Lateralus takes on Progressive Rock Classification

CAP6610 Final Report
May 2019

Hadi Abdullah	Shreya Singh	Nicholas Kroeger	Ratna Suthar	Washington Garcia
3137-1031	7915-4462	5224-5139	5041-0989	1385-9513

1 Consciousness

1.1 Discussion of supervenience question based on David Chalmers target chapter and the Stanford Encyclopedia exposition.

Stanford definition states that a set of properties named A supervenes upon another set B just in case no two things can differ with respect to A-properties without also differing with respect to their B-properties [2]. In slogan form, there cannot be an A-difference without a B-difference. Figure 1 below helped us visualize the concept of supervenience. Ultimately, supervenience would be the claim that things cannot differ in some respect without differing in some other respect.

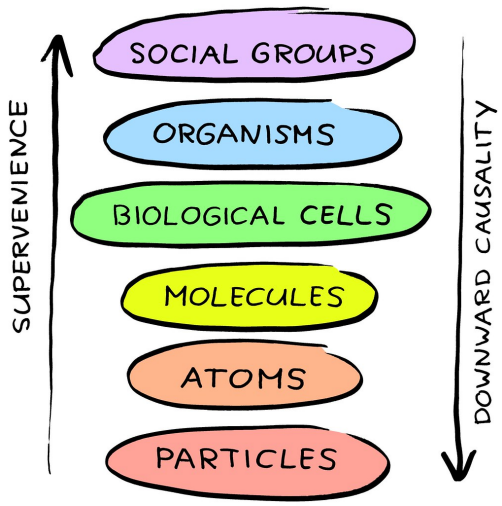

Figure 1: The upper levels on this chart can be considered to be supervenient on the lower levels [1].

1.1.1 Local and Global supervenience

Local: Which comes in the forms of "weak" and "strong" discussed later under logical vs natural [3]. If mental states locally supervene on brain states, then being in the same brain state entails being in the same mental state. The local supervenience can be understood

in terms of the relation between the properties of an object. The formal definition of local supervenience compares the properties which are in two different space-time region within the same world. In a world w and space-time regions r1 and r2 within w, if property B duplicates in region r1 and r2 and property A is locally supervenient to property B then we can conclude that property A also duplicates in the region r1 and r2.

Global: If psychological properties globally supervene on physical properties, then any two worlds physically the same will be psychologically the same. The formal definition of global supervenience states that the properties are of the exact same distribution in two different worlds w1 and w2 if and only if property A globally supervenes property B. [2]
The value of global supervenience is that it allows for supervenient properties to be determined not by local properties of an individual thing alone, but by some wider spatiotemporal distribution of things and properties. For example, something being a dollar bill depends not only on the paper and the ink it is made out of but also on a widely dispersed variety of features of the world it occupies. As another example one can say that the biological properties globally supervene on physical properties, which clearly implies that the world which is physically identical to ours would also be biologically identical.

1.1.2 Natural vs. Logical supervenience

Logical supervenience: In the literature this is referred to as "strong" supervenience [3]. Chalmers gives an example of contradiction: a world in which there exist married bachelors. He claims this world is impossible because it is not logically consistent: "bachelor-ness" is logically supervenient on the property of being not married. Logical supervenience is supported by concepts irrespective of how the actual world turns out. Bproperties logically supervene on Aproperties if any two logically possible situations that have the same Aproperties also have the same Bproperties.

Natural supervenience: Chalmers refers to this as weaker supervenience. In the reading, Chalmers gives the example of gravity. There can exist a world where which bits of matter in space are pushed away from each other instead of coming together, as they do in our world. Thus, the supervenience of gravity on is natural. Another example would be if there were a world where water vaporized at low temperatures instead of freezing into a solid form as it does in our world. Supervenience of thermodynamic laws is natural. Such worlds are not logically impossible - they just have a different set of natural laws. Thus, the class of

naturally supervenient situations is a subset of the class of logically supervenient situations.

1.2 Logical supervenience of consciousness on the physical

1.2.1 Opinion 1 - Not logically supervenient over physical

To state that consciousness is not logically supervenient to physical supervenience, it is required to consider a scenario where two entities with identical physical and psychological conditions, have different experiences. The primary assumption considered with regards to this fact is that there is no primary relationship between physical conditions and phenomenon. Several arguments ascertain the possibility of the stated hierarchy of the supervenience.

Consider the possible case of the existence of zombies, which may display identical physical and psychological behavioral characteristics but lack experiences. The claim refutes its validity by ascertaining that no contradictory theories have been known to exist which disregards it's existence.

Another possible argument to strengthen the stance is given by stating that the intuition of color is a phenomenon which buds out of consciousness. Consider a situation in an ideal world where humans knew everything about the processing of the brain and every information about colors is known. The question arises whether the known knowledge of colors is enough to identify them, without ever knowing the mapping of knowledge to actual colors. This clearly draws a boundary between the physical conditions and the phenomenon conditions and concludes that the mere knowledge of physical conditions is not a necessary and sufficient condition for intuition or phenomenon. No amount of research or understanding has been able to derive the conscious experience in animals. One cannot understand the implication of conscious in animals like mice which are known to have consciousness by virtue of the physical facts. A similar case exists in cognitive computers, where, even by learning to different things as green or red, a computer may not have any experiences about the colors.

The theorists of the belief that consciousness is not logically supervenient question on the non-reductive approach taken by the opposition to falsify the claim. Though, it can be safely concluded that the idea of consciousness being supervenient can be exemplified and reductively explained based on claims, assertions and proofs since they collectively are a part of the counter-examples for a much complex theory.

1.2.2 Opinion 2 - Consciousness is logically supervenient

To say that the consciousness supervenes on the physical is to say that there cannot be conscious differences without physical differences. This indicates that for all worlds with the same distribution of fundamental physical properties and the same natural laws and mental properties must be the same. If we consider that consciousness logically supervenes we have to restrict the claim to a local supervenience rather than a global supervenience. Since a mere physical similarity does not guarantee similar consciousness between two individuals but this argument can be debated based on the observation that a combination of several other factors along with the physical properties might result in the presence of similar consciousness. Thus a feeble claim which supports this notion can occur by considering the supervenience as local rather than global.

It is unconventional to define consciousness in terms of structural and functional properties. The fact that reductive explanations tend to exploit the structural and functional properties in order to prove the logical supervenience, any detailed explanation for reduction would fail to produce a fulfilling explanation for consciousness due to absence of structural and functional properties. This helps us achieve our objective of proving that no reductive explanation is enough to show that consciousness cannot be logically supervenient. In fact, it helps us strengthen our argument about finding bridging principles between consciousness and underlying physical processes. One such theory which explains such supervenience is the psychophysical theory which strongly believes in determining the conscious behaviour based on experiences. For example, consider the theory of color, we experience different colors due to varying hue, saturation and intensity. Properties can be covered by explaining the light waveforms which results in different colors. Hence, we can safely say that failing to find a relation between experience and conciousness is a failure towards finding meaningful reasons behind the way phenomenon occurs. Thus we might as well say that providing a non reductive explanation and the psychophysical theories together form a strong stance for conciousness being logically supervenient.

1.3 Relevance to present day AI

Artificial intelligence is a discipline which fosters the development of intelligence in machines. Theories suggest that it is possible to build a machine which can surpass human

intelligence but the question of a conscious mind remains unanswered. The subject of logical supervenience of consciousness comes into play while the modern day AI builds machines similar and efficient as humans.

Suppose we consider consciousness to be logically supervenient which means that entities with identical physical properties would a have similar conscious mind. If this stands true, one can build machines with a conscious mind and falsify the claims about machines not being conscious. To refute this notion, one may consider doing a reductive in-depth analysis of the origins of consciousness. Computers are thought off to be simply mathematically programmed machines, however computers work in a way similar to the human brain. Like neurons in human brain, the circuits and voltages which are the building blocks of computers, interact amongst each other is a causal manner. Since consciousness arises from the neurons and their interactions, it should be possible to replicate this consciousness in machines.

However, several people reject the notion of supervenience and that machine can develop a conscious mind. One such example of such a situation is that even though we communicate with computers in English language, it is the sense of understanding of the language which is absent. As another example is that, one cannot merely develop a system which can identify flaws in a program or identify cases in which program would fail. People who oppose the supervenience theory refute that the tasks of a conscious mind cannot be replicated in computers. This also interpolates to the pessimistic and reductive approach towards consciousness being supervenient which disregards the presence of physical similarity being an indicator for the presence of a conscious mind.

2 Progressive Rock Classification - Discussion of Techniques Used

The focus of this project is genre classification, a typical task from the field of music information retrieval (MIR). MIR is the study of ways to retrieve information from music for applications such as music generation, automatic music transcription, and recommender systems. Music genres are top-level descriptors of songs based on many characteristics such as cultural origin, instruments, and harmonic and melodic structure. Since many genres overlap with each other, genre classification is a nontrivial task that is embedded in ambiguity. We propose several models to distinguish a songs' genre as being either progressive rock or non-

progressive rock. The methods used are outlined below and include fully connected neural networks, long short-term memory networks, and convolutional recurrent neural networks.

2.1 Common extracted features across our models

Machine learning models require relatively good features, specifically features that are independent, informative, and discriminatory for the algorithm to be effective. Some features suspected to be useful for genre classification include raw audio, signal spectrograms, and mel-frequency cepstral coefficients.

2.1.1 Mel-Frequency Cepstral Coefficient

Current state of the art features for genre classification include Mel-Frequency Cepstral Coefficients (MFCCs), although their origins lie within speech recognition. In many of our models, we used these MFCC features and statistics (mean and covariance) from these features derived from an audio signal. Figure 2 shows the order of calculations to produce the MFCCs.

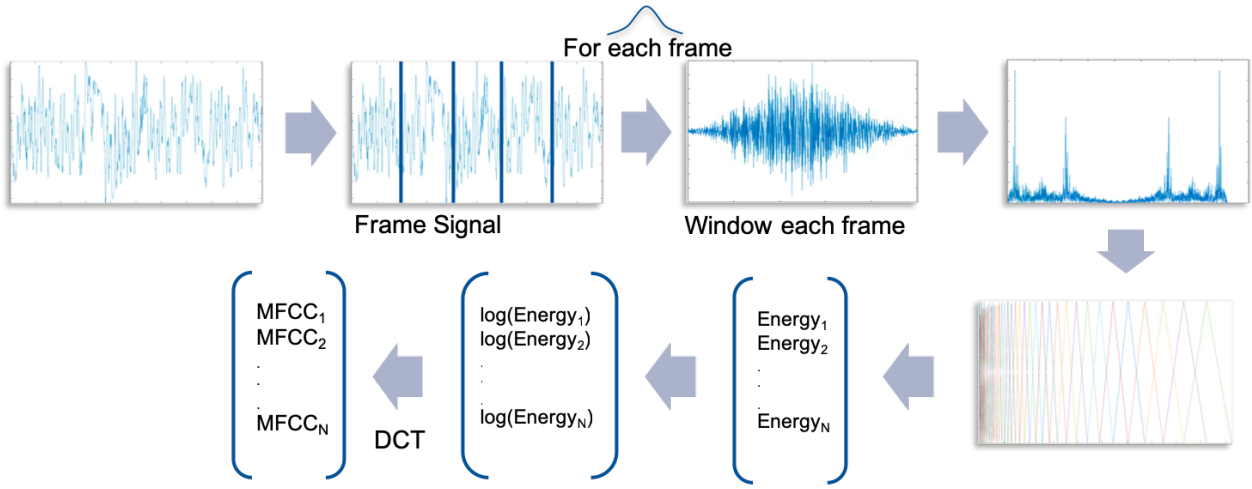

Figure 2: A flow diagram of how MFCC features are calculated from a raw audio signal. (DCT is the Discrete Cosine Transform)

2.1.2 Spectrogram

Another important feature is a time-frequency representation of the audio signal, which shows how the frequencies vary over time. The spectrogram in Figure 3 is represented as a heat map where the color represents the intensity of the frequency at that bin at a given time instant.

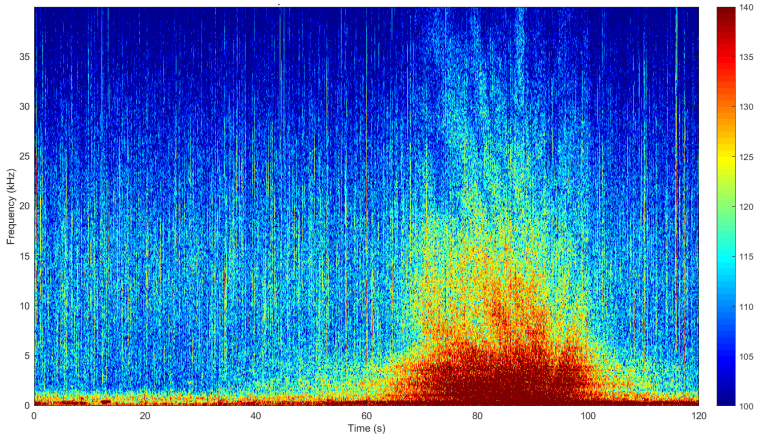

Figure 3: An example of a spectrogram showing frequencies on the y-axis as they vary across time on the x-axis.

2.2 Fully Connected Neural Network

We developed a fully connected neural network to discriminate between progressive and non-progressive rock. The features used for this model include the average of the MFCCs for each frame across time as well as the covariance of all MFCCs across time. We used 20 MFCCs per frame and thus 210 covariance values (since the covariance matrix is symmetric, take half of the matrix values).

The model architecture included an input layer of size 230 nodes, two hidden layers of size 200 nodes and 50 nodes, respectively, and an output layer of one node. Between each layer was a rectified linear unit activation function and the loss function used was the mean squared error. The labels for progressive rock was arbitrarily chosen to be one and a negative one for non-progressive rock. To classify a song, the decision was if the network output value less than or equal to 0, it was considered to be non-progressive rock, and progressive rock otherwise. Some network auxiliary parameters included a learning rate of .001 and an Adam optimizer. The stopping criteria for the optimal network weights were set using the validation

set provided. The network stopped updating weights if the validation loss increased as the training loss decreased; this method helps minimize overfitting to provide optimal model generalization.

2.2.1 Dimensionality Reduction

One dimensionality reduction technique investigated was Principal Component Analysis (PCA). PCA reduces dimensionality by projecting to a lower subspace with maximum variance. This procedure takes the correlated variables finds a smaller representation of them that are uncorrelated.

2.3 CNN RNN DC Model

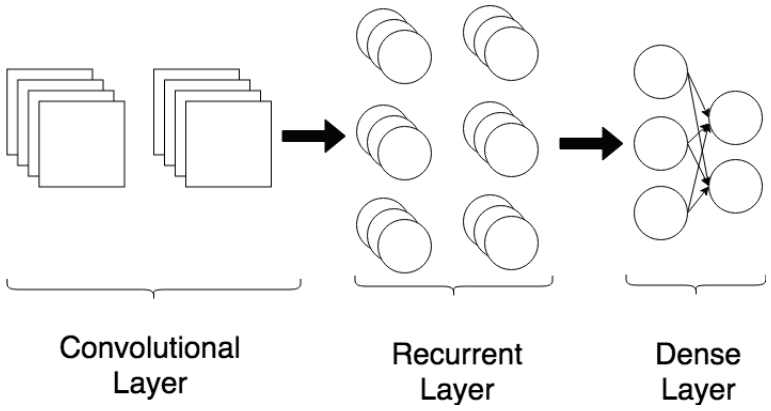

Figure 4: The CNN-RNN-DS Pipeline. The MFCC vector is passed on the the convolutional layers for feature extraction. The features are then passed to the recurrent layer to capture temporal information. Lastly, the outputs are propagated to the dense layer for label assignment.

The task of progressive music identification is a classification problem in the audio domain. One of the more complex problems in the audio domain is that of speech recognition. We wanted to investigate whether we could use the same techniques used for speech recognition to music identification. Modern speech recognition systems are comprised of multiple steps. First, the audio passed through a feature extraction function. Next, the extracted features are given to a convolutional layer for further information extraction. Next, this output is

passed onto the recurrent layer to capture information about the context and past time steps. Lastly, the output of the recurrent layers is passed to the dense layer for label assignment.

We implemented the above-described architecture, using the MFCC algorithm for feature extraction. Next, the features are passed onto the actual neural network, shown in Figure 4. The features passed to the three convolutional layers, each with 14 filters of size 3x3. Next, the output is passed to the six recurrent layers of size seven. The recurrent layers comprise of LSTM neurons. Lastly, the result is sent to the dense layer. For brevity, we omit particular results for this architecture, as it served more as a learning experience for future architectures.

2.4 Encoder-Decoder LSTM with Windowed MFCC Mean Features

Based on discussion during the later class lectures, we were interested in expanding the previously mentioned architectures into a deep neural network. In order to process very long audio sequences, we follow the general seq2seq framework which proposes the encoder-decoder architecture [9]. The encoder-decoder architecture leverages the principle of feature bottlenecks between recurrent layers to achieve acceptable performance in sequence-to-sequence translation tasks. To obtain a usable input sequence, each song is broken into equally spaced windows. Empirically, we observed that 15-second windows had the best performance. For each window, the top-20 MFCC features are taken, then averages over the cepstral coefficients. This process yields a sequence of 20-vectors that are suitable for the seq2seq framework [1]. However, in our task, the output sequence is a 1-vector which consists of a single binary classification.

The specific architecture for our encoder-decoder network was chosen empirically. We found that a single LSTM layer with 512 hidden units had best performance for the encoder, while a similar 512 hidden unit LSTM layer was used for the decoder. Two fully connected layers are used to downscale the decoder's hidden layer output to 32 and then 2 hidden units across the sequence, after which a softmax is applied. For all LSTM layers in our experiments, we use batch normalization in either 1D or 2D, based on the LSTM input, and bidirectional hidden units. For our case, each direction is summed rather than concatenated, for ease of

[1] PCA was attempted for the raw MFCC matrix, but it yielded indistinguishable improvement with the penalty of higher computation overhead.

layer portability. In all encoder-decoder experiments, we use Negative Log-likelihood as the criterion, and Adam as the gradient update algorithm. We denote the basic encoder-decoder architecture as **ED** in later charts and figures.

2.5 Residual Encoder-Decoder LSTM with Self-Attention and Windowed MFCC Mean Features

Although the seq2seq framework is acknowledged for suitable performance, it can be improved by further refining the input features of the recurrent layers, and adding a form of unsupervised learning to decide sequential importance. We run experiments on two extra variants of the encoder-decoder architecture.

2.5.1 Residual Convolutions

Before entering the recurrent layers, the MFCC features are passed through a series of 1D residual blocks performing convolutional upscaling. The blocks perform convolutions over the MFCC features by treating each MFCC coefficient (20 in total) as a channel. Thus, the original MFCC feature can be upscaled into a higher latent feature space of arbitrary dimensionality. In our architecture, we upscale the 20-vectors into 1024-vectors, which are passed through three additional residual blocks. In fact, this approach is popular in video captioning networks, which exploit feature upscaling to discover latent features [10, 7, 6]. For the specific residual block configuration, we use ReLU activations after each convolution as it would tend to be more stable in terms of validation accuracy than *tanh*.

From the convolutional layers, the features are fed to the recurrent layers directly. In our architecture, we observed that best performance is achieved by leveraging the encoder as a bottleneck down to 512 hidden units. The hidden units are unchanged when passed to the decoder. This variant of the encoder-decoder is referred to as **ED+Res** in charts and figures.

2.5.2 Self-Attention

The ability to weigh certain windows more than others arises naturally when modelling music as sequences. However, the lack of sequential ground truths in the provided dataset poses a challenge: how can the classification of each window be performed if the classification

only applies to the whole sequence, rather than the parts? This leads to a semi-supervised method of attention, known as self-attention [4]. Compared to the version of attention proposed by Luong [5] and transformer networks [8], self-attention is relatively straightforward to implement, as it is essentially a 2-layer multi-layer perceptron (MLP) acting on the hidden units of the decoder output. Self-attention offers tuning in the form of the MLP dimensionality, in addition to a concept known as "context hops", which allow different contexts to be modelled from the same sequence (which arises naturally in the framework of language translation, where the importance of words can be weighted based on their context). Although very interesting, for brevity we experiment with only one context hop (and thus each sequence is allowed only one set of attention weights). An intriguing propert of self-attention is the ability to interpret the output by visualizng the attention weights over the original sequence. This allows one to view directly which spans of the sequence were weighted more or less by the model.

For evaluation, we refer to this variant of the encoder-decoder architecture as **ED+Res+Att**.

3 Evaluation

For most of our methods which were deep learning models, they require a sufficiently large training dataset; and in general, the more data we have, the better the model will perform and generalize. Thus, one idea we had was to augment our dataset with more songs, but the training set had some ambiguity between class labels. Although some songs were labeled online as a progressive rock genre, they were put into the non-progressive rock folder. Because of the oracle-labeling bias, it made it challenging to augment the training set. In the remaining subsections, we provide confusion matrices showing the number of correct and incorrect model predictions for each class. In addition, we give the model predictions for the songs provided in the djent folder.

3.1 Fully Connected Neural Network

The validation set accuracy was 72% and the test set accuracy was 55% (see figures 5 and 6 below). In both testing cases, the network performed well in deciding whether a song was progressive rock but had difficulty labeling songs as a non-progressive rock.

3.1.1 Confusion matrix results

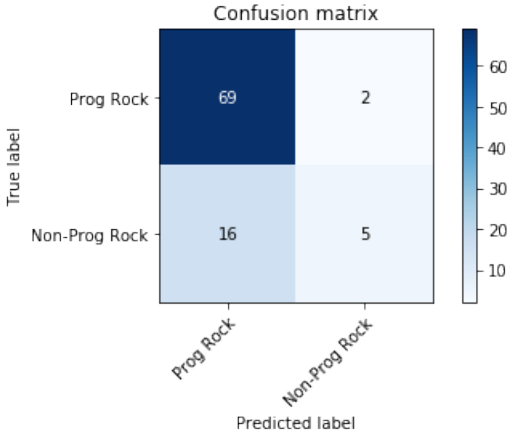

Figure 5: Validation set confusion matrix of the model predictions after training on the full training data set.

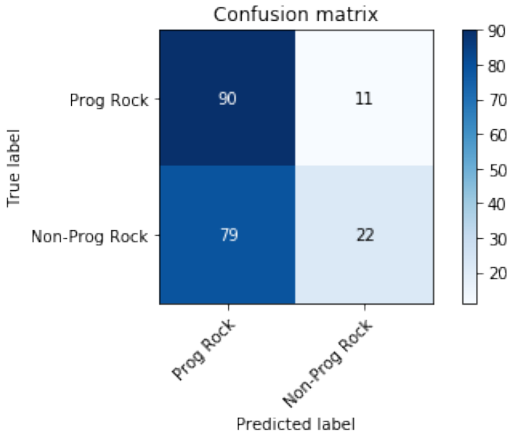

Figure 6: Test set confusion matrix of the model predictions after training on the full training data set with validation set stopping criteria.

3.1.2 Predictions on the djent audio files

Prediction	Song Name
Prog	CHIMP SPANNER - Bad Code
NonProg	BORN OF OSIRIS - Divergency
NonProg	AFTER THE BURIAL - A Wolf Amongst Ravens
NonProg	05. Physical Education
NonProg	Cloudkicker - Let yourself be huge
NonProg	Veil Of Maya - Punisher
NonProg	The Algorithm - Isometry
NonProg	Textures - Laments Of An Icarus
NonProg	SikTh - Hold My Finger
NonProg	PERIPHERY - Zyglrox
NonProg	MONUMENTS - I, The Creator
NonProg	Meshuggah- Soul Burn
NonProg	HEART OF A COWARD - Hollow
NonProg	Hacktivist - DECEIVE AND DEFY

3.1.3 Principal Component Analysis

Given the 230-dimensional input feature vector, we reduced dimensionality to 200, 100, 50, 20, 10, and 5 dimensions and then tested each reduced feature vector. We found that reducing the dimensionality from 230 to any of the previously mentioned dimensions had virtually no effect on the model accuracy. In Figure 7 the MFCC features do not have high separability in three dimensions which can imply low model accuracy.

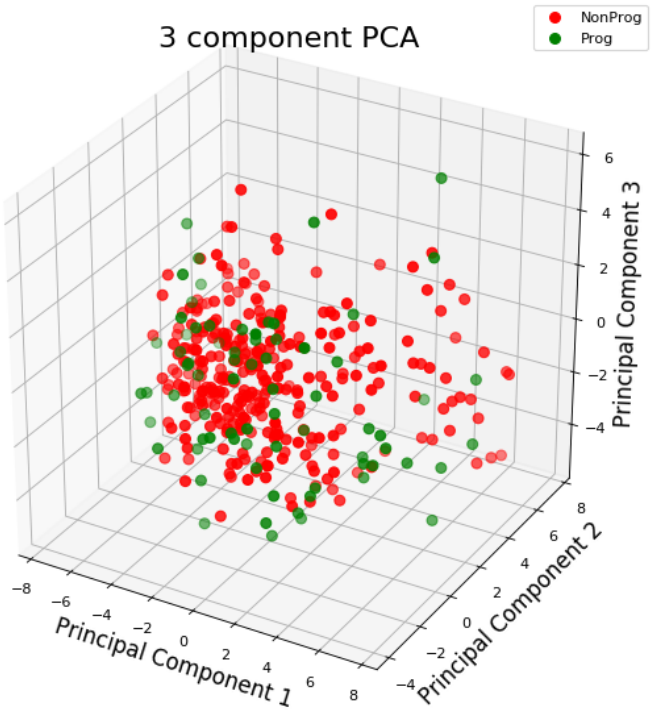

Figure 7: Training test set reduced to three dimensions using PCA.

3.2 Encoder-Decoder Medley

3.2.1 Basic Encoder-Decoder (ED)

We run the basic encoder-decoder architecture (ED) described previously on each set of data for comparison in Figure 8. We observe the expected trend of better performance on the training set than the test set. However, as a baseline, the encoder-decoder surpasses the Fully Connected network and achieves a reasonable 90% accuracy on the validation set, which corresponded to 60% accuracy on the test set.

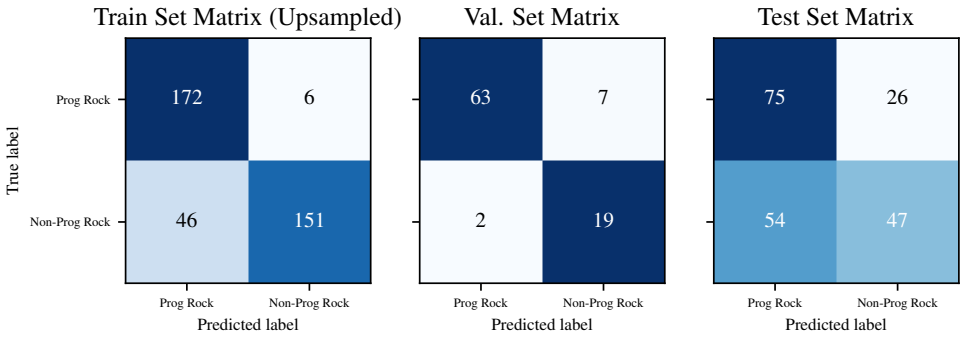

Figure 8: Confusion matrices for the basic Encoder-Decoder model across the upsampled training set, validation set, and test set.

3.2.2 Residual Encoder-Decoder (ED+Res)

Next, we add the residual convolutions described earlier in front of the recurrent layers of the encoder-decoder, and show the results in Figure 9. Interestingly, validation set accuracy increased to 91%, but test set accuracy dropped to 59%, so it difficult to determine whether the residual connections offer an advantage on their own. Empirically, the residual blocks only marginally improved performance despite tweaking the number of residual blocks and their width[2]

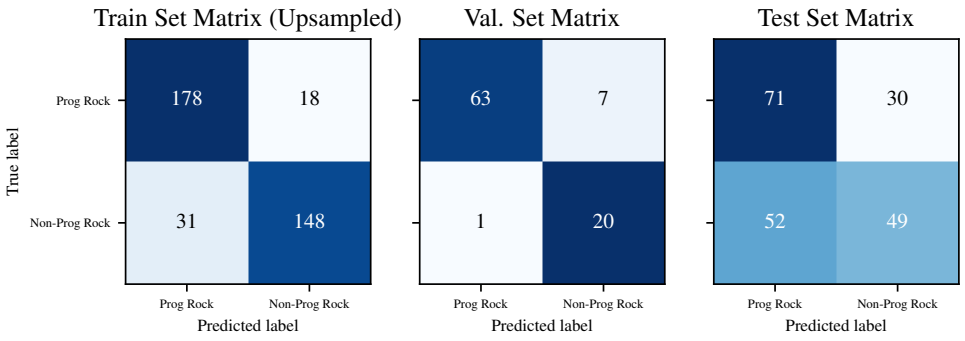

Figure 9: Confusion matrices for the basic Encoder-Decoder model with residual convolutions on the upsampled training set, validation set, and test set.

[2]Increasing residual block amount tended to drastically increase number of parameters of the model, leading to overfitting.

3.2.3 Residual Encoder-Decoder with Self-Attention (ED+Res+Att)

The self-attention mechanism is added to the output of the decoder, with one context hop and an attention dimensionality of 32 (chosen empirically). The confusion matrices of this new architecture are shown in Figure 10. Although training and validation performance are quite similar to before, the test set performance is improved dramatically. Validation accuracy of 91% and test set accuracy of 68% are achieved, which is an 8% increase over the ED baseline, and 9% increase over the ED+Res architecture. The test set results are summarized in Table 1, showing the stark increase in test accuracy with the introduction of self-attention.

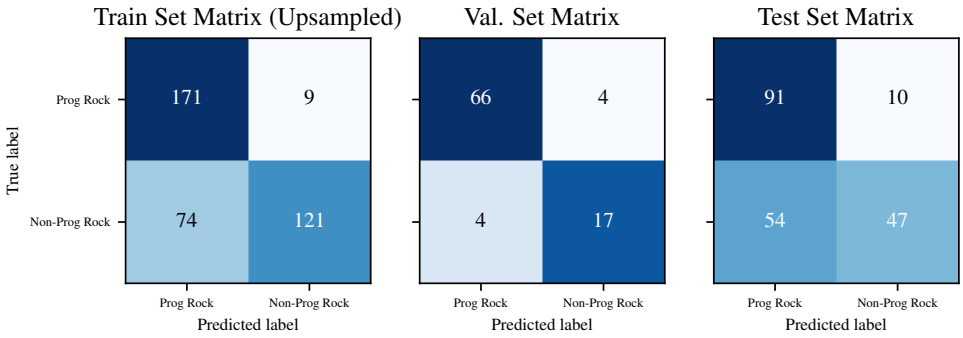

Figure 10: Confusion matrices for the Encoder-Decoder model with both residual convolutions and self-attention on the upsampled training set, validation set, and test set.

	Accuracy	Precision	Recall
ED	0.60	0.64	0.45
ED+Res	0.59	0.62	**0.49**
ED+Res+Att	**0.68**	**0.82**	0.46

Table 1: Test set metrics across each encoder-decoder variant. The combination of residual convolutions and self-attention yields a 8% boost in test set accuracy over the baseline.

3.2.4 Training & Stability Concerns

We briefly examine the training behavior of each encoder-decoder variant to better understand their stability and learning ability. Figure 11 shows both historic plots of training

accuracy and criterion value over training epochs. Notably, the ED and ED+Res models tend to achieve very high validation set accuracy early during training, but then flatline, whereas the ED+Res+Att model starts very low, and gradually gains accuracy as the parameters are adjusted. We remark that this behavior is observed despite the same batch size and optimization schedule (learning rate scheduling was used during long-term training, but it did not activate until 60 epochs) across each variant. Of course, the models are not trained on the validation samples, so we take this as a measure of added stability offered by the attention mechanism.

Further, we observe in the right subfigure that the NLL criterion falls much lower than the criterion for ED and ED+Res. Thus, gains are made in both test set performance and training effectiveness with the addition of self-attention.

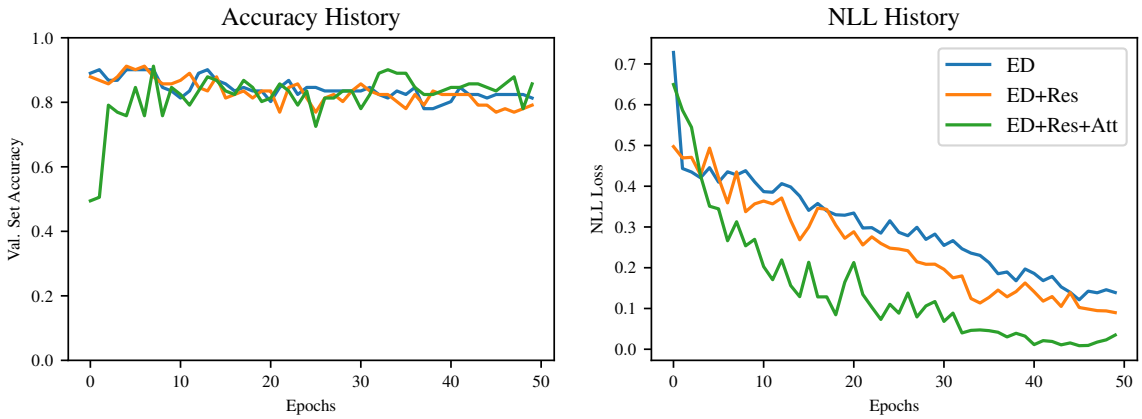

Figure 11: (left) Accuracy on the validation set across each encoder-decoder variation over the first 50 epochs, the typical convergence time. (right) Negative Log-Likelihood loss across epochs for each encoder-decoder variation.

4 Discussion & Closing Remarks

We observe that the encoder-decoder framework is able to improve on our early experiments with the fully connected neural network. The self-attention mechanism further improves on this in terms of test set accuracy. We close by discussing the qualitative measurements that the self-attention mechanism enables.

Figures 12, 13, and 14 show the attention weights for a selection of songs from the non-progressive rock, progressive rock, and djent categories of the test set, along with their

predicted class. We observe that the attention mechanism tends to weigh silence when a difficult song is encountered, as seen in the 'djent/Meshuggah - Soul Burn...' sample of Figure 14. Interestingly, the attention mechanism correctly identifies sections of music that would generally be considered as progressive rock, such as the flowing section of 'djent/05. Physical Education' in Figure 14. The intro of 'ProgRock/02. Desert Girl...' can be considered representative of the genre in Figure 13, which is correctly identified and highlighted by the attention mechanism.

To find pieces of non-progressive rock, it suffices to examine 'NonProgRock/into the void' in Figure 12, which is a heavily distorted vocal and guitar chorus with a constant rhythm. For an example of non-prog. classified as progressive rock, we can examine 'NonProgRock/03 I am a God (...' in the same Figure, which is highlighting a section of the song driven by a distorted synth melody and led by the lead vocal track.

Although the genre classification task between progressive rock and non-progressive rock is challenging even for humans, we find that it is possible to take an initial step using deep recurrent neural networks. With the aid of the self-attention mechanism, we can gain better insight into what may or may not be confused as progressive rock when the topic arises on online music forums.

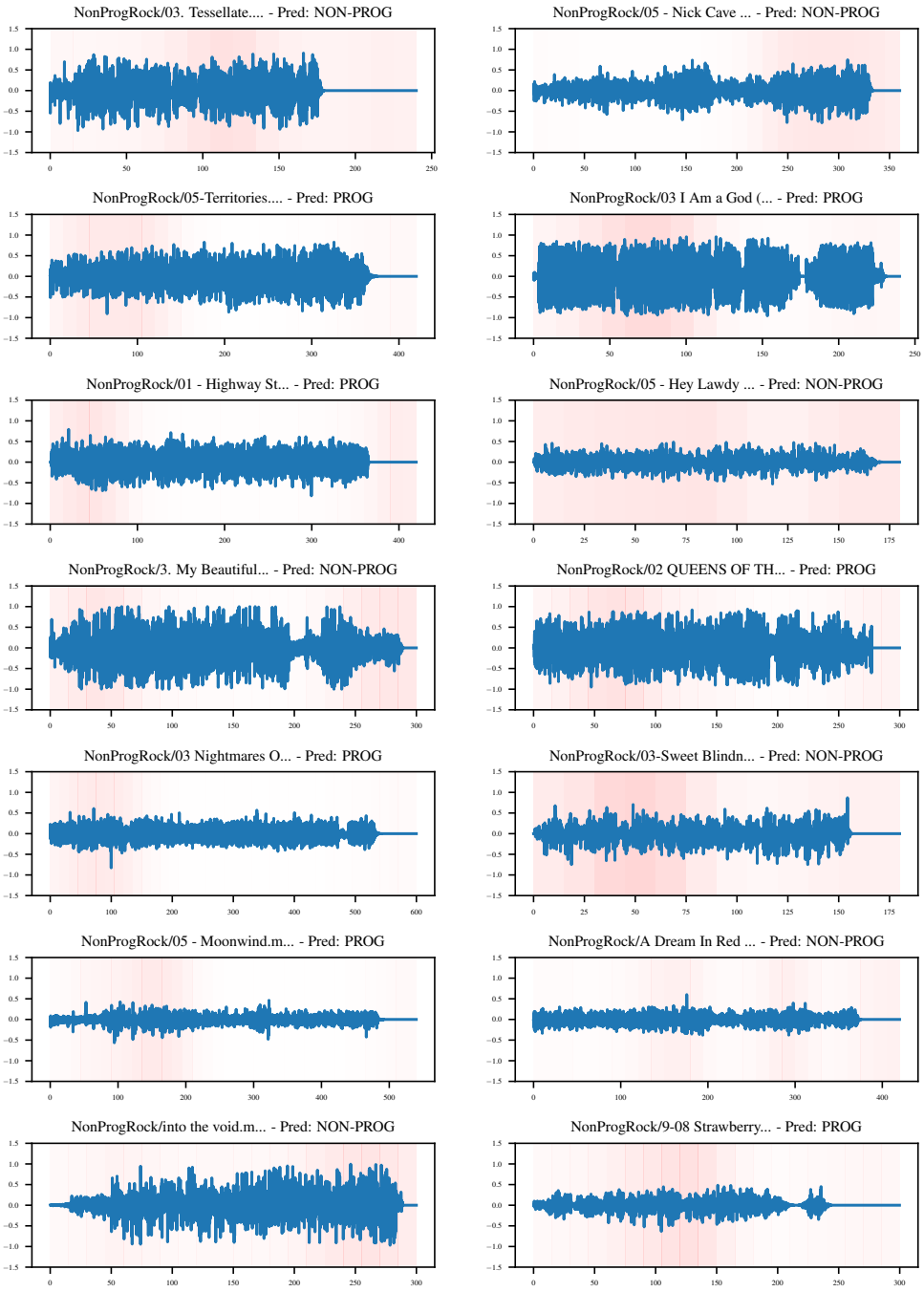

Figure 12: Self-Attention weights and classification of test set Non-Prog. Rock Songs.

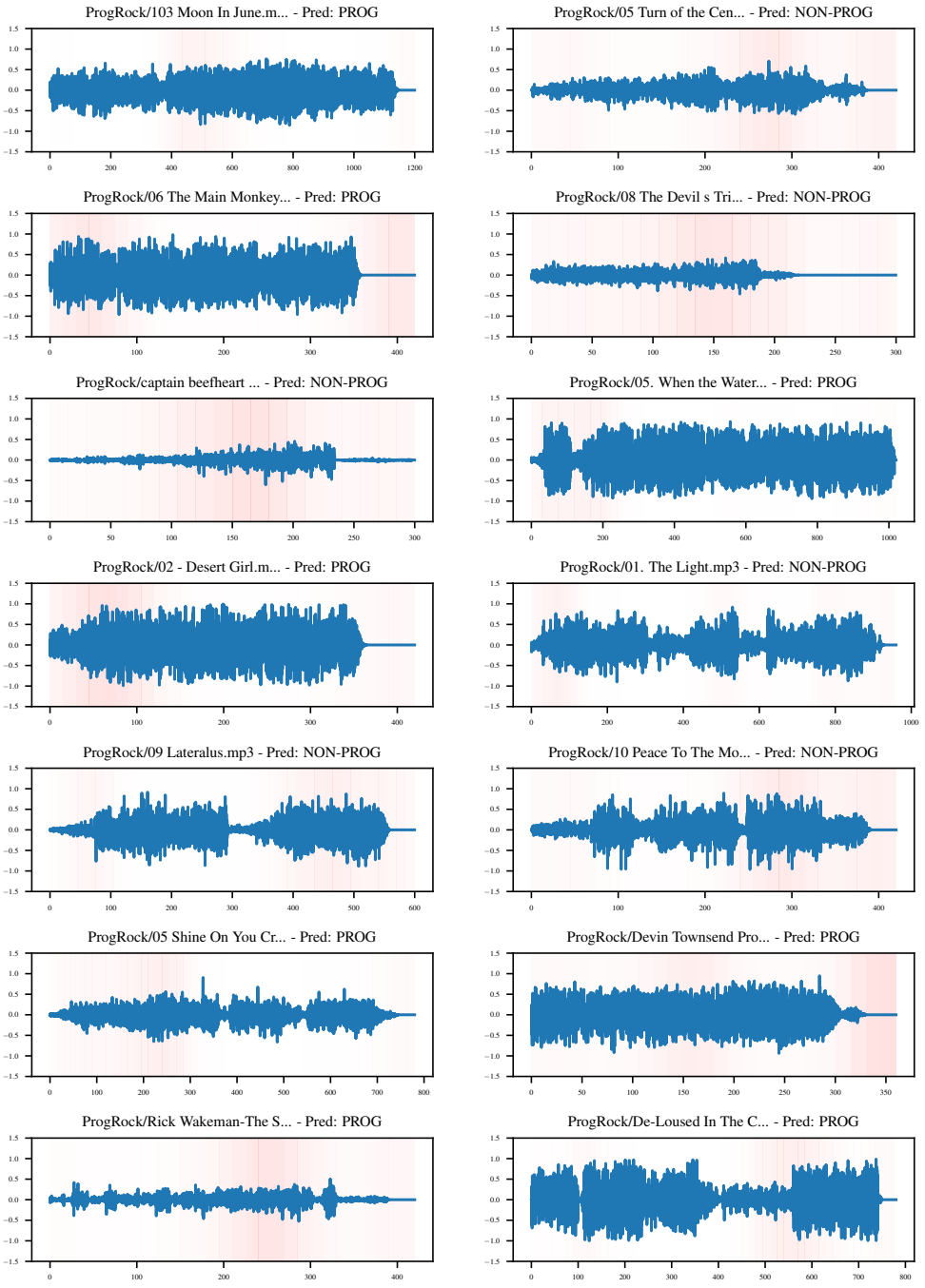

Figure 13: Self-Attention weights and classification of test set Prog. Rock Songs.

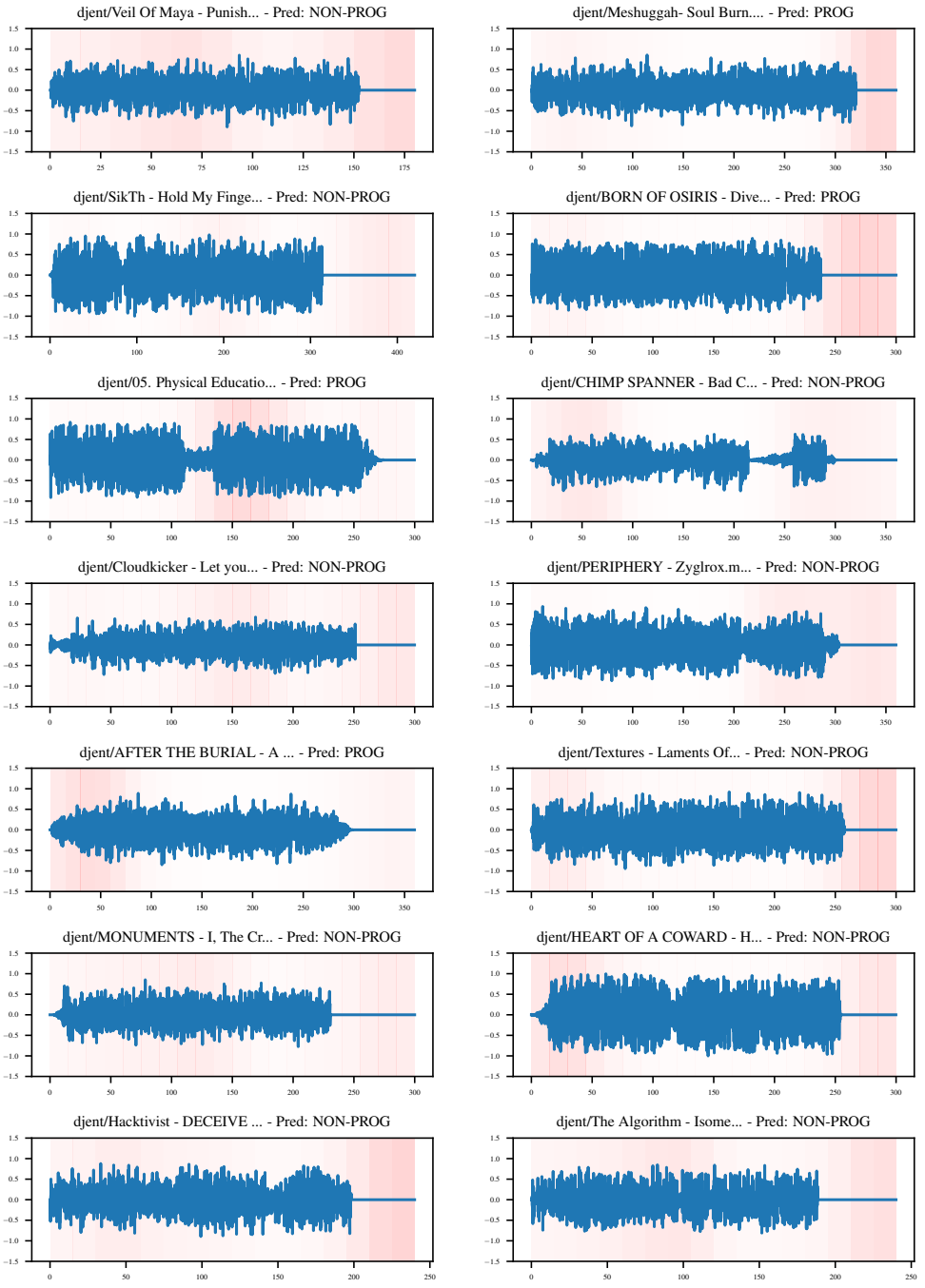

Figure 14: Self-Attention weights and classification of Djent Songs.

References

[1] Jurgen Appelo. *Management 3.0: leading Agile developers, developing Agile leaders*. Pearson Education, 2011.

[2] Stanford Encyclopedia. Supervenience.

[3] Tae-Ryang Kim. How does consciousness merely naturally supervene on the physical? 2012.

[4] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[5] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

[6] Oliver Nina, Washington Garcia, Scott Clouse, and Alper Yilmaz. MTLE: A multitask learning encoder of visual feature representations for video and movie description. *CoRR*, abs/1809.07257, 2018.

[7] Ramakanth Pasunuru and Mohit Bansal. Multi-task video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1273–1283, 2017.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *neural information processing systems*, pages 5998–6008, 2017.

[9] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence – video to text. In *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.