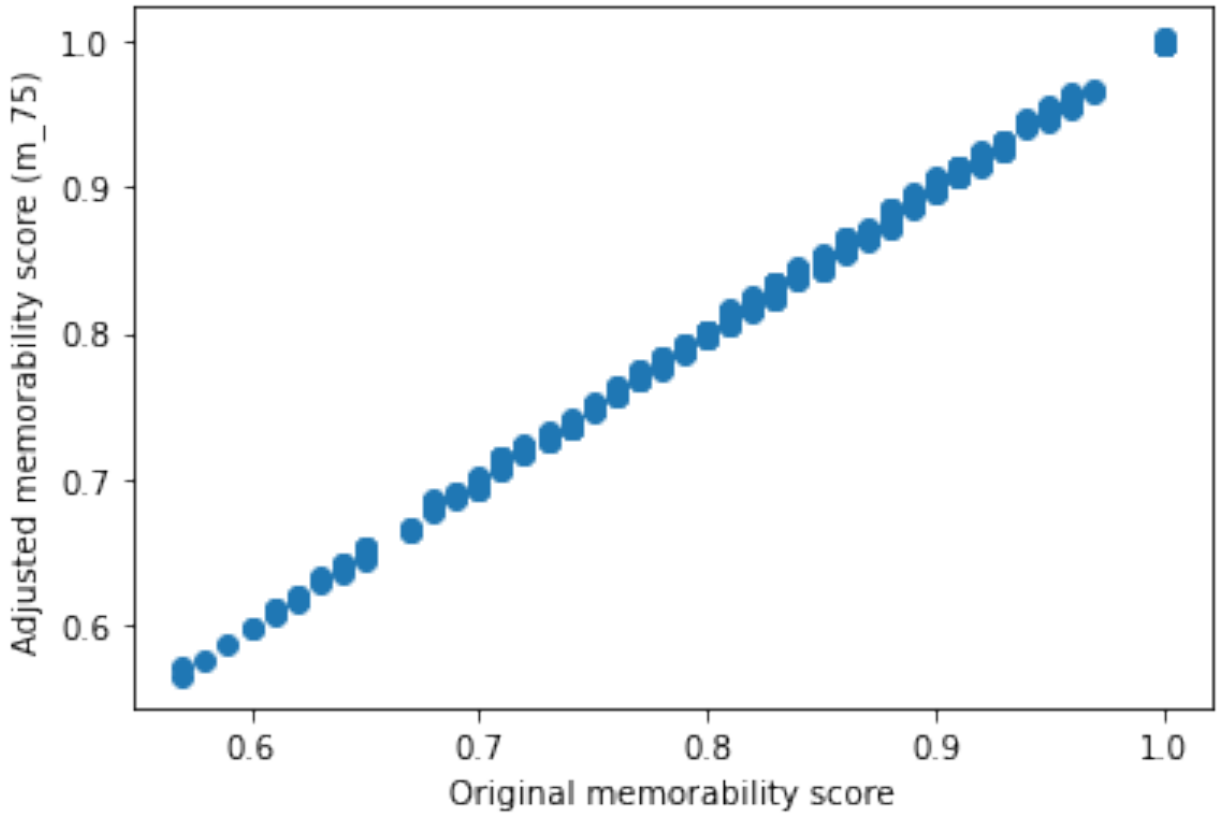

Figure 3.3: Original memorability scores and corresponding adjusted memorability scores

and calculating decay rate α and memorability m_T at duration T

$$\alpha \leftarrow \frac{\sum_{i=1}^{N} \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} \log(\frac{t_j^{(i)}}{T})[x_j^{(i)} - m_T^{(i)}]}{\sum_{i=1}^{N} \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} [\log(\frac{t_j^{(i)}}{T})]^2}$$

$$m_T^{(i)} \leftarrow \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} [x_j^{(i)} - \alpha \log(\frac{t_j^{(i)}}{T})]$$

where we have $n^{(i)}$ observations for video i given by $x^{(i)} \in \{0, 1\}$ and $t_j^{(i)}$ where $x_j = 1$ implies that the repeated video was correctly detected when shown after time t_j. The average t was 74.96, so we adjusted the target memorability score for each video to be m_{75} after 10 iterations with α converging to -0.0264. Overall, we did not notice any significant difference with or without the memorability score adjustments (See Figure 3.3). We hypothesize that memory score adjustment would be far more significant in cases where the time interval between repetition of the same element has high variance.

3.2 Approach

Models were trained on a 80-20 training-validation split on video id. Since concatenating multiple multi-modal features resulted in extremely high-dimensional feature vectors which were difficult to train with, we trained support vector regressor (SVR), Bayesian Ridge regressor, and linear models on each individual feature independently to obtain feature-specific, and therefore modality-specific, models.

We also explored several prediction aggregation functions for when a video has multiple embeddings of a specific feature, such as several different model predictions due to multiple captions for a given video. While previous work generally defaulted to a simple average, we discovered that taking the median value performed consistently better than either mean, max, or min. We hypothesize this behavior is due to the high variance in the output of the models. In the opposite scenario where a video has no feature such as a soundless video for audio-based models, we defaulted to using the average of all predicted memorability scores.

We noticed high variance in the performance of the models, which we attributed to the relatively small dataset. Therefore we ran the feature models over 5 random seeds and took the best performing features of each modality (VGGish for audio, ResNet152 for image, C3D for video, GloVe for text). The predictions of these models were then used to perform grid-search over permutations of buckets of 5% to calculate a weighted average as our final ensemble model.

Audio features

Since previous work on the topic of predicting video memorability has not included the use of audio features, we present a novel contribution in the usage of pre-trained CNNs to extract audio embeddings as features. This approach is largely inspired from similar architectural patterns with extracting image embeddings. From each video, we first extracted the audio content of the file. Using VGGish [10], a pre-trained CNN model (trained on AudioSet[7]), we extracted 128-dimensional embeddings for each second of video audio. We hypothesized that such embeddings would be a useful feature as most of the videos did not contain speech but instead background noise of the activity at hand. We noticed that highest scoring videos tend to have either music or loud explosive sound effects.

Each video had on average had 5.6 embeddings extracted, with one soundless video having none. Bayesian Ridge Regressor provided the best performance on these features. See Figure 3.4.

Image/Video Features

Local Binary Patterns (LBP), VGG [16], and Convolutional 3D (C3D) proved to have notable performance among the provided features. We discovered that Support Vector Regressor [3] models produced the best results with these features.

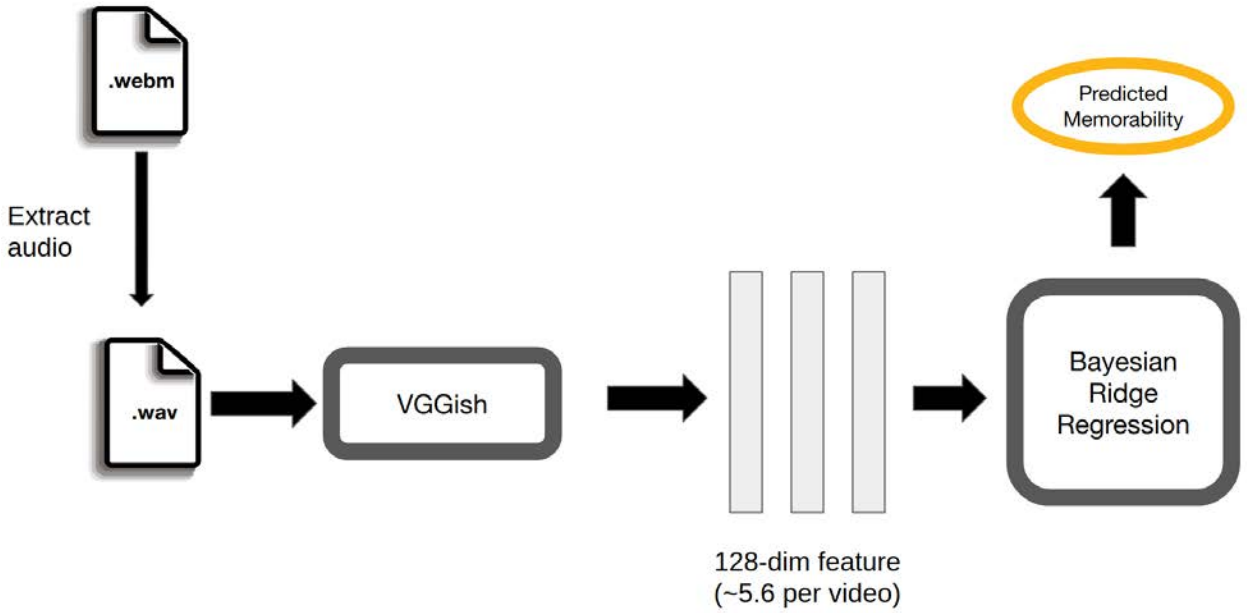

Figure 3.4: Audio feature extraction and model

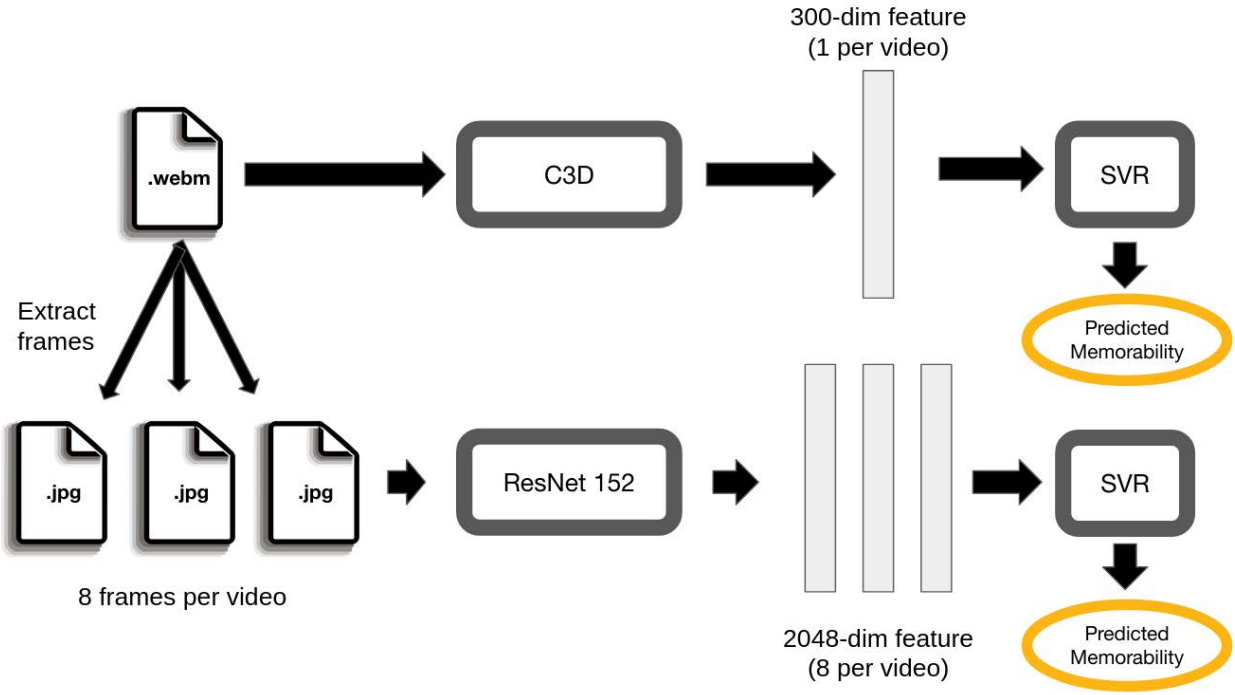

Figure 3.5: Image/Video feature extraction and models

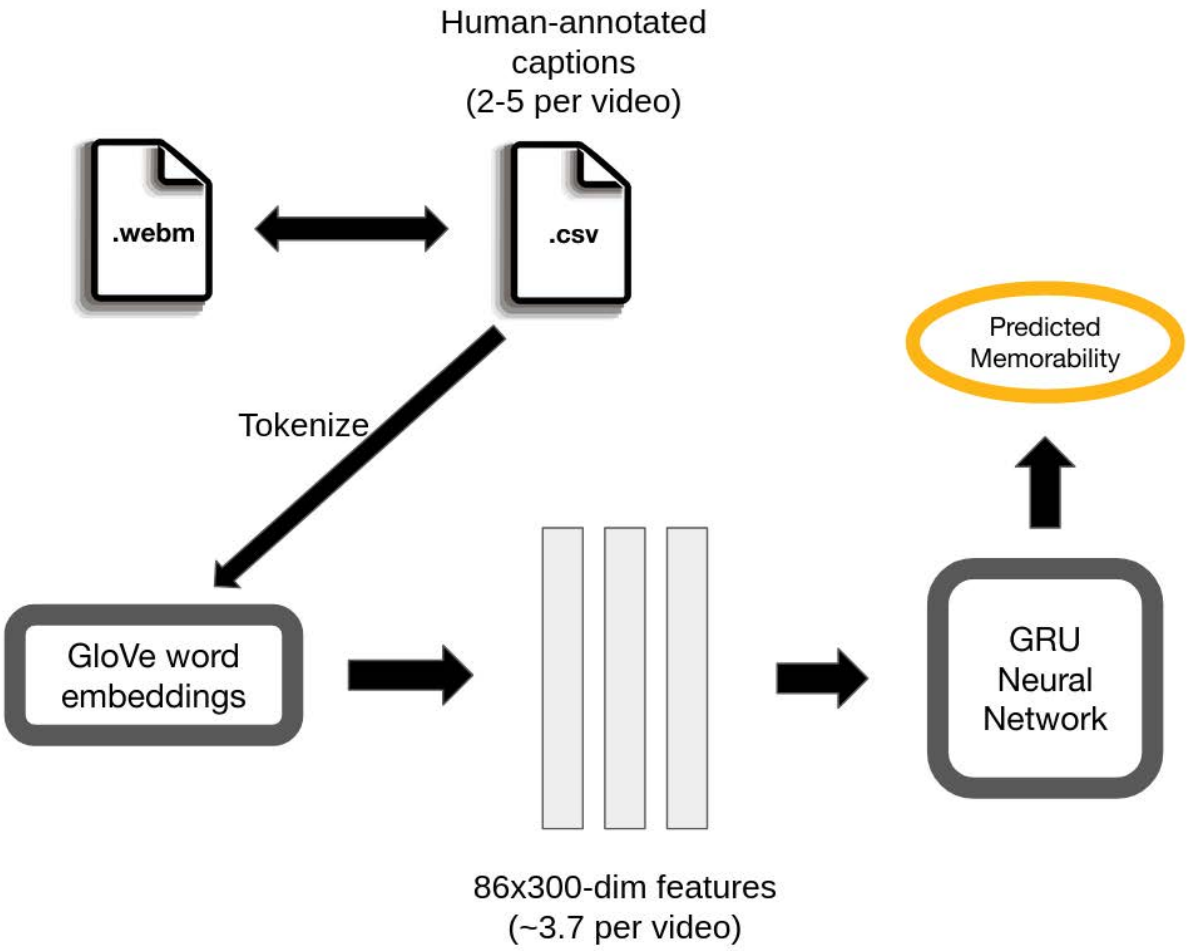

Figure 3.6: Text model

We also extracted 8 equally spaced frames from each video, which were used for our own image-based feature extraction. The penultimate layer of a pre-trained ResNet-152 [9] (trained on ImageNet) was used to extract a 2048-dimensional feature vector. Similarly, we found that the best performance came from Support Vector Regressor models trained on the extracted features. See Figure 3.5.

Following the methods outlined in [2], pre-trained facial-emotion models were used to extract emotion-based features on the video frames. Ultimately, these features were not used in any of the final models, as only approximately a third of the videos had detectable faces as well as overall poor performance from all emotion-based models.

Table 3.1: Spearman's rank correlation coefficient (SRCC) of features for short-term memorability over 5 runs trained and evaluated on training set (80-20 train-validation split on 590 videos)

Modality	Feature	Model	Mean	Variance
video	**C3D**	SVR	0.152	0.081
image	**ResNet152**	SVR	0.233	0.096
image	VGG	SVR	0.167	0.058
image	LBP	SVR	0.139	0.072
audio	**VGGish**	Bayesian Ridge	**0.246**	**0.017**
text	**GloVe**	GRU	0.200	0.029

Text Features

Provided with several human-annotated captions for each video, we explored several different methods to extract semantic features. Past work suggested that simpler methods like bag-of-words could outperform more sophisticated methods in terms of both effectiveness and efficiency [21]. We vectorized the captions using bag-of-words before training with ordinary least squares, ridge, and lasso regression models. Bag-of-words vectorization and linear models performed the worst among our text-based approaches and were not used in the final model. We hypothesize that given the unique nature of the videos, simpler models do not capture the positional information in the captions that are helpful in predicting memorability.

We tokenized the captions before extracting 300-dimensional GloVe [18] word embeddings (trained on Wikipedia 2014 and Gigaword 5) to vectorize the caption tokens. These vectors were used as input to train a recurrent neural network with gated recurrent units (GRU) [4]. Our final text-based model had 64 units in the initial GRU layer with a dropout of 0.8, followed by 4 dense layers with a dropout of 0.25 and ReLU activation, trained for 150 epochs with early stopping, a learning rate of 0.001, batch size of 64, and an Adam optimizer. See Figure 3.6.

We also explored machine-generated captions [17] to augment our text-based approaches. After generating captions for each video based on the first, middle, and last frames and mixing the generated captions with the human-annotated captions, we discovered that the performance improvement was insignificant. These automatic captions were not included in our final models.

3.3 Results and Analysis

The notable features are seen in Table 3.1, with the bolded features being selected for our final ensemble models, seen in Table 3.3 and Table 3.4. Given the small dataset, we observe relatively high variance in performance.

Table 3.2: Spearman's rank correlation coefficient (SRCC) of features for long-term memorability over 5 runs trained and evaluated on training set (80-20 train-validation split on 590 videos)

Modality	Feature	Model	Mean	Variance
video	**C3D**	SVR	0.076	0.059
image	**ResNet152**	SVR	0.133	0.066
image	VGG	SVR	**0.144**	0.092
image	LBP	SVR	0.024	0.067
audio	**VGGish**	Bayesian Ridge	0.059	**0.026**
text	**GloVe**	GRU	0.104	0.091

Table 3.3: Short-term memorability ensemble models, Validation SRCC on 20% of training set, Test SRCC on testing set

Model	C3D	ResNet152	VGGish	GloVe	Validation	Test
1	0.20	0.35	0.45	0.00	0.343	**0.136**
2	0.00	0.20	0.35	0.45	0.345	0.116
3	0.05	0.00	0.50	0.45	**0.370**	0.085
4	0.00	0.50	0.15	0.35	0.357	0.091
5	0.35	0.00	0.30	0.35	0.317	0.102

Table 3.4: Long-term memorability ensemble models, Validation SRCC on 20% of training set, Test SRCC on testing set

Model	C3D	ResNet152	VGGish	GloVe	Validation	Test
1	0.00	0.40	0.15	0.45	**0.289**	0.012
2	0.55	0.10	0.00	0.35	0.192	0.076
3	0.00	0.55	0.45	0.00	0.118	0.044
4	0.25	0.35	0.20	0.20	0.168	**0.077**
5	0.30	0.05	0.00	0.65	0.201	0.056

Our audio features from VGGish had the best performance, smallest variance, and thus presumably the best generalization for short-term memorability. However, the same audio features tended to perform poorly for predicting long-term memorability despite maintaining low variance. An explanation may be that visual memory may be more resilient to degradation over the long-term than auditory memory in humans.

The difference in validation and test Spearman's rank correlation in our final ensemble models are likely due in part to overfitting after performing grid-search over a relatively small dataset. In addition, we noticed that the quality of annotations between datasets were different, which may cause distribution differences in memorability scores and consequently poorer performance at test time.

3.4 Discussion

Our main contribution to the field of media memorability is the demonstration that audio-based models perform well for predicting short-term memorability and can generalize much more readily than other methods with the dataset. This may be in part due to the low dimensionality of the extracted audio embeddings, which are only have 128 dimensions while our ResNet152 image embeddings are 2048-dimensional. While we still believe that visual memory is the least resilient to degradation over time, we reiterate that ensembling models of different modalities still achieve the best performance as each model can represent a different high-level abstraction of the data.

Qualitatively, we noticed that videos with high memorability (See Figure 3.1) scores tended to be highly unusual and distinct from other videos in the dataset. On the other hand, videos with low memorability scores (See Figure 3.2) tended to be clustered around visually similar topics such as sporting events. These visually similar videos in the distribution may be difficult for human annotators to identify memorably distinct features, resulting in lower memorability scores.

As for our audio model, some cursory exploration seems to indicate that it rates videos with violent, explosive noises and videos with music very highly in predicted memorability. Videos with low predicted memorability seemed to have a lot of background noise, such as crowds in a sportscast video. We hypothesize that the audio model is in part rating how surprising the audio from a video is. More systematic evaluation of the behavior of audio features would necessitate adding auditory tags to the videos.

Chapter 4

Style

4.1 Setup

For our work on evaluating style transfer models, we elected to used the Speech Accent Archive [26] as well as the Voice Cloning Toolkit (VCTK) [27]. The Speech Accent Archive is comprised of a diverse collection of recordings of various accents enunciating the same passage. The files are relatively long at about 30 seconds in duration. On the other hand, VCTK is a dataset that was designed for voice cloning in mind, and thus elects to have much higher quality, but shorter utterances. Each speaker in VCTK have several hundred utterances that are just a few seconds long. However, the diversity of the accents in VCTK is comparably less extensive than Speech Accent Archive.

Thus, evaluating the performance on VCTK helps show how well the models can preserve the information in high-quality, in-distribution examples. On the other hand, the Speech Accent Archive presents more noisy and consequently more challenging audio data, helping to illustrate how well these models can generalize and perform in the real world.

We assess the performance and accuracy of in terms of the average word error rate (WER), a standard measure of discrepancy between human and machine transcriptions [13]. With N words in the ground truth transcription, WER is formally defined as

$$\text{WER} = \frac{S + D + I}{N} \tag{4.1}$$

where S, D, and I are respectively the number of word substitutions, deletions, and insertions between the ground truth and generated transcription. Thus a higher WER would indicated a greater discrepancy between two transcription and consequently worse automatic speech recognition (ASR) performance according to our evaluation.

We elected to use a state-of-the-art voice conversion algorithm as our style transfer model. FragmentVC is an attention-based approach, extracting and synthesizing source and target utterance attributes based on the attention mechanism of transformers [15].

For our ASR model, we used Silero, an enterprise-grade speech-to-text model [25]. Our

evaluation approach would thus be comparing the ground truth transcription with the decoded output of Silero.

4.2 Approach

First, we confirmed the accent disparities found by Koenecke et al. [14] by grouping by either the provided native language and accent for each speaker in the Speech Accent Archive and VCTK respectively. We discarded any group with significantly low number of distinct speakers to control the variability of our experiments, as exceptionally poor or great performance of a given speaker would otherwise add significant bias to our evaluation. We then established a baseline WER for the group by average the WER of all speakers contained in the group by extracting transcription hypotheses. For example, the baseline WER for American speakers would be the average WER between the ground truth and the Silero transcription hypothesis of all speakers identified as American. We designated poor performing groups as source groups while better performing groups are designated as our target groups.

We then randomly sampled uniformly from both the source groups and target groups to create source-target speaker pairs. Using our style transfer models, we extracted the style of the target speaker and combine with the content of the source speaker utterances. Finally, with the synthesized utterances, we extracted transcription hypothesises using our ASR model. The WER is calculated between the ground truth and the hypothesis and finally compared to the WER of the original source utterance.

When evaluating our model on Speech Accent Archive, we realized that the data was significantly distorted to the point where it was unrealistic to expect either human or speech-to-text models to transcribe. This is likely due to the longer file format as well as the noisier samples in the dataset. Therefore we could not obtain consistent results for audio style transfer and elected to omit them in our report.

4.3 Results and Analysis

We confirmed that there is a large disparity between the performance of our ASR model native English speakers and non-native English speakers in the Speech Accent Archive as shown in Figure 4.1. Likewise, we identified that Indian accents performed poorly compared to other accents in the VCTK Corpus. Irish and Scottish accents to lesser degree also performed poorly as seen in Figure 4.2.

We noticed that the WER among utterances in Speech Accent Archive was higher than those from VCTK, which we attributed to the respective formats of the two datasets. The shorter duration and higher quality in the recordings of VCTK likely were easier for the ASR model to transcribe than those from the Speech Accent Archive.

Generally, we see that style transfer almost invariably causes WER to increase. In a few particular instances, we did observe a decrease in WER, but these were the exception. The

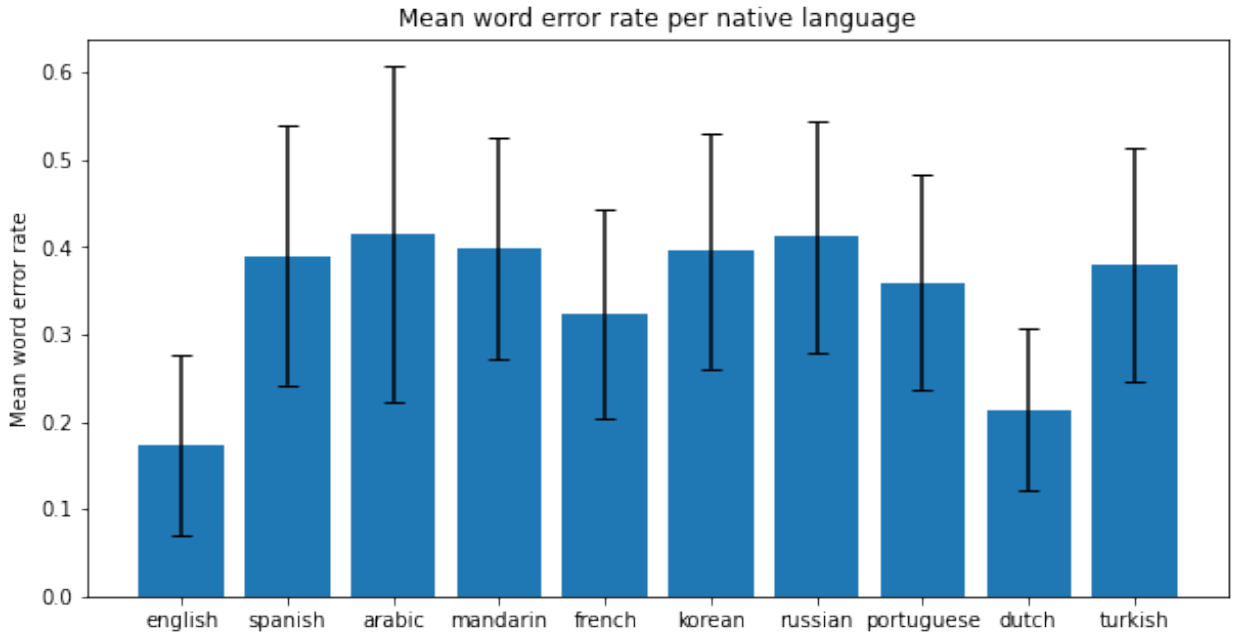

Figure 4.1: Mean error rate per native language in the Speech Accent Archive

Table 4.1: Word error rate difference after audio style transfer with FragmentVC

Source Speaker	Target Speaker	Original WER	Stylized WER	Relative Difference
Indian	American	0.28132	0.61938	120.16%
Indian	English	0.36966	0.75491	104.21%
American	American (Identical)	0.08511	0.25391	198.31%
American	American	0.09483	0.28654	202.17%
Irish	American	0.20263	0.47637	135.09%
Scottish	American	0.22282	0.46183	107.27%
English	English (Identical)	0.12750	0.33904	165.91%
English	English	0.09405	0.32091	241.20%

relative WER ranking of the source speaker was overall preserved by style transfer.

4.4 Discussion

From the results in 4.1, we see that the word error rate significantly increases after style transfer. On inspection of the outputs of style transfer, we qualitatively affirm that the algorithm could successfully synthesize reasonable waveforms with the content of the source speaker and the style of the content speaker, albeit with some audible distortions. This

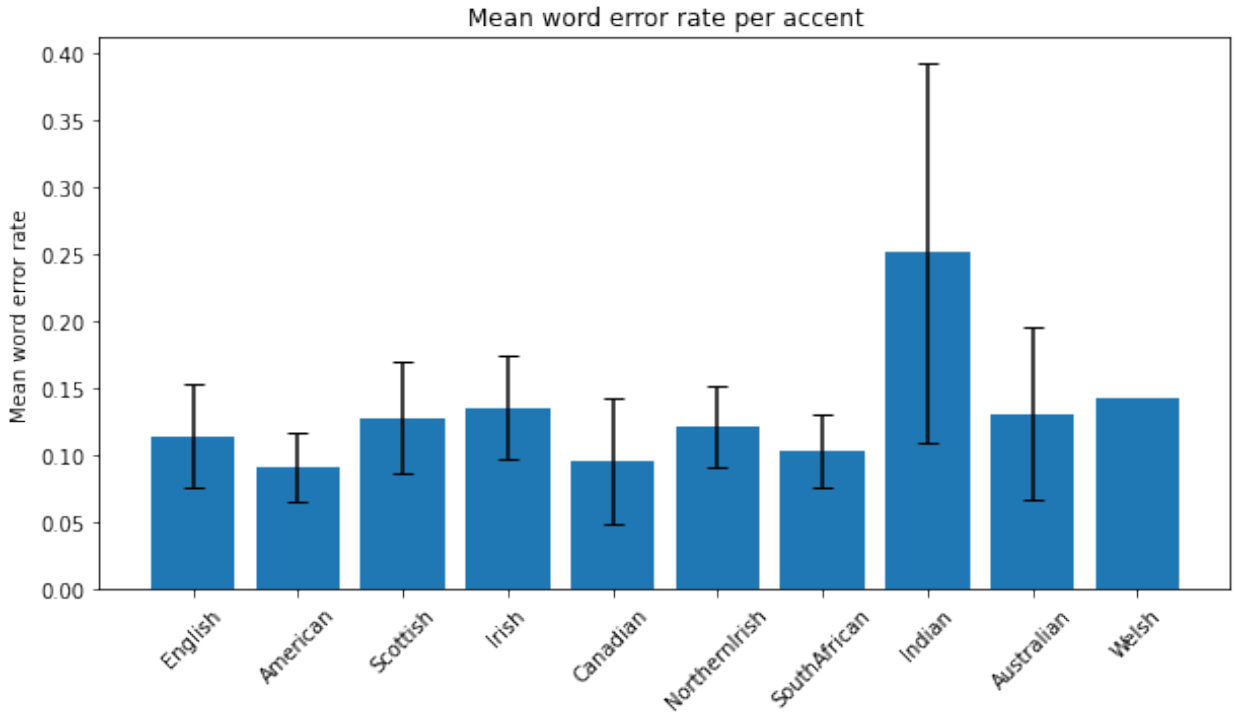

Figure 4.2: Mean error rate per native language in VCTK Corpus

indicates that the artifacts introduced by the algorithms greatly affected the accuracy the ASR models, which are unlikely to have been trained with such noise.

We propose that the ability to preserve information measured by the mean word error rate can be a valid evaluation method for how well style transfer models can disentangle content and style. This evaluation approach would directly tie the definition of content to the linguistic attributes of speech. Especially in the case where the source speaker and target speaker are identical, better style transfer models should expect relative difference in word error rate to approach 0.

One finding that was of particular interest was that the relative increase for well-performing speakers as both source and target is higher than that of a conversion of poor-performing speakers to well-performing speakers. This indicates a baseline level of WER increase caused by the distortions introduced during style transfer. Since the relative difference for poor-performing speakers to well-performing speakers is below this baseline, we hypothesize that lowering WER via our method will become viable as style transfer models become better at synthesizing speech.

Our results indicate that speech style transfer is not a completely solved problem, as even synthesis with the source and target speaker being identical reflected similar performance decreases. These results also indicate a lack of robustness and generalizablilty of ASR systems. Since our style transfer model achieved reasonable results with human interpreters,

this indicates that ASR models are weak against the auditory distortions introduced by style transfer. A potential application of such phenomena can be used in privacy, where users may want to prevent their voice from being processed accurately by speech recognition systems.

Chapter 5

Conclusion

In this report, we introduced audio embeddings as a valid feature for the task of predicting video memorability. However, predicting video memorability remains a difficult task due to the high variance of the data and inherent obstacles in identifying salient features. While image memorability prediction has developed well enough to be an evaluation metric, predicting video memorability as a task remains difficult for models to generalize and perform well on.

Future work would ideally iterate on our findings for much larger and more diverse datasets, as the size of the dataset was notably smaller than similar datasets for the task of predicting media memorability. Consideration should also be given to the variability in dataset elements, as having multiple similar videos may make some videos in similar topics more difficult to recognize distinctly.

We have also highlighted challenges facing audio style transfer models when evaluating them in an automatic speech recognition context. State-of-the-art voice cloning models have achieved decent mean opinion scores, a subjective metric of the overall quality of the output. However, such success does not translate well into preserving linguistic content as measured by the word error rate in speech-to-text systems. Future work in privacy may include extending this phenomena adversarial to create algorithms that preserve content for human listeners but can exploit failure points in ASR models to cloak both the identity and content of the speaker.

One area of inquiry opened by this discussion is the idea of dependent automatic speech recognition and audio style transfer models. A promising direction for future work would be exploring training both acoustic models and audio transfer models with shared content embedding layers.

Finally, once work in memorability and style have matured significantly, we are optimistic about the possibility of augmenting and optimizing digital multimedia content. Given accurate models for predicting memorability as well as style transfer, future work could be able to tweak digital content in a systematic and algorithmic way to maximize memorability and potentially other abstract and artistic attributes.

Bibliography

[1] George Awad et al. *TRECVID 2019: An Evaluation Campaign to Benchmark Video Activity Detection, Video Captioning and Matching, and Video Search and Retrieval*. 2020.

[2] David Azcona et al. "Predicting Media Memorability Using Ensemble Models". In: *Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019*. Vol. 2670. CEUR Workshop Proceedings. CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2670/MediaEval%5C_19%5C_paper%5C_15.pdf.

[3] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 27:1–27:27.

[4] Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: https://www.aclweb.org/anthology/D14-1179.

[5] Romain Cohendet et al. "VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. "A Neural Algorithm of Artistic Style". In: *CoRR* abs/1508.06576 (2015). arXiv: 1508.06576. URL: http://arxiv.org/abs/1508.06576.

[7] Jort F. Gemmeke et al. "Audio Set: An ontology and human-labeled dataset for audio events". In: *Proc. IEEE ICASSP 2017*. New Orleans, LA, 2017.

[8] Eric Grinstein et al. "Audio style transfer". In: *CoRR* abs/1710.11385 (2017). arXiv: 1710.11385. URL: http://arxiv.org/abs/1710.11385.

[9] K. He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[10] Shawn Hershey et al. "CNN Architectures for Large-Scale Audio Classification". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. URL: https://arxiv.org/abs/1609.09430.

[11] Phillip Isola et al. "What makes a photograph memorable?" In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.7 (2014), pp. 1469–1482.

[12] A. Khosla et al. "Understanding and Predicting Image Memorability at a Large Scale". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2390–2398. DOI: 10.1109/ICCV.2015.275.

[13] Dietrich Klakow and Jochen Peters. "Testing the correlation of word error rate and perplexity". In: *Speech Communication* 38.1 (2002), pp. 19–28. ISSN: 0167-6393. DOI: https://doi.org/10.1016/S0167-6393(01)00041-3. URL: https://www.sciencedirect.com/science/article/pii/S0167639301000413.

[14] Allison Koenecke et al. "Racial disparities in automated speech recognition". In: *Proceedings of the National Academy of Sciences* 117.14 (2020), pp. 7684–7689. ISSN: 0027-8424. DOI: 10.1073/pnas.1915768117. eprint: https://www.pnas.org/content/117/14/7684.full.pdf. URL: https://www.pnas.org/content/117/14/7684.

[15] Yist Y. Lin et al. *FragmentVC: Any-to-Any Voice Conversion by End-to-End Extracting and Fusing Fine-Grained Voice Fragments With Attention*. 2020. arXiv: 2010.14150 [eess.AS].

[16] S. Liu and W. Deng. "Very deep convolutional neural network based image classification using small training sample size". In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015, pp. 730–734. DOI: 10.1109/ACPR.2015.7486599.

[17] R. Luo et al. "Discriminability Objective for Training Descriptive Captions". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6964–6974. DOI: 10.1109/CVPR.2018.00728.

[18] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[19] Kaizhi Qian et al. *AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss*. 2019. arXiv: 1905.05879 [eess.AS].

[20] Kacper Radzikowski et al. "Accent modification for speech recognition of non-native speakers using neural style transfer". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2021 (Feb. 2021). DOI: 10.1186/s13636-021-00199-3.

[21] Alison Reboud et al. "Combining textual and visual modeling for predicting media memorability". In: *MediaEval 2019, 10th MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop, 27-29 October 2019, Sophia Antipolis, France*. S, Oct. 2019. URL: http://www.eurecom.fr/publication/6062.

[22] Ricardo Manhães Savii, Samuel Felipe dos Santos, and Jurandy Almeida. "GIBIS at MediaEval 2018: Predicting Media Memorability Task". In: *Working Notes Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29-31 October 2018*. Vol. 2283. CEUR Workshop Proceedings. CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2283/MediaEval%5C_18%5C_paper%5C_40.pdf.

[23] Sumit Shekhar et al. "Show and Recall: Learning What Makes Videos Memorable". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2730–2739.

[24] Aliaksandr Siarohin et al. "Increasing Image Memorability with Neural Style Transfer". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 15.2 (June 2019). ISSN: 1551-6857. DOI: 10.1145/3311781. URL: https://doi.org/10.1145/3311781.

[25] Silero Team. *Silero Models: pre-trained enterprise-grade STT / TTS models and benchmarks*. https://github.com/snakers4/silero-models. 2021.

[26] Steven Weinberger. *Speech Accent Archive*. 2015. URL: https://accent.gmu.edu/.

[27] Junichi Yamagishi, Christophe Veaux, and Kirsten. MacDonald. *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)*. 2019. DOI: https://doi.org/10.7488/ds/2645.