Analyzing 18th-20th Century Art and Music with Contrastive Cross-Modal Learning

Vivien Nguyen
Ren Ng, Ed.
Alexei (Alyosha) Efros, Ed.

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2020-160
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-160.html

August 14, 2020

Analyzing 18th-20th Century Art and Music with Contrastive Cross-Modal Learning

by

Vivien Nguyen

A thesis submitted in partial satisfaction of the

requirements for the degree of

Masters of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ren Ng, Chair
Professor Alyosha Efros

Summer 2020

Analyzing 18th - 20th Century Art and Music with Contrastive Cross-Modal Learning

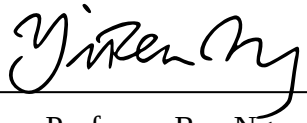
by Vivien Nguyen

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

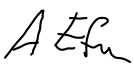


Professor Ren Ng
Research Advisor

August 13th, 2020

(Date)

* * * * * * *



Professor Alyosha Efros
Second Reader

August 10th, 2020

(Date)

Analyzing 18th-20th Century Art and Music with Contrastive Cross-Modal Learning

Abstract

Analyzing 18th-20th Century Art and Music with Contrastive Cross-Modal Learning

by

Vivien Nguyen

Masters of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Ren Ng, Chair

The **relationship between art and music** goes back at least as far as the depiction of instruments and musicians on ancient walls and vases. In more recent centuries, some artists, composers, and theorists have tried to define and explore the relationship between visual and musical concepts more abstractly.

Why should two seemingly completely separate creative domains be related to one another at all? It turns out that even the typical human is generally able to form relationships between different sensory inputs, even when it is not always clear how those relationships are formed or what they are based off of.

In the world of artificial intelligence, this insight has led to a long line of work in exploring **multimodal machine learning**. These works are built on the idea that, for machines to more successfully reason about and navigate the human world, models need to be able to process and interpret multimodal signals.

In this work, we are interested in exploring the relationship between art and music, and more broadly, are motivated by questions of cross-modal perception. We apply techniques from multimodal machine learning to a novel domain, paintings and classical music, in order to learn a shared representation between two different creative modalities. Our results demonstrate that such a representation can be achieved even with limited supervision.

Our embedding space is one that is chronologically organized; works that were created close in time to one another lie close to one another in this embedding space, regardless of their modality (paintings or music).

We hypothesize that future work can improve upon and use such a representation to propose relationships between works from these two domains. Doing so could provide valuable insights about the shared culture two works come from, or about the basis of cross-modal perception.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I must thank my advisor, Professor Ren Ng, for all his advice and support over the past several years. I am so grateful for his encouragement when I first expressed interest in research, and I deeply admire Professor Ng for cultivating such a supportive and engaging group culture. I know I will continue to embrace all the lessons he has taught me about defining problems, developing plans, and communication as I move forward in my academic journey. I could not ask for a more understanding and empathetic advisor.

I am also incredibly grateful to Professor Scott Shenker and Professor Jonathan Ragan-Kelley, who have both given me tremendously valuable advice and have also been so supportive over the past few years.

To Cecilia Zhang, thank you so much for allowing me my first foray into research and for being such an amazing mentor and role model. I don't think I can ever really express how much you have inspired me, and how grateful I am for your patience, advice, and friendship. I pledge my eternal loyalty to you, and I would follow you to the end of the earth.

Of course, thank you as well to all of my amazing office-mates: Ben, Grace, Matt, Pratul, and Utkarsh. Thank you for being so welcoming, for your encouragement, for answering all my stupid questions, for all the boba runs, for the laughs, for teaching me the importance of "breathing through snacks". You all amaze me.

Thank you to Shiry Ginosar for all of her wisdom and advice, not only on this specific project, but on life in general.

To all of my friends who have listened to me cry, complain, and ramble my way through the last 5 years, thank you for your unconditional support, and for keeping me sane.

Finally, and most importantly – my parents. How could I possibly thank you for everything you have done for me? Everything that I am today, I owe to you. Thank you for always encouraging me to chase my passions, and for letting me be me.

Chapter 1

Introduction

Throughout history, people have turned to various forms of creative output such as painting, music, literature and more to express everything from stories to emotions to ideologies. The artwork produced by a particular culture is undoubtedly influenced by socioeconomic circumstances, current politics, religion, philosophical beliefs, and so on. Indeed, these influences may manifest themselves in various ways depending on the output media.

In the visual arts (painting, architecture, etc.), studying the **style** of a particular work, or group of works from a particular period, cultural group, etc. is a common method of understanding those influences (Figure 1.1). Although this methodology has come under attack in more recent decades, it dominated academic discussion in the 19th and 20th centuries and remains a popular method of learning about art particularly for the general public.

(a) *Venus and Adonis*, Paul Rubens, 1635.

(b) *Ecstasy of Saint Teresa*, Gian Lorenzo Bernini, 1647-1652

(c) *Santa Maria della Salute*, Baldassare Longhena, 1631-1687

Figure 1.1: Examples of visual works that are classified under the Baroque art style.

The notion of a "**cross-media artistic style** is one that can be hypothetically exhibited by works of art in more than one medium" [26]. For example, the terms *baroque* or

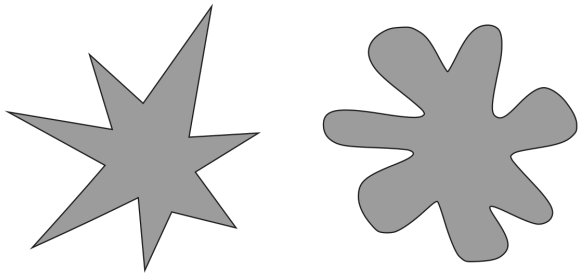


Figure 1.2: Images shown in the Kiki (generally associated with pointed shape on the left) vs. Bouba (generally associated with pointed shape on the right) experiment

impressionist are applied to both painting and music, and is used to name historical periods in both modalities.

But even within the visual arts, there has been debate on whether or not the same style label can or should be applied to works from different media, such as painting and architecture, let alone whether or not correspondences can be found between more distant forms of media, such as painting and music.

That being said, it's been shown that humans are generally able to form relationships between inputs of different modalities, such as visual and audio signals. Sensory integration can lead to synthesized cultural or emotional experiences, [20] [18]. These can become standardized to the point of a feeling of "correctness". Consider, for example, the colloquial notion of a genre: the phrase "pop music" and "rock music" evoke distinct pairings of visuals and sounds; to swap one with the other would induce some cognitive dissonance.

The interaction between visual and auditory elements that creates such experiences occurs at a high level, making use of cultural understanding or emotional cues to associate, for example, a song that "sounds sad" – perhaps a slow song in a lower tone – with an image that "looks sad" – perhaps a rainy scene with dark figures.

However, the interaction between visual and auditory elements can also occur at a "low level". Multisensory illusions such as the McGurk effect [22] or ventriloquism [13] show how integrating these senses creates new *perceptual* experiences. Another example is the "Kiki vs. Bouba" experiment [29]. This effect, first demonstrated by Wolfgang Kohler in 1929 and replicated by Ramachandran and Hubbard in 2001, suggests that "there may be natural constraints on the ways in which sounds are mapped on to objects" that is consistent across people. Subjects are asked which shape (Figure 1.2) corresponds to the name "Kiki", and which corresponds to the name "Bouba". Most people agree that the pointed shape on the left should be paired with "Kiki", while the rounder shape on the right should be paired with "Bouba"

These interactions have a real effect on the way people perceive and understand the world around them.

Whether cross-media styles truly exist due to combinations of underlying, low-level cross-modal perception, only exist due to such high level cultural understanding, or don't exist at all, it remains true that a single culture outputs artifacts in multiple forms of media. The same set of cultural, historical, or artistic developments can be manifested (consciously or not) in different modes of human creative output.

For that reason, there may be unique insights revealed by studying, for example, visual art, music, and literature not as separate and disjoint art forms, but as a related set of creative outputs resulting at least in part from the same set of cultural inputs.

Meanwhile, deep learning and data science have been proven valuable tools in studying and analyzing diverse modes of data. Depending on the capacity of our models or the labels available to us, certain nuances and specifics may be lost in this process. For example, conventional art analysis may rely on in-depth knowledge of culture or biographical information, or depend on physically analyzing the object in question. Nevertheless, there is a different kind of value to being able to analyze mass amounts of data at once. Data-driven methods can be used to look for patterns across data that is both varied and large in scale. These patterns can then be further analyzed in greater depth by specialists with domain knowledge, or be used to provide quantifiable examples of qualitative labels originally provided by human analysts.

In this work, we aim to explore methods for connecting two modes of creative output, painting and music. In particular, our goal is to construct and explore a joint embedding space that would allow us to take a data-driven approach to analyzing and understanding the relationship between art and music.

Though there have been several successful methods for classifying *either* art *or* music according to style, artist, or other features, we aim to explore these aesthetic forms *cross-modally*. For example, we would like to understand the relationship *between* visual features in art and auditory features in music as different styles emerge over time.

In this work we first collect a cross-modal dataset, gathering paintings and classical music from the 18th - 20th centuries. We then use this dataset to explore a method for connecting painting and music in the absence of explicit paired labels or one-to-one correspondences. We examine the results of learning a coordinated embedding space for painting and music, and compare to the results of single-modality learning. Finally, we discuss how such a representation could be improved be useful for understanding the relationship between art and music, and more generally, how that relationship could provide insights into the nature of cross-modal perception.

Chapter 2

Background and Related Work

2.1 Psychological Cross-modal Perception

Cross-modal or multisensory perception describes the ability of humans to combine information from multiple inputs into a single representation. Cross-modal illusions such as the ones described in [27] or the McGurk effect. Show us how the synthesis of input signals results in a unique perceptual experience.

However, many things seem to influence the process by which these various streams are connected. In the literature, these are described (in a non-mutually exclusive manner) as statistical, structural, semantic, or emotional correspondences.

Statistical correspondences result from "internalization of the statistical regularities of the environment". For example, the Kiki-Bouba illusion described in the Introduction is often believed to result from the association of the spoken sounds to the visual appearance of the mouth shape required to produce those sounds. *Structural* correspondences result from the sensory processing systems we are born with to organize various stimuli in order of, for example, increasing magnitude. *Semantic* correspondences, sometimes also called *linguistic* correspondences, refer to stimuli that may result from our use of particular language terms to refer to things from different modalities, i.e. a "bright" sound and a "bright" color [11] [33].

Many psychophysical studies have explored the cross-modal correspondences of "low-level" audio-visual features, i.e. loudness and brightness or loudness and size, that may fall under the *statistical* or *structural* categorizations.

Studies show that participants have consistent cross-modal associations from intervals and chords to colors as well as instrument timbres to colors [34] [37] [11]. Another line of work explores how these associations may be semantic correspondences, developing the "Emotional Mediation Hypothesis" [28]. According to this hypothesis, the relationship between sounds and visuals is linked through the emotion. For example, a person may associate a particular sound with a particular color because they associate them both with the emotion "happy".

2.2 The Relationship Between Art and Music

Both "low-level" features and "high-level" semantic concepts have a place in categorizing and analyzing art and music. For example, in the field of art history, pieces of art may be discussed in terms of aesthetic formalism, which is concerned with elements such as color, line, shape or texture, or in terms of its iconography AKA how it depicts its content or a particular meaning.

We might therefore expect a relationship between visual art and music that could also be described by a combination of statistical and semantic correspondences.

For example, cultural context or emotions could create a semantic correspondence between visual images and music. Modern music videos are a strong example of this (though this is by no means a modern phenomenon), where the combination of visuals, music, and language create a powerful narrative and/or emotional experience.

Similar to the "Emotional Mediation Hypothesis" explored by Palmer in [18], several studies explore the emotional connection between music and paintings specifically, as well as the effect of music and paintings on an observer's mood. [24] [18]

Less often explored is how statistical correspondences may account for a perceived relationship between a certain piece of music and an image. For example, the notion of a "cross-media artistic style is one that can be hypothetically exhibited by works of art in more than one medium". Statistical correspondences could be the basis of a cross-media artistic style. When describing a particular painting as *baroque*, for example, the analyst may point to not necessarily a particular color, but the use of color combinations overall; or to the use of a particular type of linework or textural detail, and the visual effect or impact created by these things. Similarly, when categorizing a particular musical composition as *baroque*, the analyst may describe the musical "texture" of the piece or use of tones.

However, it's unclear how (or if) these features correspond to one another across modalities. Indeed, there is some skepticism on the existence of cross-media styles at all for example by Merriman in [23] (summarized here by [26]):

> the evidence generally given for the existence of cross-media styles is based on "improper sampling, metaphorical transfer of terms, arbitrary conversion, sheer subjectivity and more or less free associationalism." Improper "pick and choose" sampling is rampant among crossmedia theorists: Parallels between the arts are "demonstrated" by choosing the two most similar works out of many thousands. An example of metaphorical transfer of terms would be calling both the sonnets and the sculptures of Michelangelo "jagged." Free-associationalism refers to the tendency to equate works in different media because they were created contemporaneously.

While these criticisms are valid, [26] demonstrates that naive (untrained) participants could often correctly group works of music, poetry, and art that were made in the same named style period. This suggests that these "associations" or "transfers" are consistent across people. Indeed, the long line of work in psychophysical cross-modal perception (discussed in

Section 2.1 suggests that what Merriman calls the "metaphorical transfer of terms" actually has a consistent perceptual effect, i.e. semantic correspondence.

Similarly, [1] explores the existence of cross-modal associations between "highly complex stimuli", namely between materic painting and classical Spanish guitar. This study asks subjects to match paintings and music clips to adjectives describing the stimuli, but also to match paintings and music clips directly to one another.

2.3 Cross-modal Learning

There continues to be rapid growth in the amount and types of multimodal data (image, video, audio, text, etc.) available. Cross-modal learning aims to combine and connect information across these different modalities, motivated by the fact that humans often complete complex tasks by making use of multisensory information.

Cross-modal methods include, for example: cross-modal representations, where "heterogeneous data" is projected into a common subspace organized by semantics rather than data modality [12]; cross-modal generation, where data from one modality is generated from another; or cross-modal learning as a method for unsupervised/weakly supervised learning.

Cross-modal applications commonly work with relating images and sketches, images and text, text and speech audio, video and speech, and so on. In particular, the relationship between images and music audio is not often explored. This is likely because there are not many examples of paired or grouped images and music, which itself may be because of limited interest in the relationship between generic images and music.

Prior work that *does* focus on the cross-modal relationship between image data and music is typically related to music videos and cross-modal retrieval, which is useful in the music industry for music similarity and recommendation algorithms [25]. It is also used as a creative tool, for example to add semantically meaningful music in the background of a video or photo slideshow [3]. In each of these cases, the model is typically able to learn from given pairs of music and visuals due to the availability of paired data from existing music videos or music album covers [2] [39]. This paired data (and the relationships learned from it) is also directly meaningful for these end tasks. In this work, we try to learn a more general relationship between the two modalities and without paired data for the specific domains we are interested in.

2.4 Computational Art and Music Analysis

Prior work in computational art and music analysis have used a wide variety of techniques.

Studying the influences and development of various artists or composers is a common question in art and music analysis. [8] is one such work that constructs "ecological" networks based off documented relationships and influences between composers. [30], on the other

hand, uses a Bag of Words approach to encode a painting as a set of local, semantic-level features.

Other works have demonstrated success in classifying paintings by style or classifying objects in paintings using CNNs. [7][9] Similarly, CNNs have successfully been used to classify music into genres or by emotion. [19][5]

There are of course other works that use domain knowledge to hand-pick features for analysis. For example, [6] is a cross-modal work that attempts to explore the connection between French painting and music, and Russian painting and music from 1870 - 1920. They do this by measuring and comparing visual value (degree of brightness) to auditory pitch. However, due to the limited number of features studied, it can be difficult to find a specific correlation. Hand picked features are challenging to explore exhaustively, and the relationship across domains may only be explained by several features in combination.

While our goals are similar to those presented in [6], we aim to approach the problem of finding a cross-modal relationship between music and art using learning-based techniques discussed in Section 2.3.

Chapter 3

Dataset

One contribution of this work is the collection and use of a cross-modal painting and classical music dataset.

First, we define the scope of our dataset. We are interested in the domains of Western painting and art music (often colloquially referred to as "classical music") from roughly the 18th - 20th centuries. The time frame is chosen based on the greater availability of data and information on works from this period.

Next, it is important to consider the metadata available or desired in each modality as this determines the possible ways of connecting the two modalities. For example, since we are interested in how different features emerge jointly in each modality over time, we want to collect the year that each piece of art or music was created.

Another label that might be useful is the style classification of the art or music (Baroque, Classical, etc.). While we do collect this information, we ultimately chose not to use it as a supervisory training signal. Firstly, since we want to discover how different styles emerge over time, we want to avoid the bias from directly using these pre-existing subjective labels. Secondly, the granularity of style labels easily available differs greatly between art and music.

Another aspect to consider, particularly when it comes to music, is how the data itself will be represented. There are two main options for audio data: audio recordings of real performances, or a symbolic representation of music such as MIDI. MIDI files explicitly encode notes and other musical features such as tempo and volume. We ultimately decided to use audio recordings of real performances as a more "real" representation of sound.

With all these aspects in mind, we want to collect a dataset that is at least roughly balanced across time (i.e. style periods) and across modalities. That is, we want to ensure there is sufficient data across the 200 year span we are interested in, and we want to ensure that there is comparable amounts of data in both music and art modalities for each time period.

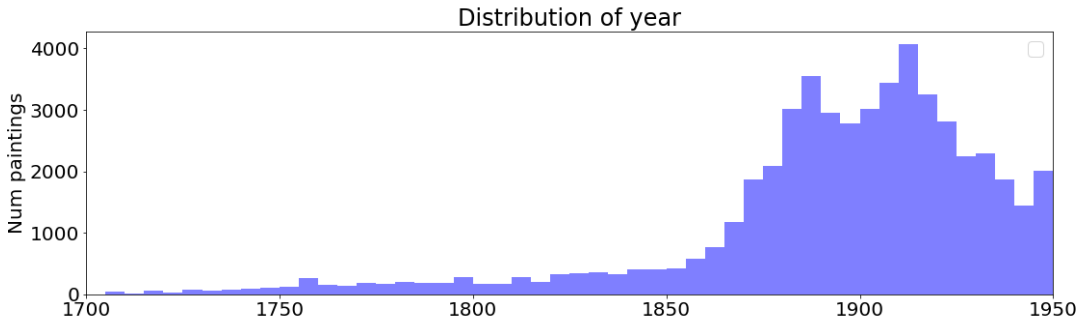


Figure 3.1: Distribution of data across time of WikiArt dataset

3.1 Data Collection

WikiPaintings

WikiPaintings[17] is a set of metadata collected from WikiArt.org. Since it already contains information such as the title, artist, year, and style, we can immediately use this as our metadata and use the included URLs to download the artwork.

Despite its name, the WikiPaintings metadata still contains works that are not paintings, such as images of sculptures or conceptual art. We filter those out, along with works from non-Western artistic traditions, such as traditional Chinese painting.

Figure 3.1 visualizes the distribution of resulting paintings across time. Notably, the dataset contains significantly more paintings from after 1850 than before 1850.

MAESTRO Dataset

The MAESTRO dataset is a collection of 200 hours of piano performance recordings of classical music from the International Piano-e-Competition [14]. This dataset is a great place to start, as the recordings are very high quality.

The metadata associated with this dataset does not include the year the piece was composed, nor the style of the piece. We use the International Music Score Library Project, or IMSLP, database to find that information [16]. Several pieces only have an approximate date or date range. Many pieces also do not have the style directly associated with it, so we use the general style period associated with the composer, which we also collect from IMSLP.

Another issue with the MAESTRO dataset is that it is a relatively small dataset and is not well-distributed throughout the time frame we're working in.

Figure 3.2 visualizes the distribution of pieces in the MAESTRO dataset across time, using the style period associated with the composer as a proxy for the composition's style to color the histogram.

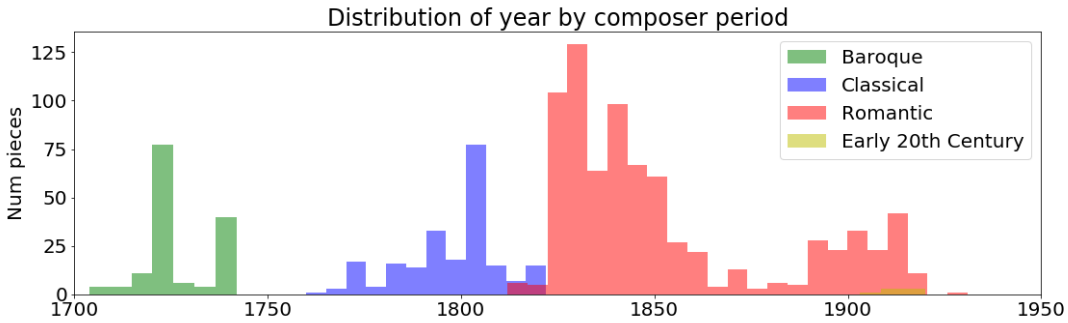


Figure 3.2: Distribution of data across time of original MAESTRO dataset

Yale-Classical Archives Corpus and YouTube

To supplement the MAESTRO dataset, we utilize the metadata from the Yale-Classical Archives Corpus (YCAC), which contains information such as piece title, composer, and year composed [38]. To obtain the style, we continue using the general style period associated with the composer.

We then use the metadata to search and download audio from YouTube. Though we don't employ any rigorous data cleaning, in practice, we notice that the top search result is typically a high quality album or live performance recording of the correct piece.

Aggregating the data

With these various data sources in mind, we can now aggregate our final, cross-modal dataset. First, we must establish the time frame for which we have both painting and music data. It is difficult to find recordings of classical music composed before 1700, and after 1940. Thus, though the WikiArt dataset is extensive, we will only consider paintings made between 1700 and 1940 to mirror the availability of music data.

We further resample the dataset in order to combat the imbalance between paintings created before and after 1850. The resulting distribution is shown in Figure 3.3.

We use the YCAC metadata to supplement the MAESTRO dataset by adding pieces to increase the overall amount of data, and also balance time periods that were rarer in the original MAESTRO dataset. To avoid variable length inputs, we split each audio file into 30-second clips, which yields the final distribution shown in Figure 3.4.

While there are still some disparities between the distribution of music clips and paintings, aggregating the dataset in this way alleviates the sparsity of data from some time periods while maintaining aspects of the original data distribution.

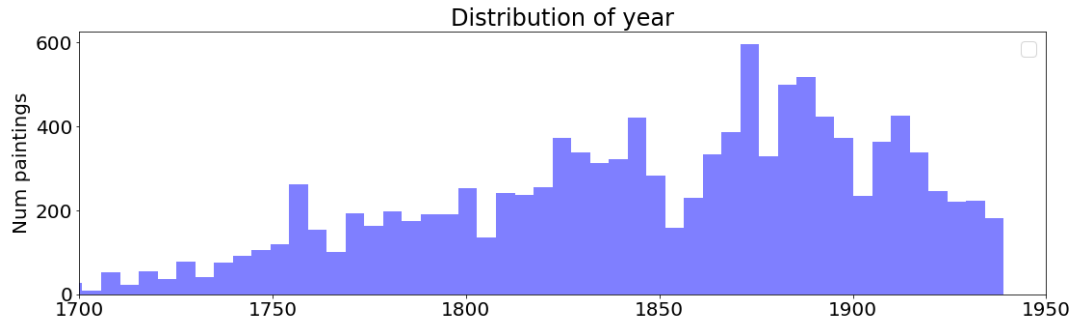


Figure 3.3: Distribution of data across time of final painting dataset

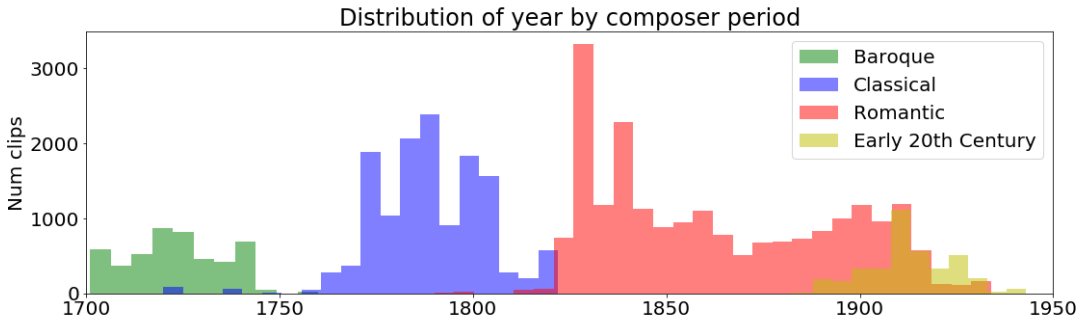


Figure 3.4: Distribution of data across time of aggregated music dataset. Note that the y-axis now represents number of 30-second clips.

3.2 Data Pre-processing

This section explains the various ways our data is pre-processed for input into the neural network architectures we use, which will be further are described in Chapter 4.

Splitting the Dataset

We divide our dataset into train, validation, and test splits following an 80-10-10 ratio for music and art separately. This gives the dataset splits listed in Table 3.1.

Art Pre-processing

As originally described in [32] and implemented in [21], to obtain a fixed-size 224×224 image, training images are re-scaled and randomly cropped. Other data augmentations,

	Train	Validation	Test	Total
Art (paintings)	6291	1161	1162	8714
Music (clips)	34888	4161	4162	43211

Table 3.1: Number of paintings for each split in art dataset and number of clips for each split in music dataset.

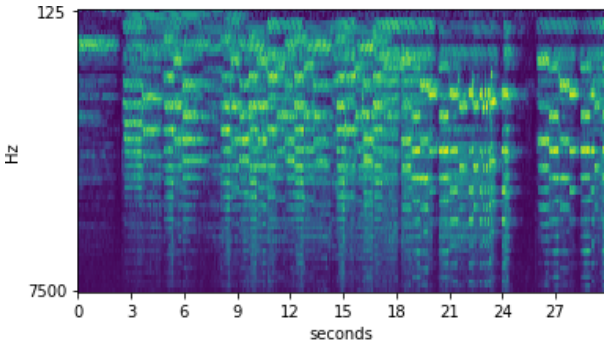
such as flipping the image horizontally, applying a small rotation, and subtracting a mean RGB value, are also applied.

Music Pre-processing

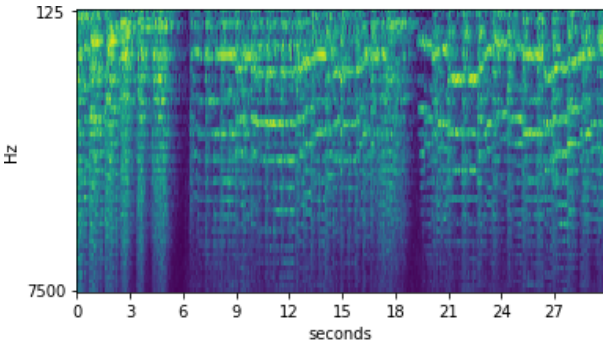
To preprocess data for MusicNet, starting from a 30 second audio clip, we first convert the audio into a log-mel spectrogram. These spectrograms are computed by first dividing the 30 second audio clip into 960 ms frames, then applying a short-time Fourier transform using 25 ms windows every 10 ms and 64 frequency bands, as in the original work. This means a 30 second audio clip yields 31 96 $\times$ 64 log-mel spectrogram patches. The spectrogram can be visualized by "stitching" together each of these patches, as seen in Figure 3.5.

The x-axis of the spectrogram is time (in seconds), while the y-axis of the spectrogram represents the frequency bands used bin the spectrogram. VGGish specifies that there are 64 bands in the range 125 Hz to 7500 Hz.

It can be challenging to mentally "translate" between a spectrogram and the original sound/audio, but some do have clearly visible melodic patterns and chords.



(a) L'Histoire du Soldat 3 Music to Scene II Undeterminable, Stravinsky, 1918



(b) Concerto for 2 Clarinets and Orchestra in Bb Edition for 2 Clarinets and Piano 2 Andante Moderato F major, Stamitz, 1765

Figure 3.5: Spectrograms allow us to create 2D visualizations of audio signals.

Chapter 4

Learning Cross-Modal Relationships

4.1 Goals: Learning a Chronological Embedding

Our goal is to learn about the relationship between music and paintings created during the same time periods. We want to avoid trying to use a style label as supervision, for reasons discussed in Chapter 3. We also don't necessarily want to directly regress to the year the music or painting was created, partly because our date labels are noisy and imperfect. We also want to be able to encapsulate some notion of similarity between works, both within and across modalities.

Constructing triples

For those reasons, we take a metric learning approach. This allows us to model the notion of relative similarity or dissimilarity between two pieces of data. It also fits the roughly continuous way that style changes over time.

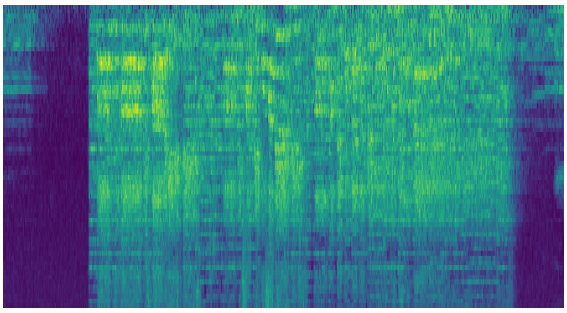
Another design aspect of our network is how to construct triples consisting of an anchor, positive example and negative example. By definition, the anchor and positive example are meant to be semantically closer to one another or "more similar", while the anchor and negative example are meant to be "less similar".

We chose to construct triplets based on the hypothesis that, stylistically, art and music created in the same time period are more similar to one another than art and music from different time periods. We defined a **positive pair** of works to be any two created within 10 years of one another, while a **negative pair** of works is any two created at least 20 years apart from one another.

By allowing positive pairs to be created within a span of years, we account for slightly mis-dated works as well as the fact that styles and techniques change gradually over time.

We combine subnetworks in order to process data from different modalities.

Sub-networks that take the inputs from the same data modality (art or music) will share weights with one another, also known as "Siamese networks" [10]. At any given training step, the same data input, whether it serves as the "positive" example in one triple or the



(a) Messa da Requiem a tre voci d'uomo 4 Dies irae D minor, Perosi, 1892

(b) Landscape at Midday, Cezanne, 1887

(c) Mademoiselle de Clermont as a Sultana, Nattier, 1733

Figure 4.1: Example of network inputs for a (music, art, art) triple

"negative" example in another triple, should be transformed into the same embedding by the sub-network.

4.2 Network Architecture

Base networks

We need two types of sub-networks, one for each modality, which we will refer to as ArtNet and MusicNet.

One of the challenges of deep learning with art or music is the limited amount of data. Typical practice is to overcome this by finetuning pre-trained network. Thus, for both ArtNet and MusicNet, we use a pre-trained state-of-the-art network as the base of the network that will be finetuned according to our objective. On top of these pre-trained networks, we add a few fully connected layers that will be trained from scratch and allow us to flexibly control the final embedding size.

ArtNet

ArtNet uses VGG-16 [32] as its pre-trained base and three fully-connected layers with output sizes 512, 256, and 16. VGG-16 has been used in prior work to classify artwork with demonstrated success ([7]). To confirm this, we trained ArtNet by adding an additional classification layer to predict the painting's style label as a sanity-check task.

MusicNet

MusicNet uses VGGish [15] as its pre-trained base and three fully-connected layers with output sizes 1024, 256, and 16. As with ArtNet, we validated the reasonableness of VGGish

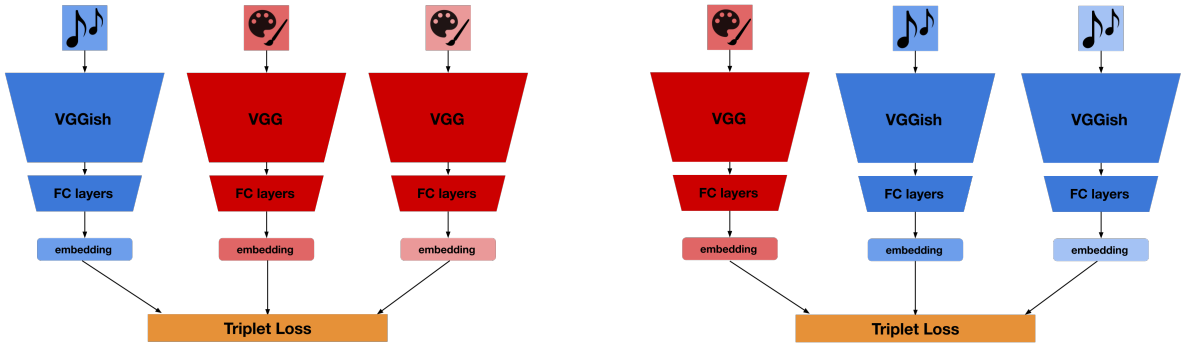
features by adding a classification layer to predict each music piece's style label.

VGGish processes each of these 31 patches independently; they are combined back together before entering the additional fully connected layers.

Triplet Network

We chose to use a Triplet Network architecture, which can be seen as three sub-networks representing an anchor, positive example, and negative example. The Triplet Network then consists of three subnetworks, which can be any combination of ArtNets and/or MusicNets.

We construct **cross-modal triplets** for our main experiment, where the anchor comes from one modality (art or music) and both the positive and negative examples come from the other modality. This yields the two possible Triplet Network configurations shown in Figure 4.2.



(a) Network architecture for (music, art, art) triples.

(b) Network architecture for (art, music, music) triples.

Figure 4.2: Network architectures for cross-modal triplet networks.

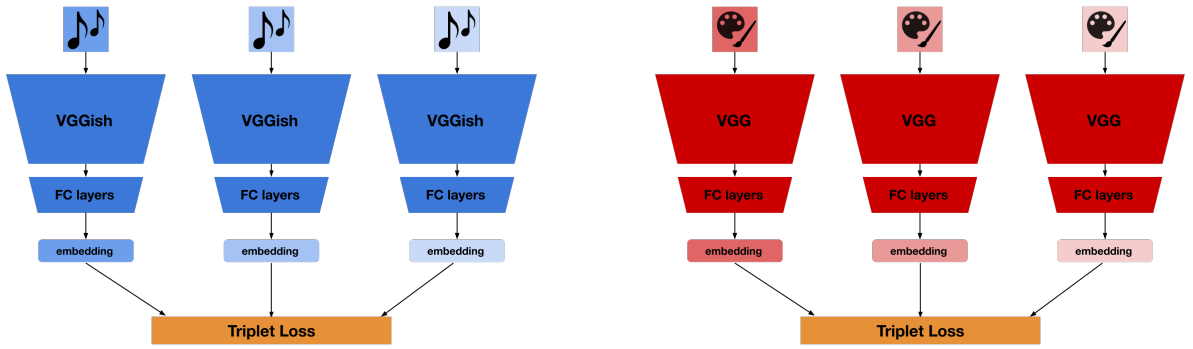
For comparisons, we also train networks based on single modality triples, where the anchor, positive example, and negative example all come from the same modality, which yields the two Triplet Network configurations shown in Figure 4.3.

Due to the Triplet Network's modular construction, any particular painting or music input will be embedded independently by the appropriate subnetwork. However, during crossmodal training, both subnetworks jointly update their weights since the triplet loss takes all three embeddings as input.

The Triplet Loss

To supervise the network, we use the triplet loss:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} max\{\|emb_i^{anc} - emb_i^{pos}\|_2^2 - \|emb_i^{anc} - emb_i^{neg}\|_2^2 + \alpha, 0\}$$



(a) Network architecture for (music, music, music) triples.

(b) Network architecture for (art, art, art) triples.

Figure 4.3: Network architectures for single-modality triplet networks.

where $emb_i^{anc}, emb_i^{pos}, emb_i^{neg}$ are the resulting embeddings from the anchor input, positive input, and negative input respectively for the ith triplet in a batch size of N. This has the effect of encouraging the network to embed the anchor input and positive input closer to one another than the anchor input and negative input.

The parameter α is frequently referred to as the **margin**, and specifies *how much* closer the anchor and positive embeddings should be to one another compared to the anchor and negative embeddings. Thus, a loss value greater than the margin α means the network actually embedded the anchor and positive example *further* away from one another than the anchor and negative example. A non-zero loss less than α means the network successfully embedded the anchor and positive example closer to one another than the anchor and negative example, but not with the provided margin. Zero loss means the network embedded the anchor and positive example at least α closer together than the anchor and negative example.

4.3 Training Details

For all the experiments listed in this chapter, we train our Triplet Network using a **batch size** of 4, the **Adam optimizer**, an initial **learning rate** of 1^{-5}.

Output Embeddings

We experiment with two formats for the cross-modal output embeddings. For both, we fix the output **embedding size** to 128. For the first, we additionally constrain the embedding to have unit ℓ_2-norm. Prior work have experienced empirical success with these parameters

	train loss	val loss	test loss
unconstrained cross-modal (music-art-art)	5.589	9.554	8.794
unconstrained cross-modal (art-music-music)	6.363	5.923	5.757
constrained cross-modal (music-art-art)	0.533	0.549	0.572
constrained cross-modal (art-music-music)	0.540	0.546	0.536
single-modality art	0.155	0.357	0.493
single-modality music	0.463	0.512	0.345

Table 4.1: Final loss values for each model (noisy due to the nature of triplet training).

for representation learning and discussed the benefits of having features lie on the unit hypersphere [36] [35]. In this case, we use $\alpha = 0.8$ in the triplet loss (the **margin**).

The second format has no constraints on the output embedding, and uses an $\alpha = 10$ in the triplet loss. (Note that this is the margin parameter of the loss function, and unrelated to the ± 10 year period from which "positive" examples are mined from to generate input triplets).

We will refer to these two embedding spaces as the **constrained** and **unconstrained** cross-modal embeddings, respectively.

For our single modality triplet networks, we only trained **constrained** embeddings.

Cross-modal Training Procedure

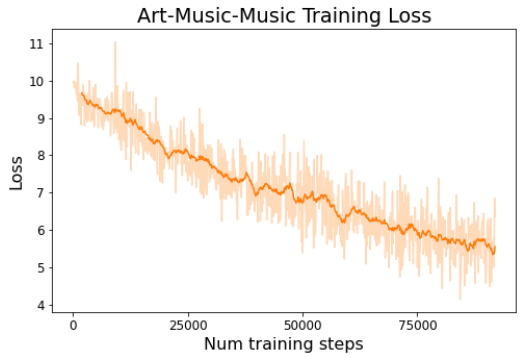
Our training process involves alternating between two configurations of the Triplet Network. We alternate between triplets of the form **art, music, music** and **music, art, art** for the anchor modality, positive example modality, and negative example modality, respectively. Initial results suggested that in practice, this improves the network's ability to learn a well-formed embedding space for *both* modalities.

Since the art modality and music modality have different amounts of training data, we set one "epoch" to be the number of iterations it takes to go through the smaller modality (art) – approximately 2300 iterations.

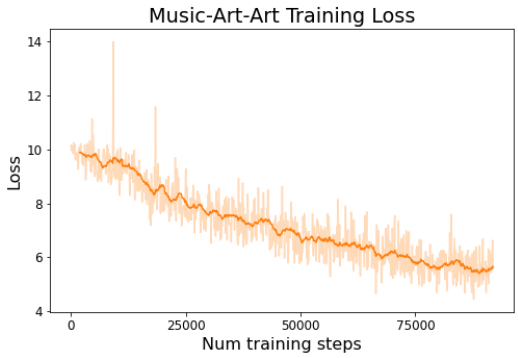
4.4 Training Quantitative Results

Aside from standard loss curves, we compute additional quantitative metrics, described in the following sections. All metrics are summarized in Tables 4.1, 4.2, and 4.3. Recall that the loss value is dependent on the margin, which is set differently for our constrained and unconstrained embedding models.

Training loss curves from training the unconstrained and constrained cross-modal embeddings can be seen in Figures 4.4 and 4.5. Training loss curves from training constrained



(a) Training loss values from only the (art, music, music) triples.



(b) Training loss values from only the (music, art, art) triples.

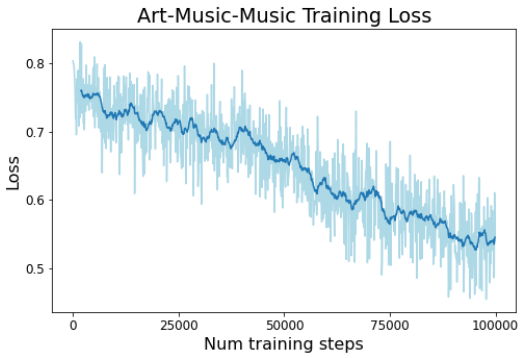
(c) Training loss curve, interleaving values from the alternating training scheme.

Figure 4.4: Training curves for the cross-modal, unconstrained embeddings model.

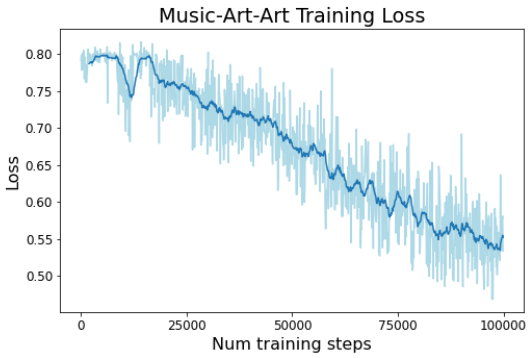
single modality art embeddings can be seen in Figure 4.6, and training loss curves from constrained single modality music embeddings can be seen in Figure 4.7.

After every epoch (as defined in the previous subsection), we validate our network's progress by holding the weights constant and embedding the entire validation dataset. This is done by allowing each example from the validation dataset to serve as the anchor once (rather than random sampling). However, to collect a meaningful loss value, we still require that the positive example and negative example are sampled correctly.

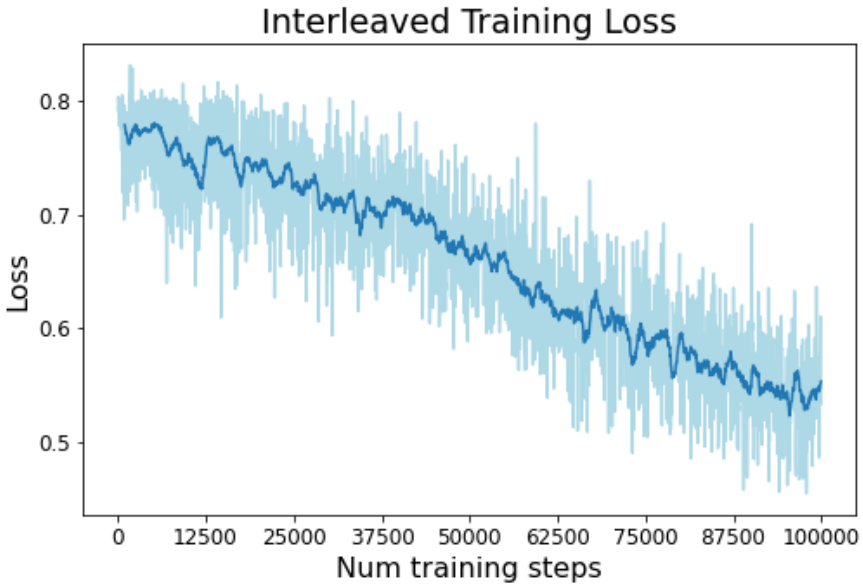
Validation loss curves from training the unconstrained and constrained cross-modal em-



(a) Training loss values from only the (art, music, music) triples.



(b) Training loss values from only the (music, art, art) triples.



(c) Training loss curve, interleaving values from the alternating training scheme.

Figure 4.5: Training curves from the cross-modal, constrained embeddings model.

beddings can be seen in Figures 4.8 and 4.9. Validation loss curves from training constrained single modality art embeddings can be seen in Figure 4.10, and validation loss curves from constrained single modality music embeddings can be seen in Figure 4.11.

Additional metrics

Aside from validation loss, we compute two other metrics, Precision @ K and Average Difference @ K. Both of these first require us to embed the entirety of our validation set

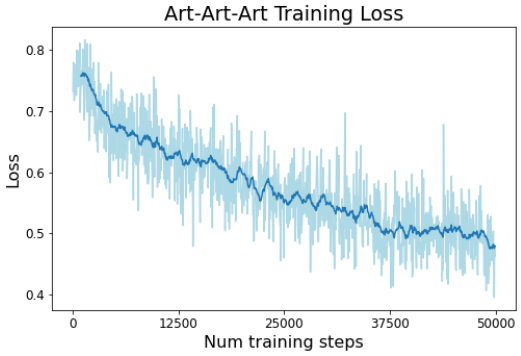


Figure 4.6: Training loss curve for single-modality art embeddings.

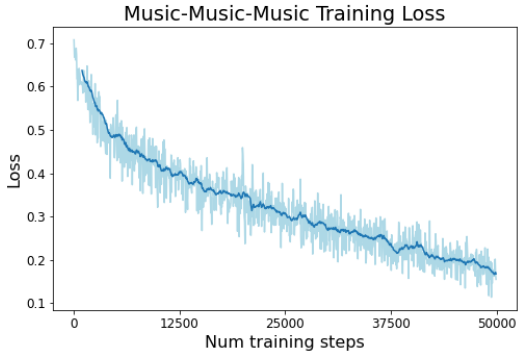


Figure 4.7: Training loss curve for single-modality music embeddings.

	art-art	art-mus	mus-art	mus-mus
unconstrained cross-modal (validation)	0.201	0.208	0.204	0.305
unconstrained cross-modal (test)	0.202	0.216	0.191	0.300
constrained cross-modal (validation)	0.242	0.191	0.196	0.374
constrained cross-modal (test)	0.172	0.195	0.202	0.292
single-modality art (validation)	0.242	N/A	N/A	N/A
single-modality art (test)	0.242	N/A	N/A	N/A
single-modality music (validation)	N/A	N/A	N/A	0.374
single-modality music (test)	N/A	N/A	N/A	0.376

Table 4.2: Precision summary

(for all modalities involved). Then, we designate a **source** modality and **search** modality (which can be either the same modality or the other modality). Each metric is therefore computed 4 times: art to art, art to music, music to music, and music to art.

Both metrics involve finding the K nearest neighbors in the search modality for a given embedding from the source modality. For each embedding in the search modality, we compute the L_2-distance to every embedding in the source modality, then take the closest K embeddings.

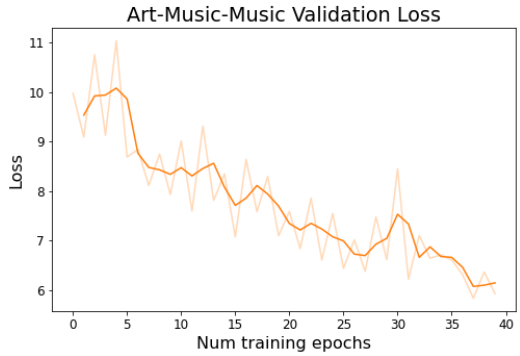
Precision @ K

Our network is trained under the hypothesis that music and art created within 10 years of one another are similar. It is encouraged to place such pairs close to one another in the embedding space (by construction of the input triplets).

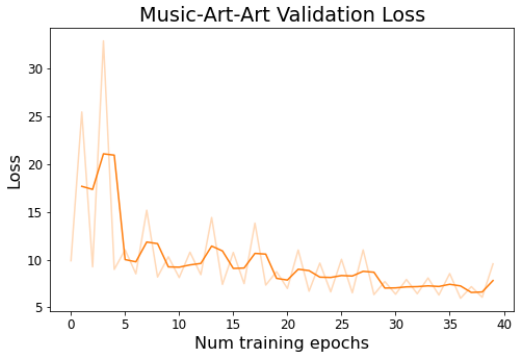
For a particular embedded work of music or art (the source), this metric measures the

	art-art	art-mus	mus-art	mus-mus
unconstrained cross-modal (validation)	39.05	35.41	41.36	34.62
unconstrained cross-modal (test)	39.44	35.44	41.79	35.04
constrained cross-modal (validation)	35.78	37.94	40.84	27.96
constrained cross-modal (test)	45.27	38.52	40.70	35.59
single-modality art (validation)	36.22	N/A	N/A	N/A
single-modality art (test)	35.85	N/A	N/A	N/A
single-modality music (validation)	N/A	N/A	N/A	27.96
single-modality music (test)	N/A	N/A	N/A	27.39

Table 4.3: Average Difference summary



(a) Validation loss values over the course of training, using art as the anchor.



(b) Validation loss values over the course of training, using music as the anchor.

Figure 4.8: Validation loss values for the unconstrained embeddings. These are computed using both the (art, music, music) and (music, art, art) configurations after every training epoch.

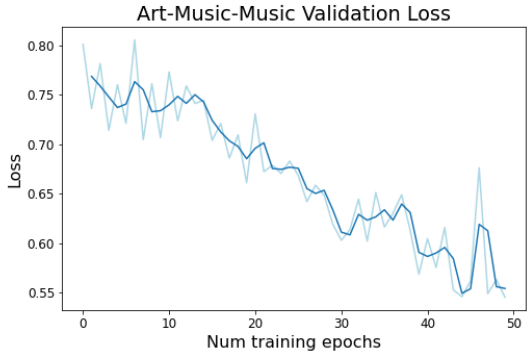
percentage of K nearest-neighbors that were created within 10 years of the source. Precision @ K is a metric frequently used by content retrieval and recommender systems, such as [39].

To compute this metric, for each embedding of the validation data, we find the K nearest neighbors. We then count the number of embeddings n in the closest K embeddings that were created within 10 years of the source, and compute the average over all possible source embeddings.

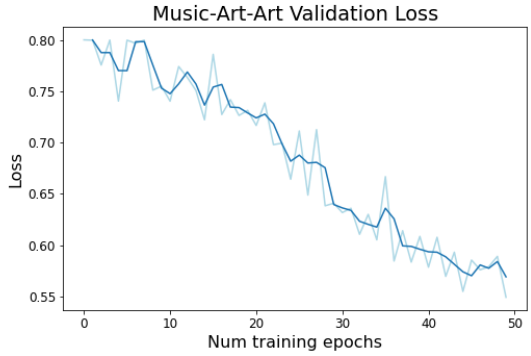
Precisely, we define

$$P@K = \frac{1}{N}\sum_{i=1}^{N}\frac{n_i}{K}$$

where N is the total number of source embeddings, and n_i is the number of embeddings



(a) Validation loss values over the course of training, using art as the anchor.



(b) Validation loss values over the course of training, using music as the anchor.

Figure 4.9: Validation loss values for the constrained embeddings. These are computed using both the (art, music, music) and (music, art, art) configurations after every training epoch.

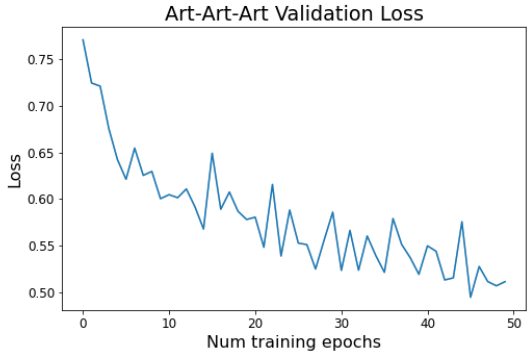


Figure 4.10: Validation loss curve for single-modality art embeddings.

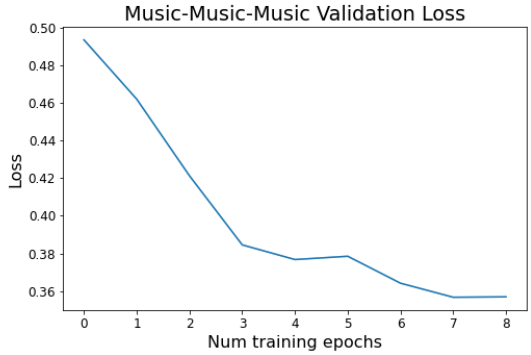


Figure 4.11: Validation loss curve for single-modality music embeddings.

within the ith source embedding's K nearest neighbors that has a creation date within 10 years of the ith embedding's creation. A higher P@K value is better.

Note that since K is a fixed constant, even a perfect embedding may yield a value less than 1, if there are not at least K other works created within 10 years of the source.

Figure 4.12 visualizes the Precision @ K=25 across training epochs for the unconstrained embeddings, and Figure 4.13 visualizes the same metric for the constrained embeddings. Figures 4.14 and 4.15 visualize Precision @ K=25 for training the constrained art and music single-modality embeddings.

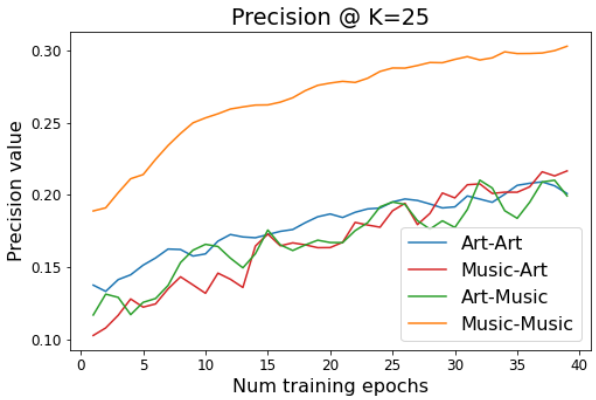


Figure 4.12: Precision metric over the course of training for unconstrained embeddings.

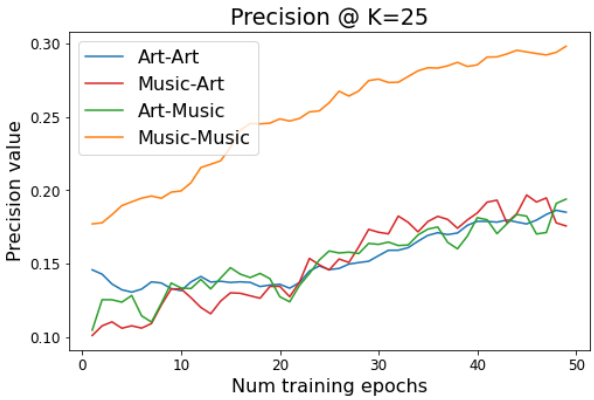


Figure 4.13: Precision metric over the course of training for constrained embeddings.

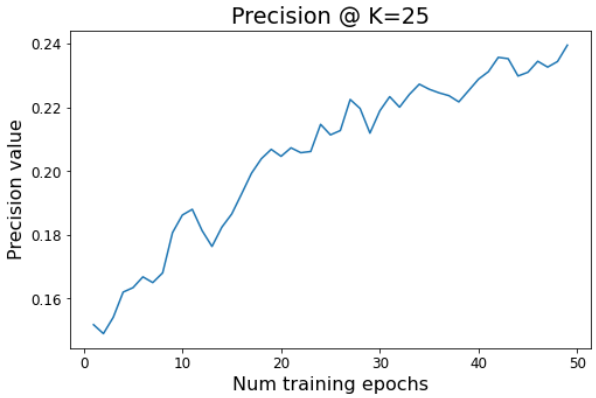


Figure 4.14: Precision metric over the course of training for single-modality art embeddings.

Figure 4.15: Precision metric over the course of training for single-modality music embeddings.

Average Difference @ K

A similar metric is to compute the average (absolute) difference in years of creation between the source embedding and each of its K nearest-neighbors, AD@K.

This allows us to see how localized to a particular time period a neighborhood of embeddings is. A lower AD@K value is better.

Figure 4.16 visualizes the Average Difference @ K=25 across training epochs for the unconstrained embeddings, and Figure 4.17 visualizes the same metric for the constrained embeddings. Figures 4.18 and 4.19 visualize Average Difference @ K=25 for training the constrained art and music single-modality embeddings.

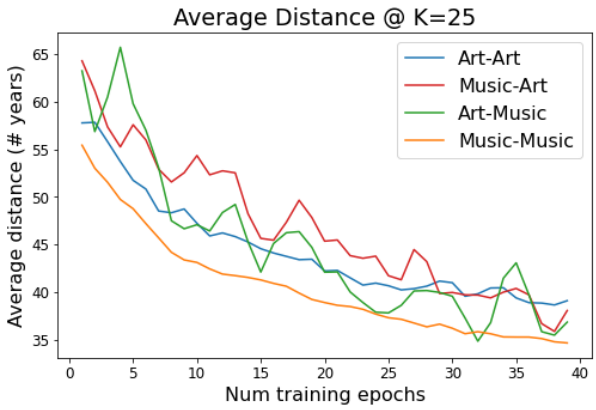


Figure 4.16: Average difference metric over the course of training unconstrained embeddings.

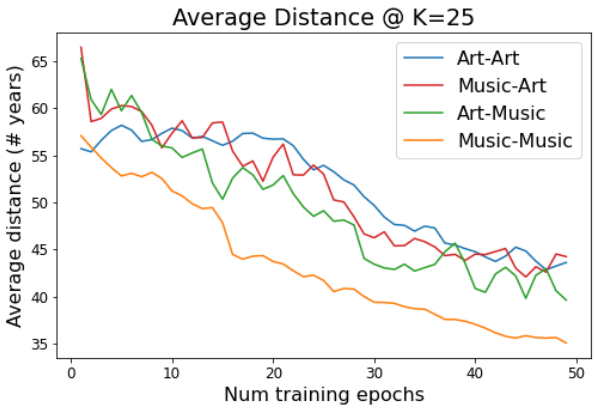


Figure 4.17: Average difference metric over the course of training constrained embeddings.

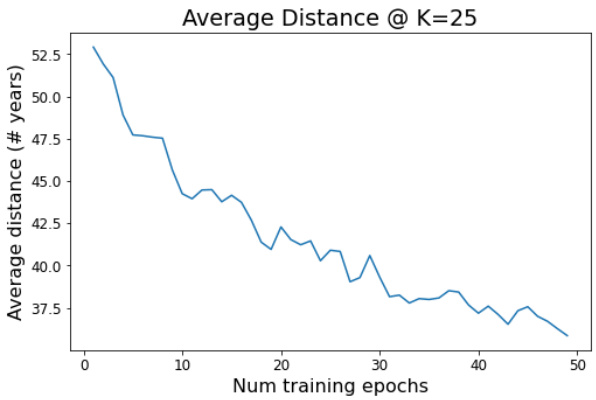


Figure 4.18: Average difference metric over the course of training for single-modality art embeddings.

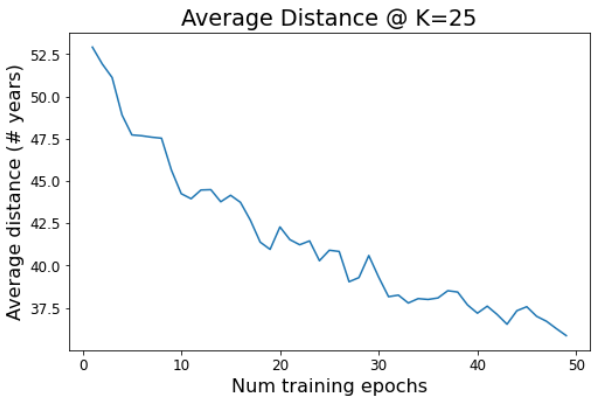


Figure 4.19: Average difference metric over the course of training for single-modality music embeddings.

Chapter 5

Analysis and Discussion

5.1 Analysis

First, we observe that for all experiments, the training and validation loss curves and the final test loss values are reasonable for a model that is learning to complete its task without overfitting to the training set.

Comparing the cross-modal networks to the single-modality networks, we observe that both single-modality networks seem to converge to better loss values than the cross-modality network. In particular, the network trained only on music converges to a significantly lower loss value.

This was not expected, partly since the art dataset was of more consistent quality. Moreover, the effectiveness of VGGish as a general audio feature extractor, or the adaptability of VGGish features to different domains has not been demonstrated to the same extent as VGG. However, the music training dataset we used was of higher *quantity* than the art dataset.

Similarly, within the cross-modal network, the precision and average difference metrics are best when using the music modality as both the source and search space. This suggests that within the cross-modal embedding space, the music portions are "better organized" compared to the art portions, or both put together.

5.2 Visualizing Our Chronological Embedding Space

To begin inspecting our learned embedding space, we can reduce the dimensionality of our embeddings using principal components analysis (PCA), which allows us to visually inspect which embeddings are nearby one another.

In Figure 5.1, we plot the results of reducing our constrained, 128-D embeddings into 2 components. The points are colored according to the year of creation, where colors close to white represent works made close to 1700, while colors close to red or blue represent works

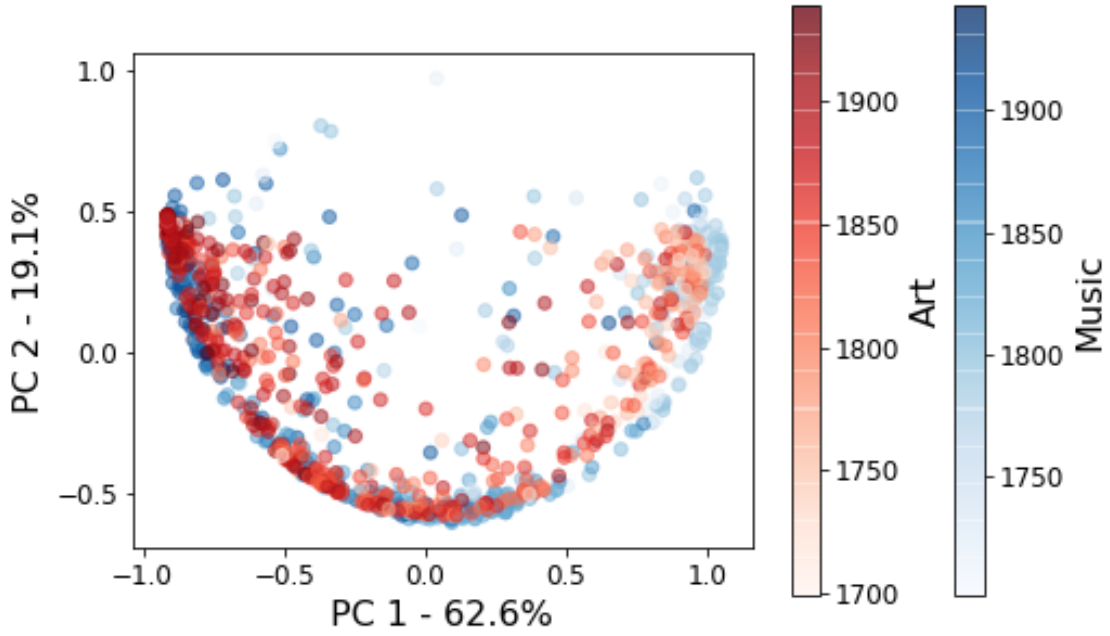


Figure 5.1: Plotting the top two principal components for constrained cross-modal embeddings of validation data

made close to 1950 in the art and music modalities respectively. (For visualization purposes, we plot only a random subset of our data points).

The plot also demonstrates that the difference between modalities is not a significant part of, or does not "explain", at least 81.7% of variance in the joint embedding space. This suggests that our triplet network indeed maps both modalities into a common representation space.

In comparison, Figure 5.2 visualizes the embedding space of the unconstrained embeddings (again, after transformation into the top 2 principal components). We note that there is less symmetry between the art embeddings and the music embeddings. This suggests that modality specific features have more influence in the overall shape of the embedding space under this training regime (compared to the constrained embeddings).

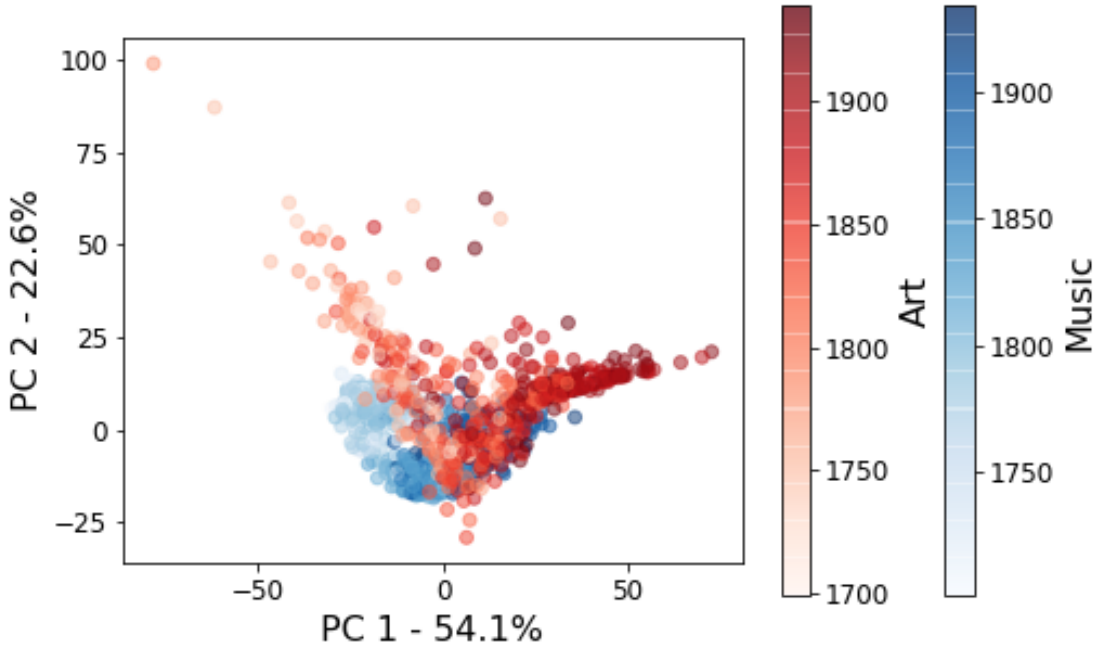


Figure 5.2: Plotting the top two principal components for unconstrained cross-modal embeddings of validation data

5.3 Usefulness of Learned Embeddings for Downstream Dating Tasks

Next, to see whether or not the network has learned a useful representation of the data in a quantifiable way, we can fit a simple linear model to the learned embeddings.

To do so, we first embed the entire training set and validation set, for both art and music modalities, using our cross-modal triplet network. Then, we fit three separate linear regression models: (1) using the training set embeddings from both art and music, (2) using the training set embeddings from just the art modality, and (3) using the training set embeddings from just the music modality. Next, for each of these simple linear regression models, we see how well it fits the validation set embeddings (1) from both art and music, (2) from just the art modality, and (3) from just the music modality.

Finally, for comparison, we embed the entire training set and validation set using the single-modality triplet networks. Then, we fit a linear regression model to the art embeddings and a model to the music embeddings, and use these two models to predict their respective validation embeddings.

	Test on cross	Test on art (from cross)	Test on mus (from cross)
Fit on cross	29.24	33.70	28.05
Fit on art	250.32	33.54	308.03
Fit on mus	29.22	33.66	28.03

Table 5.1: Using constrained embeddings from cross-modal network to regress to year. Values represent mean absolute error of predictions on validation data.

	Test on cross	Test on art (from cross)	Test on mus (from cross)
Fit on cross	27.20	30.41	26.35
Fit on art	212.43	29.36	261.17
Fit on mus	45.63	25.96	119.96

Table 5.2: Using unconstrained embeddings from cross-modal network to regress to year. Values represent mean absolute error of predictions on validation data.

	Test on art (from art)	Test on music (from music)
Fit on art	27.78	N/A
Fit on mus	N/A	23.76

Table 5.3: Using constrained embeddings from single modality networks to regress to year. Values represent mean absolute error of predictions on validation data.

The goal of this comparison is to see whether or not the "support" of the additional modality in the joint cross-modal representation helps generate embeddings that are more generalizable.

Based on the results presented in Tables 5.1, 5.2, and 5.3, the embeddings generated by the cross-modal network are not better than the single-modality network for regressing the creation date, but they do yield similar results. The exception to this is the result from fitting a linear regression model to the constrained art embeddings generated by the cross-modal network, then trying to predict values for the constrained music embeddings generated by the same cross-modal network.

In general, however, while we are not able to demonstrate an improvement over single-modality learning, we are able to align the embedding spaces of art and music without losing too much representation power.

5.4 Visual Explanations: What Does the Network See?

We want to be able to understand what features of the original data the network is utilizing to embed each piece of music or art. These features are useful to look at, since they were how the network determined that a particular piece of music and art were "similar" to one another, according to the metric where two works are "similar" if created close in time to one another. Thus, they could be good candidates for the basis of cross-media style.

Computing Activation Maps

We take the approach of [4] to generate activation maps. Typical gradient-weighted class activation maps (Grad-CAMs) are used to visualize the "important" regions of the input image to a CNN by using the gradients with respect to the target class [31]. With some adaptation, we can generate activation maps for our embedding networks, even without class labels.

To compute the activation map for a single anchor image (or spectrogram), we need the k feature maps A^k output by the last layer of the internal CNN during the forward pass, as well as the gradient of our loss function with respect to those feature maps,

$$g(A^k) = \frac{\delta \mathcal{L}_{tri}}{\delta A^k}$$

The "grad-weights" are then calculated exactly as in the original Grad-CAM work:

$$a_k = \frac{1}{Z} \sum_i \sum_j g(A^k)_{i,j}$$

where i and j are spatial positions in the feature map A^k and Z is the total number of spatial positions (i.e. $i \times j$).

To create the final activation map for visualization, we multiply each feature map by its corresponding grad-weight and sum these weighted maps together. The result is then scaled (in resolution) to the size of the input image, and overlayed on the original input image as a heat map.

In practice, and as described in [4], we take the 50 feature maps (out of the 512 output by the CNN) with the highest corresponding grad-weights.

Additionally, since the triplet loss relies on the embeddings of all three inputs, not just the anchor (for which we are computing the activation map), we compute an average activation map by running this process 15 times.

Visualizations

Figure 5.3 shows, for four input images, the activation map resulting from VGG (before fine-tuning), the activation map resulting from the single-modality art network, and the activation map resulting from the cross-modal network.

As a general trend, we see that the VGG network "looks" at high-level semantic features wherever possible, which makes sense as it is initially trained for object detection. We can also see that its activations make the most sense on paintings that *have* clearly depicted objects (i.e. are more realistic).

In comparison, we see that the specifically trained art network and the cross-modal network tend to focus less on looking at individual objects, and more either at regions where objects meet one another (i.e. edges) or at background regions. We speculate that this is because this is where the style of the artwork is most easily detectable, and style is a strong signal for the time period a painting was created (even though style is not the signal the network is explicitly trained on).

We are not always able to interpret the difference between the activation maps from the single modality network vs. the cross modal network, but we are able to note that they frequently differ substantially. We acknowledge could simply be the result of variance, but it is also possible that it demonstrates the cross-modal network is incorporating some additional signals from the other modality (music).

There are of course still cases where the cross-modal network is less successful at embedding the inputs, and where the resulting CAMs do not lend themselves to any immediate potential conclusions. We visualize two such cases in Figure 5.4. In Figure 5.4a, the activation map from the cross-modal network essentially highlights the entire image, and the highest value activations do not correspond to either any highly textured regions *or* any semantically meaningful regions. Figure 5.4b is an example of a case where the activation maps from VGG, the single-modality network, and the cross-modal network don't differ substantially from one another, in addition to the cross-modal network not being particularly successful at embedding this painting (given the loss value).

We can also visualize Grad CAMs on top of the input spectrograms for the music portion of our network (Figures 5.5 and 5.6. Interpreting the activation maps on top of spectrograms is even more challenging than interpreting it for paintings, since music is of course meant to be listened to. However, we can still easily observe the trend that the activations from the cross-modal network are substantially different from either not-finetuned VGGish or the single-modality network. Though we still can't say definitively whether or not this is a direct result of incorporating features from the art domain, it supports the idea that learning a joint embedding space impacts the specific features learned by the network.

Additionally, compared to not-finetuned VGGish, the trained single-modality music network and cross-modal network show activations that are seem more related to musical structure. Additionally, compared to the single-modality music network's activations, the cross-modal network's activations seem more spatially coherent on the spectrogram, which corresponds to more continuity in the temporal and frequency domains.

5.5 Further Considerations and Future Work

Dataset

Cross-modal learning starts with the construction of a cross-modal dataset. For this work, we compiled a cross-modal painting and classical music dataset as a first step to learning a joint embedding of these two very different modalities.

Due to limited existing metadata and labels, we used a fairly weak link between the two modalities (and therefore a fairly weakly supervised approach to this task). While that does allow us to train a network that does not directly incorporate the biases of human labels, it limits the analysis we are able to do on the results of our network.

For example, having better style labels across modalities would allow us to use classification as a downstream task to validate the representation learned by our network.

Aside from the construction of the dataset itself, as mentioned in Section 5.1, differing amounts of data in one modality compared to another could impact the quality of the learned embedding space. In future work, we might want to explore the impacts of this further by intentionally limiting the amount of training data in one modality or the other. We could see whether the joint embedding space would improve, or whether parts of it would simply become weaker.

Generating Triples

We chose to generate triples where the anchor and positive inputs were created within 10 years of each other, and the anchor and negative inputs were created at least 20 years apart from one another.

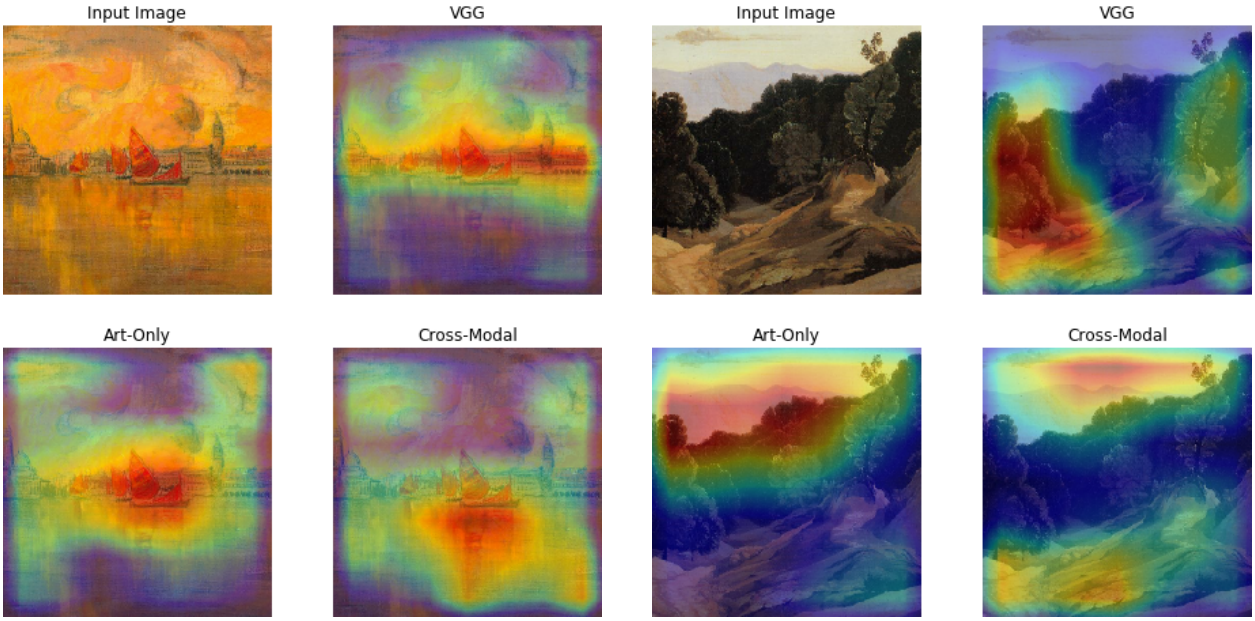
It would be interesting to explore how these "windows" impact the final representation. At the extreme, what if we essentially binarized the dataset into works created before some year and works created after that year?

Constrained ℓ_2-norm

While several works have empirically shown that constraining the embeddings to the unit hypersphere is effective, it is still not completely demonstrated why that is the case. Though we experimented with it here and similarly noted that the training process seemed smoother and more stable, it may not have been the correct decision for our specific data.

First, our work differs from others in that we are trying to contrast between and learn a representation of two completely different data types. Second, the nature of our triplet generation does not encourage an embedding space that is spherical in nature – there is a clear "start" and "end" to our data due to its chronological nature. This can be seen in Figure 5.1: only approximately half of the sphere (or circle, projectected to 2D) is utilized.

However, there could be other types of constraints or regularizations to explore that would increase training stability, while being more suited to our chronological data.



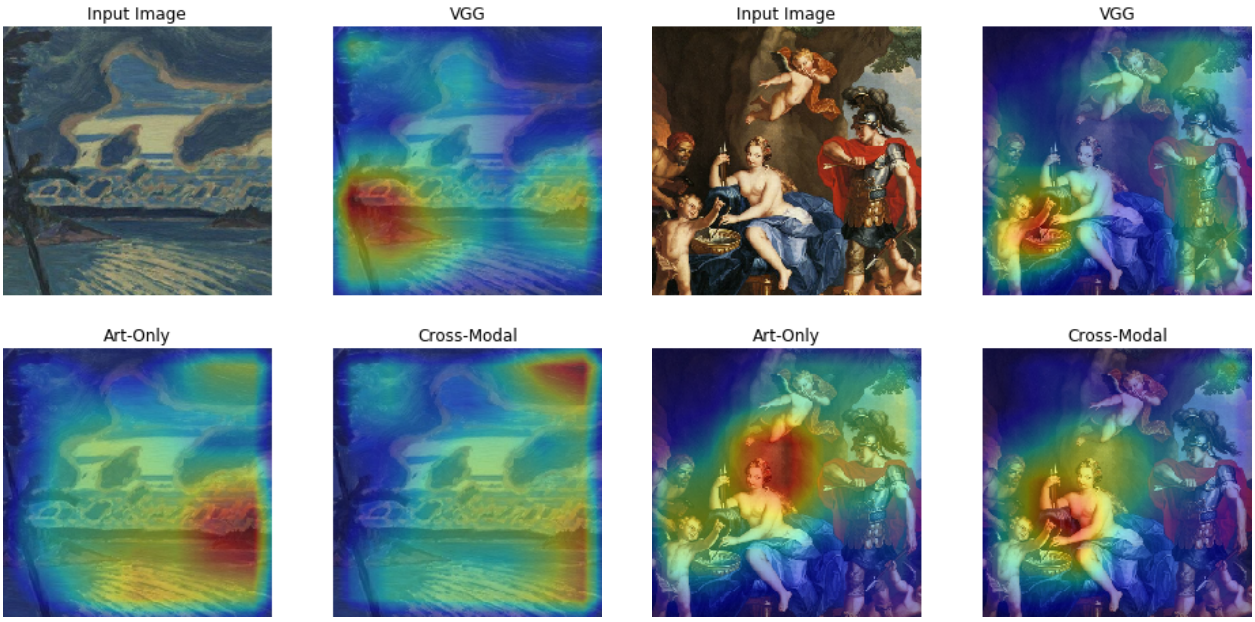
(a) " View of Venice from the sea", Cottet, 1896. Loss value: 0.49

(b) "Road through Wooded Mountains", Corot, 1830. Loss value: 0.32

(c) "Eurybates and Talthybios Lead Briseis to Agammemnon", Tiepolo, 1757. Loss value: 0.56

(d) "Portrait of an unknown man", Kiprensky, 1811. Loss value: 0.56

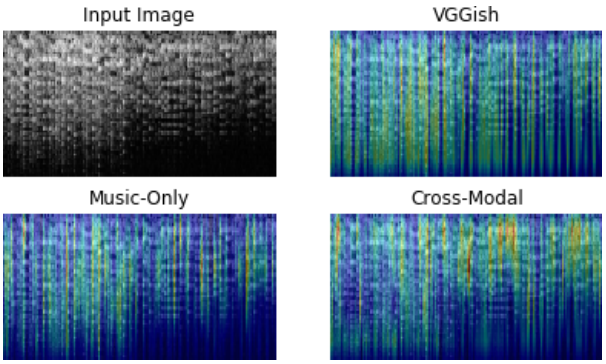
Figure 5.3: Grad-CAMs for four different input images where the cross-modal triplet network performed "well". The reported loss value is the average loss value for the 15 sampled triples used to compute the Grad-CAM. We visualize the resulting Grad-CAMs from un-finetuned VGG, the single-modality art network, and the constrained cross-modal network.



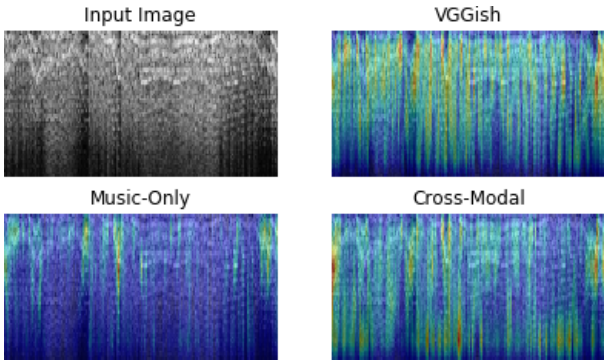
(a) "After Sunset, Georgian Bay", MacDonald, 1931. Loss value: 1.02

(b) "Mars, Venus and Vulcan: the forge of Vulcan", Copley, 1754. Loss value: 1.02

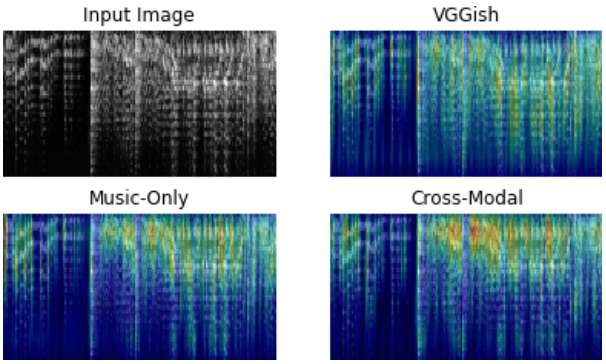
Figure 5.4: Grad-CAMs for four different input images where the cross-modal triplet network did not perform "well". The reported loss value is the average loss value for the 15 sampled triples used to compute the Grad-CAM. We visualize the resulting Grad-CAMs from un-finetuned VGGish, the single-modality music network, and the constrained cross-modal network.



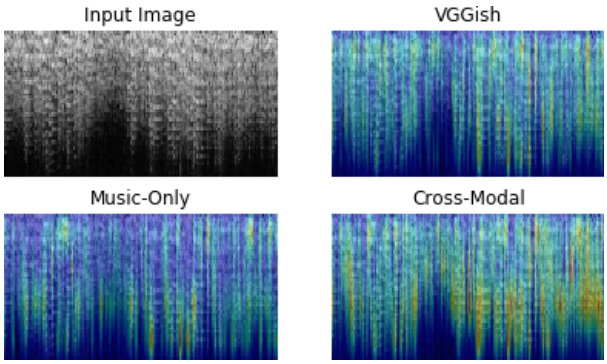
(a) "Prelude and Fugue in D-sharp Minor, WTC II, BWV 877", Bach, 1740. Loss value: 0.45



(b) "Piano Concerto No 26 in D, K 537 iii D major", Mozart, 1788. Loss value: 0.36

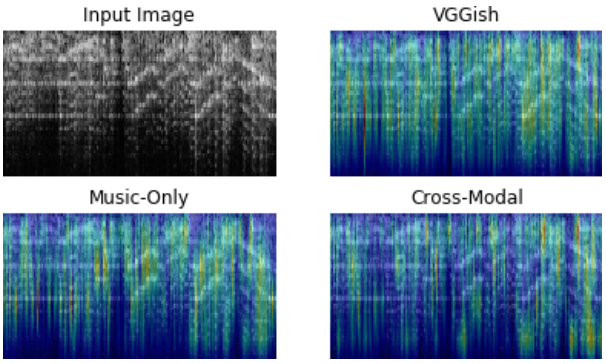


(c) "Sonata No. 18 in E-flat Major, Op. 31, No. 3 (Complete)", Beethoven, 1802. Loss value: 0.44

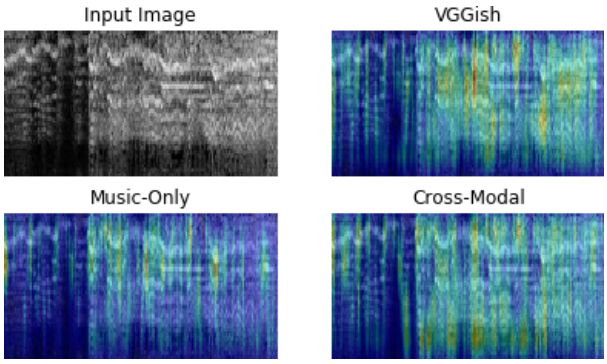


(d) "Prelude, Choral et Fugue", Franck, 1884. Loss value: 0.42

Figure 5.5: Grad-CAMs for four different input images where the cross-modal triplet network performed "well". The reported loss value is the average loss value for the 15 sampled triples used to compute the Grad-CAM. We visualize the resulting Grad-CAMs from un-finetuned VGGish, the single-modality music network, and the constrained cross-modal network.



(a) "Paganini Etude 3 S141 III Ab Minor", Liszt, 1851. Loss value: 1.43



(b) "Cosi fan tutte K588 No25 Aria E major", Mozart, 1790. Loss value: 1.04

Figure 5.6: Grad-CAMs for four different input images where the cross-modal triplet network did not perform "well". The reported loss value is the average loss value for the 15 sampled triples used to compute the Grad-CAM. We visualize the resulting Grad-CAMs from un-finetuned VGGish, the single-modality music network, and the constrained cross-modal network.

Chapter 6

Conclusion

Inspired by prior work in contrastive representation learning and cross-modal/cross-media learning, in this work, we apply these state-of-the-art techniques to a novel domain, namely classical painting and music.

To do so, we first collect an art-music dataset designed for cross-modal comparisons. We hypothesize that paintings and music created during the same time period should be related to one another, or at least more similar to one another than paintings and music created from different time periods. This formulation lends itself to the metric learning task (via the triplet loss). Under the limitations of available ground-truth labels, we learn a cross-modal embedding space to align these two very different modalities of art and music.

In doing so, we demonstrate that it is possible to learn a common representation between music and art modalities, even with this limited supervision.

We also visualize and propose a few interpretations of how different features are utilized by a network trained cross-modally, as compared to a network trained for a single modality.

Broadly motivated by the idea that the cross-modal psychological perception that exists in humans yields cross-modal *expression*, we hoped to see how analyzing cross-modal or cross-media creative works could bring insights to how cross-modal correspondences are perceived by people. However, it is evident that this itself is a "chicken and egg" type problem, and one that requires a feedback loop that incorporates actual human perceptual studies.

Nevertheless, the fact that we can create a shared representation for these two modalities is a useful tool for future studies. For example, the features extracted by our cross-modal network could then be used to isolate specific regions of paintings, or specific portions of a musical piece, and be used as part of a study on how humans perceive those features specifically.

Alternatively, the shared representation could be used to answer questions about the shared sociocultural context from which these works were created.

The flexibility of the triplet contrastive learning model allows us to use any labels that may be available to us and that are suitable to the question that we are interested in. We chose to use the year the work was created, which yielded an embedding space that we interpret as chronologically organized.

Other forms of supervision would yield embedding spaces organized in other ways. However, the nature of the triplet loss (or other metric learning losses) generates an embedding space where distance in the resulting space relates well to the original semantic meaning tied to the supervisory labels. Additionally, the notion of labeling inputs based on their similarity or dissimilarity to other inputs is a generally accessible one. Thus, researchers without specific computational knowledge, but with domain knowledge in art or music, for example, can easily create example triplets for supervision without specific technical knowledge or background. For that reason, we believe triplet-based supervision is intuitive and well-suited to applications in the digital humanities. We hope to see future models and frameworks that take advantage of this.

Bibliography

[1] Liliana Albertazzi, Luisa Canal, and Rocco Micciolo. "Cross-modal associations between materic painting and classical Spanish music". In: *Frontiers in Psychology* 6 (2015), p. 424.

[2] Eric Brochu. "The sound of an album cover: Probabilistic multimedia and IR". In: *Ninth International Workshop on Artificial Intelligence and Statistics*. 2003.

[3] Jiansong Chao et al. "A Semantic-Driven Music Recommendation Model for Digital Photo Albums". In: Jan. 2013, pp. 369–381.

[4] Lei Chen et al. "Adapting Grad-CAM for Embedding Networks". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2020.

[5] Mingwen Dong. "Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification". In: *ArXiv* abs/1802.09697 (2018).

[6] Amanda Catherine Duthie. "Do music and art influence one another? Measuring cross-modal similarities in music and art". MA thesis. Iowa State University, 2013.

[7] Ahmed Elgammal et al. *The Shape of Art History in the Eyes of the Machine*. 2018.

[8] Patrick Georges. "Western classical music development: a statistical analysis of composers similarity, differentiation and evolution". In: *Scientometrics* 112 (Apr. 2017). DOI: 10.1007/s11192-017-2387-x.

[9] Shiry Ginosar et al. "Detecting People in Cubist Art". In: *Computer Vision - ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Springer International Publishing, 2015.

[10] Ruslan Salakhutdinov Gregory Koch Richard Zemel. "Siamese neural networks for one-shot image recognition". In: *ICML Deep Learning Workshop*. 2015.

[11] William Griscom.

[12] W. Guo, J. Wang, and S. Wang. "Deep Multimodal Representation Learning: A Survey". In: *IEEE Access* 7 (2019).

[13] W. D. Hairston et al. "Visual Localization Ability Influences Cross-Modal Bias". In: *Journal of Cognitive Neuroscience* 15.1 (2003), pp. 20–29.

[14] Curtis Hawthorne et al. "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=r1lYRjC9F7.

[15] Shawn Hershey et al. "CNN Architectures for Large-Scale Audio Classification". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[16] *International Music Score Library Project*. URL: https://imslp.org/wiki/Main_Page.

[17] S Karayev et al. "Recognizing image style. BMVC". In: (2014).

[18] Wendy M. Limbert and Donald J. Polzella. "Effects of Music on the Perception of Paintings". In: *Empirical Studies of the Arts* 16.1 (1998), pp. 33–39.

[19] Xin Liu et al. "CNN based music emotion classification". In: *ArXiv* abs/1704.05665 (2017).

[20] Utkarsh Mall et al. "GeoStyle: Discovering fashion trends and events". In: *ICCV*. 2019.

[21] Martın Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: http://tensorflow.org/.

[22] Harry McGurk and John MacDonald. "Hearing lips and seeing voices". In: *Nature* 264 (Dec. 1976), pp. 746–748.

[23] James D. Merriman. "The Parallel of the Arts: Some Misgivings and a Faint Affirmation: Part 1". In: *The Journal of Aesthetics and Art Criticism* 31.2 (1972), pp. 153–164.

[24] Fred A. Minnigerode, David W. Ciancio, and Lori A. Sbarboro. "Matching Music with Paintings by Klee". In: *Perceptual and Motor Skills* 42.1 (1976), pp. 269–270.

[25] M. Mueller et al. "Cross-Modal Music Retrieval and Applications: An Overview of Key Methodologies". In: *IEEE Signal Processing Magazine* 36.1 (2019), pp. 52–62.

[26] Colin Martindale Nancy Hasenfus and Dana Birnbaum. "Psychological Reality of Cross-Media Artistic Styles". In: *Journal of Experimental Human Psychology: Human Perception and Performance* 9.6 (1983), pp. 841–863.

[27] Casey O'Callaghan. *Beyond Vision: Philosophical Essays*. 2017.

[28] Stephen E. Palmer et al. "Music–color associations are mediated by emotion". In: *Proceedings of the National Academy of Sciences* 110.22 (2013), pp. 8836–8841. ISSN: 0027-8424.

[29] V.S. Ramachandran and E.M. Hubbard. "Synaesthesia — A Window Into Perception, Thought and Language". In: *Journal of Consciousness Studies* 8.12 (2001), pp. 3–34.

[30] Babak Saleh et al. "Toward automated discovery of artistic influence". In: *Multimedia Tools and Applications* 75 (2014), pp. 3565–3591.

[31] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128 (2019), pp. 336–359.

[32] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015.

[33] Charles Spence. "Crossmodal correspondences: A tutorial review". In: *Attention, perception psychophysics* 73 (May 2011), pp. 971–95.

[34] Charles Spence. "Multisensory Perception". In: *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. American Cancer Society, 2018, pp. 1–56. ISBN: 9781119170174.

[35] Feng Wang et al. "NormFace: L2 Hypersphere Embedding for Face Verification". In: *Proceedings of the 25th ACM International Conference on Multimedia*. Association for Computing Machinery, 2017, pp. 1041–1049.

[36] Tongzhou Wang and Phillip Isola. "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere". In: *arXiv:2005.10242* (2020).

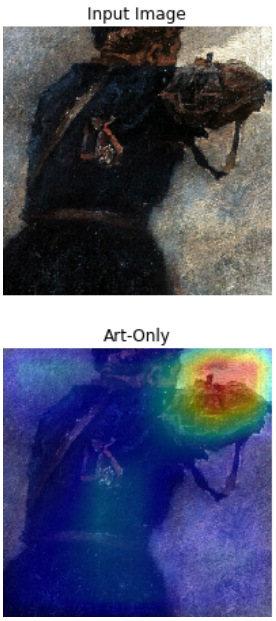
[37] Jamie Ward, Brett Huckstep, and Elias Tsakanikos. "Sound-Colour Synaesthesia: to What Extent Does it Use Cross-Modal Mechanisms Common to us All?" In: *Cortex; a journal devoted to the study of the nervous system and behavior* 42 (Mar. 2006), pp. 264–80.

[38] Christopher Wm. White and Ian Quinn. *Yale-Classical Archives Corpus*. 2014.

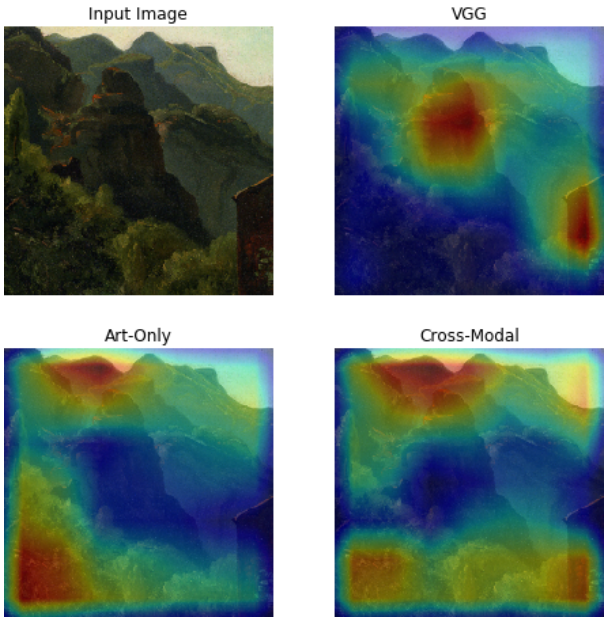
[39] D. Zeng, Y. Yu, and K. Oyama. "Audio-Visual Embedding for Cross-Modal Music Video Retrieval through Supervised Deep CCA". In: *2018 IEEE International Symposium on Multimedia (ISM)*. 2018, pp. 143–150.

Appendix A

Additional Activation Map Visualizations

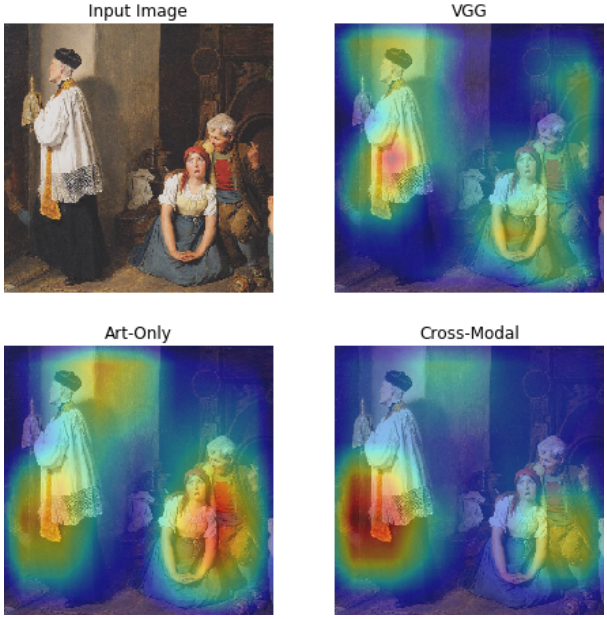


(a) "Shooting Cossack", Surikov, 1895, Loss: 0.60

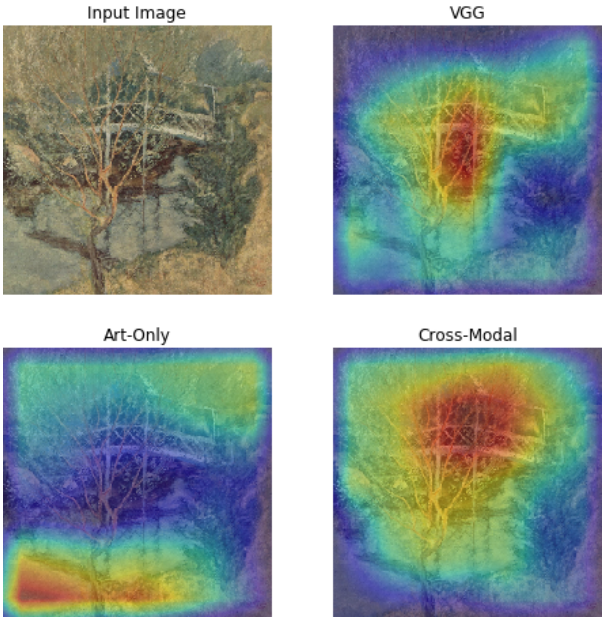


(b) "Side of the Valley of Saint-Vincent (Auvergne)", Rousseau, 1830, Loss: 0.48

(c) "Arab on Camel", Vereshchagin, 1870, Loss: 0.82

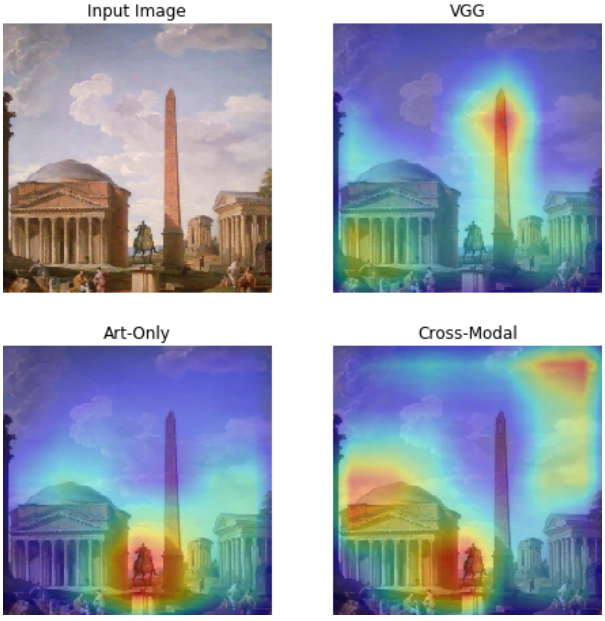


(d) "Extreme Unction", Waldmuller, 1846, Loss: 0.83

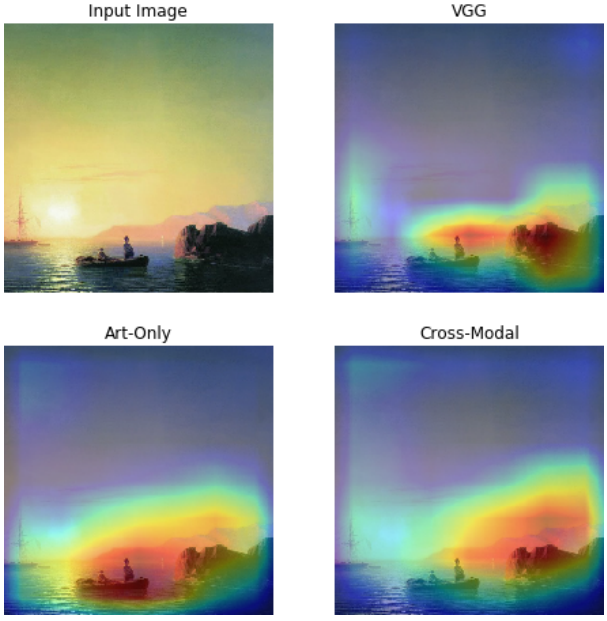


(e) "The White Bridge", Twachtwman, 1897, Loss: 0.81

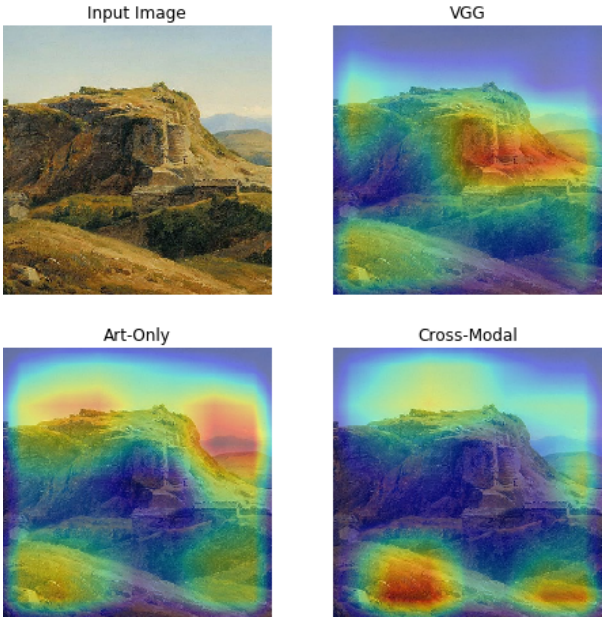
(f) "Portrait of Torsukov Ardalyon", Borovikovsky, 1795, Loss: 0.74



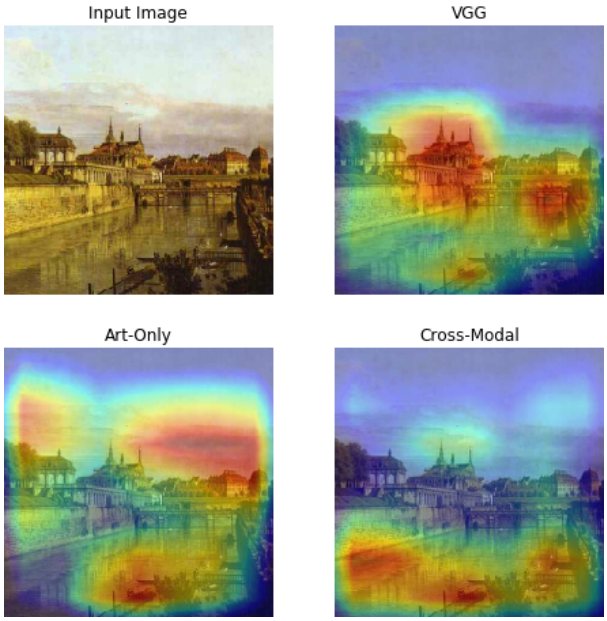
(g) "Roman Capriccio: The Pantheon and Other Monuments", Panini, 1735, Loss: 1.07



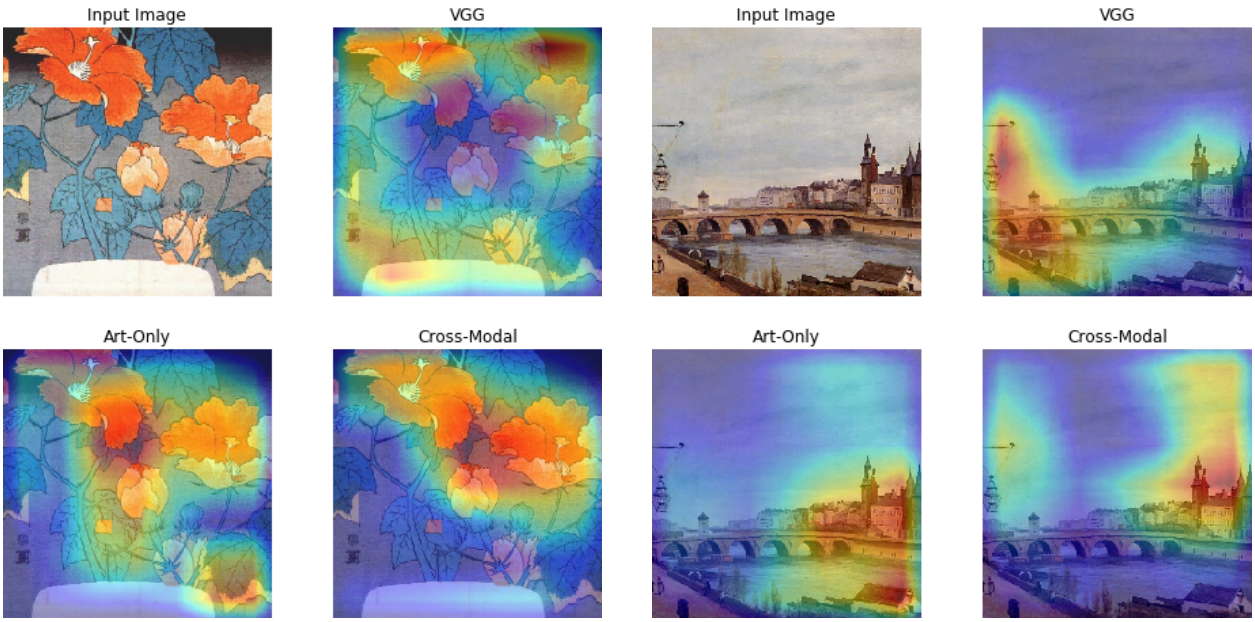
(h) "Sunset at the Crimean coast", Aivazovsky, 1856, Loss: 0.49



(i) "Hilly landscape, Auvergne", Rousseau, 1830, Loss: 0.46



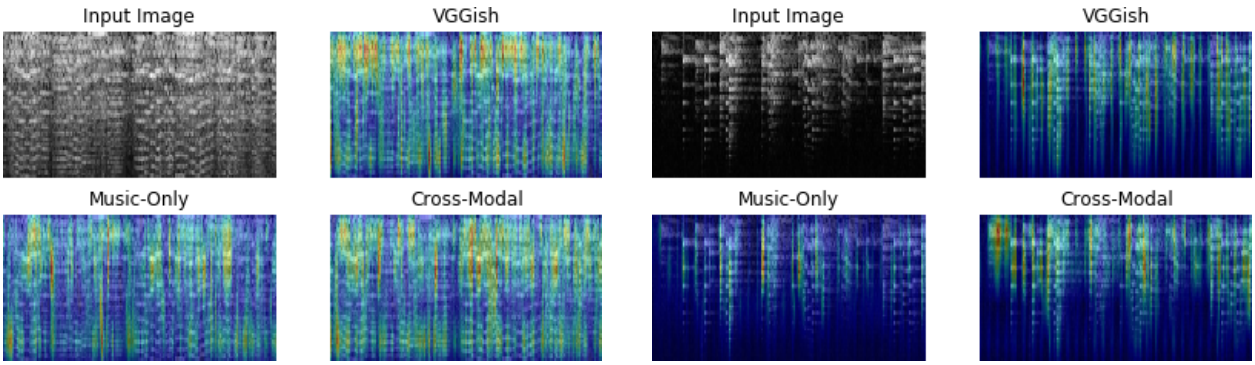
(j) "The Moat of the Zwinger in Dresden", Bellotto, 1750, Loss: 1.11



(k) "Hibiscus", Hiroshige, 1845, Loss: 0.74

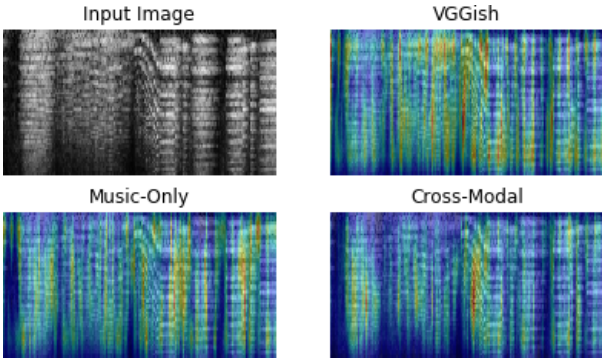
(l) "View of the Pont au Change from Quai de Gesvres", Corot, 1830, Loss: 0.52

Figure A.1: Additional activation maps for art

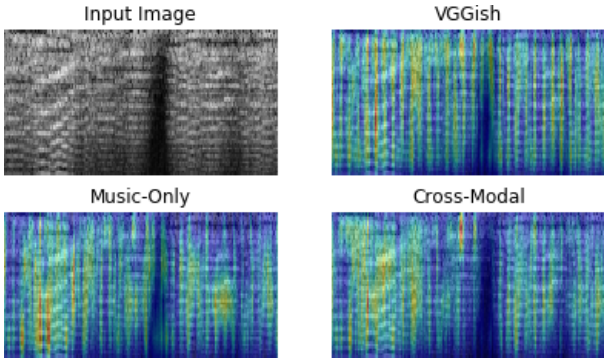


(a) "String Quartet Op55 No2 i F major", Haydn, 1788, Loss: 0.84

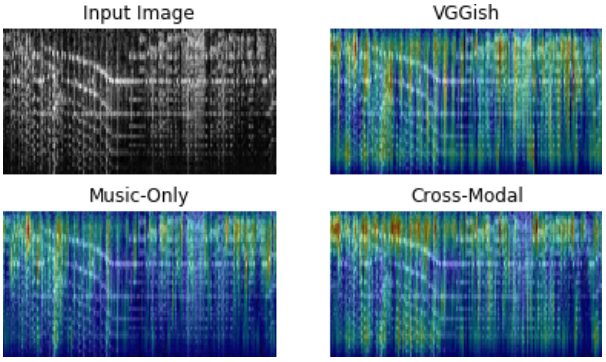
(b) "Sonata No. 4 in F-sharp Major ,Op.30", Scriabin, 1903, Loss: 0.91



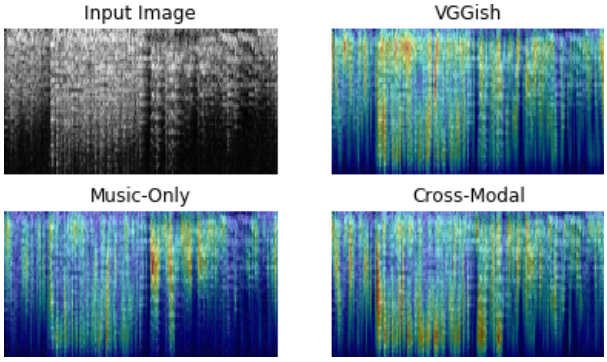
(c) "String Quartet Op74 No3 iv G minor", Haydn, 1793, Loss: 0.54



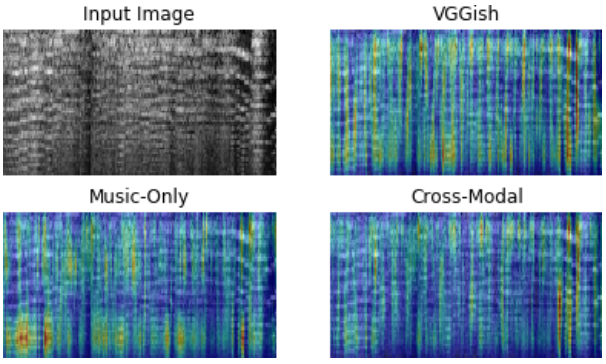
(d) "Cavalleria Rusticana-Intemezzo F major", Mascagni, 1890, Loss: 0.66



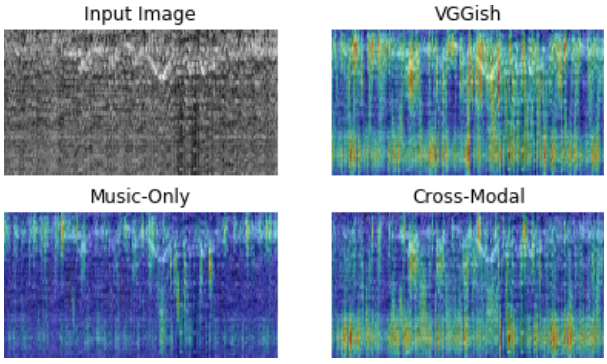
(e) "Grandes Etudes de Paganini No. 3 La Campanella, S. 141", Liszt, 1851 Loss: 0.42



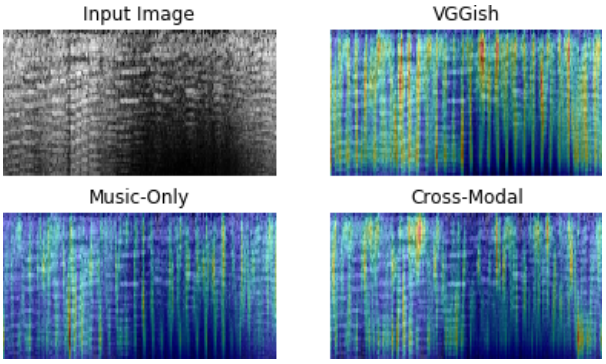
(f) "Piano Sonata No10 Hob16-1 iii C major", Haydn, 1760, Loss: 0.54



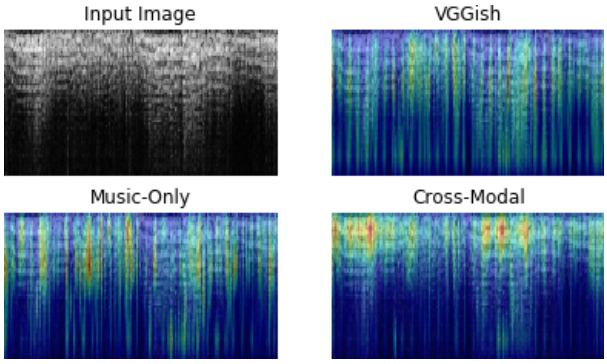
(g) "Allegro HWV 323 D Major", Handel, 1739, Loss: 0.62



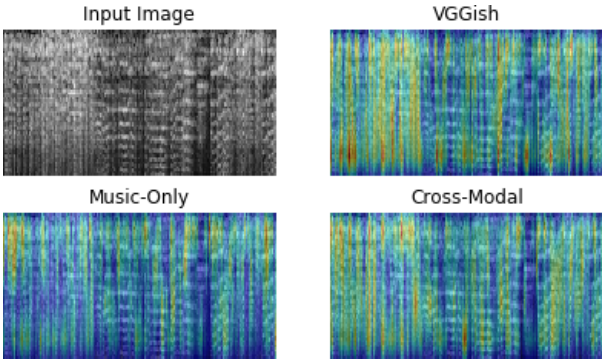
(h) "Sonata for Flute and Piano ii Eb major", Haydn, 1764, Loss: 0.62



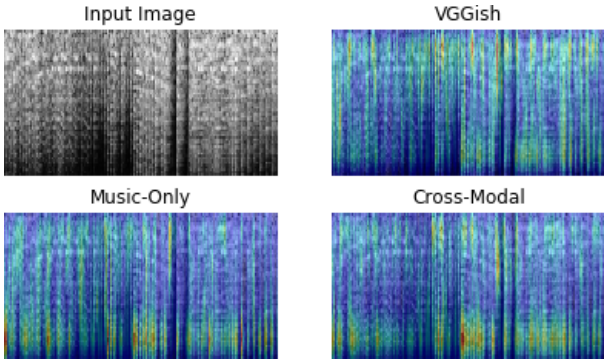
(i) "In the South Alassio Overture Op50 Eb major", Elgar, 1903, Loss: 0.53



(j) "Six Trios Allegro Op82 F major", Reicha, 1912, Loss: 1.23



(k) "Violin Concerto 3, K216 iii G major", Mozart, 1775, Loss: 0.58



(l) "Sonata 2 F-sharp Minor, Op. 2", Brahms, 1852, Loss: 0.75

Figure A.2: Additional activation maps for music