Is Spatial-Temporal-Audio Attention All You Need for Video Understanding?

David Charles Mason
Dartmouth College
david.c.dartmouth.edu

Abstract

Recent advancements in video understanding models have built upon the adaptation of the transformer architecture to specific modalities. However, current spatial-temporal transformers lack input from other types of data like audio. On the other hand, many models that fuse modalities have no built-in methods to modulate the influence of each data type. The human brain is inherently biased toward visual information. As a result, models seeking to mirror or improve on human performance on video classification tasks should use all relevant information types but avoid unneeded audio bias, for example, in the model's final output. This paper proposes a novel architecture to integrate multiple modality specific transformers using late fusion. It allows the representation of each contributing model to be changed by a hyperparameter. In the implementation, we leverage the spatial-temporal processing from the TimeSformer and the audio understanding from the Audio Spectrogram Transformer (AST). We remove the fully-connected layers, free model weights, and concatenate their outputs in a late fusion strategy, followed by fine-tuning with new fully connected layers. However, when the features are concatenated, we apply a scalar to the audio data to mirror the brain's prioritization of visual over audio data. We utilize a subset of the VGGSound dataset to validate our approach. This research has practical applications in video classification, video production, and content management. GitHub Repository : https://github.com/charliemason1/Is-Space-Time-Audio-Attention-All-You-Need.git

1 Introduction

Previous [1]scholarship has advanced the state of the art in video understanding by focusing on improving the transformer model [18]. Recent work has developed transformers that bifurcate self attention on both spatial and temporal data. These models [2] have achieved state-of-the-art performance using only self attention. However, these approaches are flawed due to the massive opportunity they are not taking advantage of. In cultural or humanities videos, there is a variety of sometimes unnoticeable objects or actions that take place in the shadows. Picture a video describing a Mardi Gras parade in New Orleans. The cameraman pans over a crowd of people dressed in purple, green, and gold. The viewer sees massive parades throwing beads to throngs of children. Then, the presenter concludes and says, "Alright, everybody, let's go eat some 'king cake.' Most models only focusing on video information would completely miss out on this very important part of the celebration. However, if audio and NLP processing was incorporated, a model could learn more nuanced information about the video, allowing it to pick up that 'eating king cake' is an extremely important part of the celebration.

[1] https://github.com/charliemason1/Is-Spatial-Temporal-Audio-Attention-All-You-Need-for-Video-Understanding/tree/main

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Returning to our example above, as the viewer watches the video, she is more likely to focus on visual features than the audio or possibly the subtitles that play along. Human brains are inherently visually biased. As You et al.[12] show, when the brain is presented with multiple modalities, it weights some more heavily than the other. Specifically, they tested the time difference between the brain's processing of visual and audio information. They found "that, the R-LRP onset to the auditory target was delayed about 91 ms when it was paired with a simultaneous presented visual target, compared to that when it was presented by itself. For the visual target, however, the R-LRP onset was comparable irrespective of whether it was paired with an auditory target or not." [12]. This phenomenon has been frequently observed and given the name 'the Colavita Visual Dominance Effect.' Researchers suggest that visual information contains a type of 'prepotent' stimuli. When asked to push a button corresponding to a variety of different uni-modality and multi-modality situations rapidly, participants in the study had no problem responding or with getting the task correct. However, they rarely pushed the auditory button, especially when audio and video were presented together.[16].

To understand why these interesting advancements in neuroscience contribute to designing multi-modal neural nets, we must present a more detailed introduction to current approaches. For a long time, the differing approaches of late vs early fusion have troubled and inspired the community. The most current models typically use an architecture known as a mixture of experts[10]. These models take advantage of the inductive biases present in modality specific transformers and connect these advances using a gating procedure based on the input of the user. This approach is currently the SOTA among late fusion. Another approach attempts to treat modalities as a unified type of data. Using a variety of different mathematical concepts such as fourier transforms, they reduce data types down to a format that the model can process equally. The data is encoded with modality specific features, which at the end of the day is the only thing that distinguishes them in the model's eyes. This architecture best represented by the Perceiver[8] is the SOTA for early fusion. In the end, late fusion has been shown to produce better results.

Therefore, building upon the research in computer science and neuroscience, we propose the $\alpha - transformer$. Our model seeks to follow the trend of late fusion approaches, but also wishes to replicate the priority of the brain's modality. About 30 percent of the brain's neurons are used for visual stimuli, with 8 percent for proprioception, and 3 percent for auditory information. Therefore, why should our model give other modalities equal computation and attention when NLP and visual tasks are so heavily prioritized by our own mental models. Thus, our logic is to fuse audio information at the end of an already excellent spatial-temporal model. In essence, we are fine-tuning the visual and temporal model on audio data. What that means is that the model will, based on the number of spatial-attention layers, tends to classify images based on its spatial and temporal contexts. However, the addition of the audio modality aims to improve that context and increase accuracy. To that end, we utilize a pretrained TimeSformer[2]and Audio Spectrogram Transformer[6].

2 Novelty and Related Works, Understanding Research in the Industry

2.1 The Origin of Transformers

This approach, modulating the features of different modalities before late fusion and fine-tuning, is a novel method to expanding the capabilities of spatial-temporal transformers. We cannot guarantee that this approach will improve the state of the art, however, we hypothesize that adding audio context will improve performance on events that lack sufficient spatial/temporal information, like the example given in the abstract.

However, this paper stands on the backs of giants and is heavily influenced by the scholarship that came before it. It is significant to understand the context that this paper sits in to better comprehend why it seeks to address the problem of multi-modality video understanding. In 2017, the transformer was introduced by Google[18] which allowed a significantly faster and more accurate way to process sequential information. Leveraging multi-head attention allowed a model to understand dependencies between data points that were very far away. For example, in the sentence

```
"My dog, Sam, went to the park the other day in a very happy mood, but a
        bigger tiger found him in a bush, and it scared him away."
```

attention allows the model to understand the relationship between ‘Sam’ and ‘him’ or ’tiger’ and ’it.’ Our approach is different from the original transformer because both the TimeSformer and the AST only use an encoder.

2.2 Adapting Transformers to Specific Modalities

After enough scholarship had informed the community on the transformer’s diverse uses, more contributors started to implement this generalizable architecture for other modalities. GPT-2 was a notable example of this phenomenon. The model predicts the next word in a given sentence using only a decoder. It exhibited zero shot learning, and showed that the full transformer architecture wasn’t needed[15]. In 2020, Dosovitsky et al. introduced the ViT (Vision Transformer)[3] which takes a 224x224 image and creates a set of patches. These typically 16x16 patches are then flattened, projected, and fed into a transformer for feature extraction. In 2021, scholars [19] designed a model which allowed a transformer to understand audio by processing spectrograms, which operated similarly. These models showed a transformer could be used to process image, text, and audio.

Our method does not use any NLP models. However, it differs from the ViT and Audio Transformer, because it implements better and more efficient versions of these models. The TimeSformer innovates on the Vision transformer by incorporating temporal attention[2]. This capability allows the model to process frames of video while maintaining a video’s structure with positional encodings. The AST[6] innovates upon other audio models because it uses Mel-spectrograms, which offer rich frequency and temporal data.

2.3 Multi-Modality Models as Fused Transformers

Given that video is a multi-modal data type, fusion of different models that specialized in a given data type was necessary for video understanding and processing. Over the past few years, a variety of new fusion techniques have been proposed. In 2018, Jiang et al. integrated a CNNs and LSTMs with a feature fusion network[9] to understand videos. More complicated methods followed. For example, The MMTM[4] includes slow modality fusion between each unimodal CNNs, reaching state-of-the-art among multimodal convolutional neural networks in 2020. These methods were extrapolated upon with transformers. For example, the V-MoE, a sparse vision transformer, eliminates the need to process every input by every parameter, cutting the computation by a half and reaching similar or better performance with the state-of-the-art. These trends were applied to perform a variety of astounding tasks, allowing scholars to classify a database of music videos based on how emotional they were[14] in 2021. Other fusion strategies have considered masking strategies. This year, Huang et al. [7] were able to integrate audio, temporal, and spatial information with their model MAViL through a masked learning approach which allowed simultaneous training for each separate modality.

The type of fusion method used impacts how the model learns the input data. Adding features means that a given model contributes as much influence as the others involved. Concatenating incorporates more complex relationships between datum, preserving all features as they enter the fully connected layers. The most innovative approach creates a learnable matrix which dictates how much influence each respective modality has on the final output. Our method chose to concatenate the TimeSformer and AST features, which in itself does not differ from previous approaches. However, the key contribution of this paper is to apply a series of transformations on the features before they are concatenated. For example, in one experiment, we take the square root of AST features before passing them to the TimeSformer.

2.4 Efficient Late-Fusion Pretrained Models

Finally, much scholarship has been focused on improving efficiency, mainly through architectures that use transformers for multiple modalities. The root of this effort can be seen in efforts to convert pre-trained fully connected layers in multiple modalities into RNN layers.[21]. In another direction, others sought to limit the architecture’s reliance on a given modality. In 2021, scholars proposed the Perceiver[8], which through cross-attention layers and fourier features, allowed a model to perform the same on information that was specific to the modality and raw input. For example, this model was able to reach the same accuracy on audio features extracted from a spectrogram as on raw audio input[8]. Finally, in 2023, authors were able to perfect fine-tuning using frozen image encoders, which allowed models to learn and generalize across modalities using existing pre-trained architectures[11]. Along with research in multi-modal transformers generalizing beyond classification[20] and the use of multi-modal transformers performing on few shot tasks [1], we can now attempt to fuse a different modality on top of a frozen pre-trained spatial-temporal architecture for video classification and generalization on other tasks.

Our approach, given in detail below, differs from these approaches through the feature transformations it performs on pre-concatenated data. Unlike the models above, which use pretrained models fused before their fully connected layers with frozen weights, our model changes the influence of connected models using scalar multiplication or mathematical transformations. Moreover, it realizes that models designed specifically for a given modality perform better than early fusion techniques like the Perceiver which do not take advantage of this characteristic.

3 Proposed Approach

The detailed approach for the completion of this project is as follows. First, we load an already trained TimeSformer[2], created on the HowTo100M[13] dataset with 64 frames, 448 spatial cropping, and a single clip coverage of 68.3 seconds. This model already achieved classification accuracy of 62.1 percent on the test set. Secondly, we will load a trained Audio Spectrogram Transformer[6] from its GitHub repository. This model was trained on AudioSet[5] with tstride of 10, fstride of 10, and weight averaging. It achieved 0.459 mAP on audioset. The reason for the selection of both model is the following:

1. The TimeSformer performed poorly on HowTo100M compared to other datasets (62.2% vs. 82.2% on Kinetics-600). We hypothesize that this result occurs because that particular dataset's audio is an explicit explanation of what is occurring. The lack of that context causes less than desirable accuracy.
2. The AST version achieved near perfect accuracy on AudioSet (0.459 mAP) which is far larger than the VGGSound dataset we use for fine-tuning. We hypothesize that an excellent audio model paired with an underperforming TimeSformer should significantly increase video classification accuracy.

After both models were loaded and processed, we remove their final fully connected layers. This action eliminates the two respective model's classification layers. Afterward, we freeze the weights on both models. Finally, we concatenate both models outputs using current fusion techniques, and pass those results into several new fully connected layers. we will then train the model on videos with both audio and visual data. Figures 22 demonstrates the architecture of the pre-trained models before they lose their fully connected layers.

The key contribution of this paper draws on extensive neuroscience research, which claims the human brain weights different modalities differently when receiving them simultaneously. The 'Colavita Visual Dominance Effect' suggests that a model simulating a human's ability to classify videos should include a bias toward visual over auditory information. To that effect, we define a set of transformations T = [sqrt(), pos-scalar, and neg-scalar()]. Before the features are concatenated, we perform the following transformations. We define the tensor of TimeSformer features (S) and tensor of AST features (A) as:

$$S = \begin{pmatrix} s_1 & \cdots & s_n \\ \vdots & \ldots & \vdots \\ s_m & \cdots & s_n \end{pmatrix} \quad A = \begin{pmatrix} a_1 & \cdots & a_n \\ \vdots & \ldots & \vdots \\ a_m & \cdots & a_n \end{pmatrix}$$

Before we perform $cat(S, A)$,, we multiply one of the sets S or A by $t_i \in T$. Which gives us

$$\texttt{combined\_features} = cat(S \cdot t_i \ \forall \ t_i \in T, A)$$

or

$$\texttt{combined\_features} = cat(S, A \cdot t_i \ \forall \ t_i \in T)$$

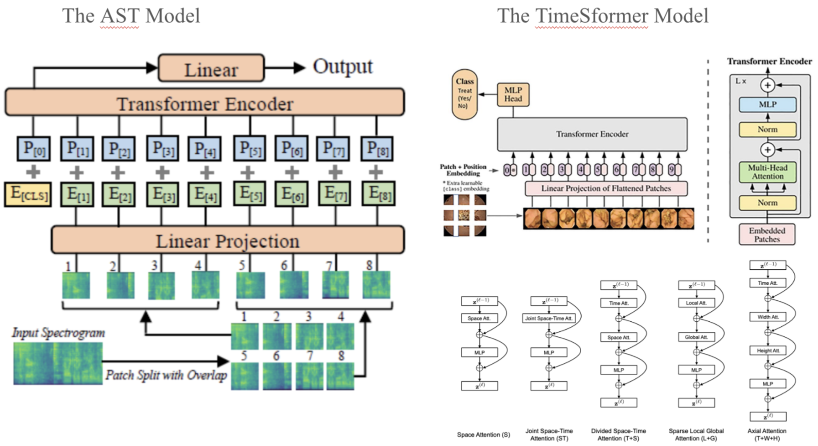

Figure 1: The AST (pictured on the left) processes a series of spectrograms, linearly projects the batch and feeds into an encoder. The TimeSformer (pictured on the top right) processes images similarly but also incorporates position encoding and temporal attention heads (pictured bottom right.)

These combined features serve as inputs to the fully connected layers, fine-tuned on the VGGSound data subset. Due to computational and time limitations, the training procedure involved randomly sampling 5% of the VGG Sound database. The selection of the subset was random and the further partitioning into valuation, training, and testing sets was also random. The results are benchmarked against the MAViL paper, as it also runs on VGGSound. Secondly, because alpha builds off the TimeSformer, there is a natural ablation study present. So, we compare our results to those obtained in their original paper[2]. The process described above can be seen in Figure 2 below.

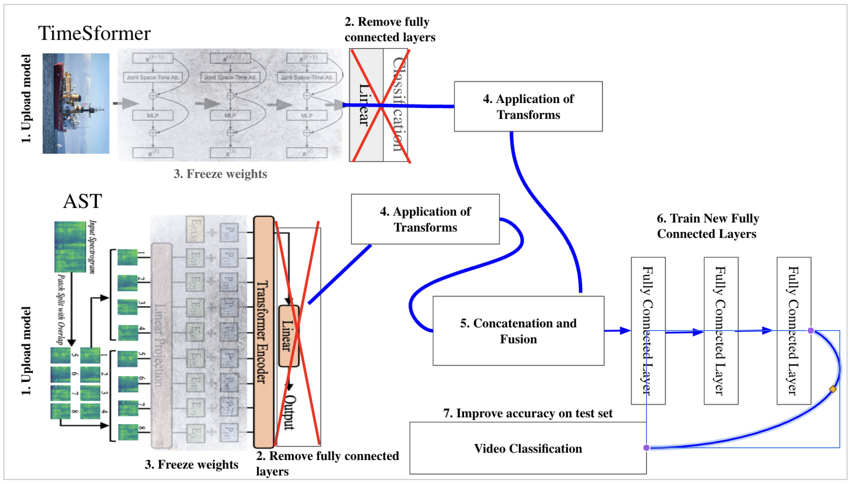
TimeSformer
1. Upload model
2. Remove fully connected layers
Linear
Classification
3. Freeze weights
4. Application of Transforms
AST
1. Upload model
Transformer Encoder
Linear
Output
4. Application of Transforms
3. Freeze weights
2. Remove fully connected layers
5. Concatenation and Fusion
6. Train New Fully Connected Layers
Fully Connected Layer
Fully Connected Layer
Fully Connected Layer
7. Improve accuracy on test set
Video Classification

Figure 2: The TimeSformer and AST are uploaded, stripped of their previous fully connected layers, and the remaining weights are frozen. The outputs of the remaining layers are joined in late fusion and then fed into new fully connected layers. The model is then fine-tuned on a dataset made up of both audio and visual videos.

4 Experimental Results

4.1 Results

To test our hypothesis about mixing varying degrees of pre-trained model influence, we trained six models. We used the same dataset for each, which was a random subsampling of VGGSound.

There were 92 classes, which averaged around one training example per class, given that the test, validation, and training subsets had 20, 10, and 100 videos in them respectively. The highest accuracy achieved was **22%** by experiments 2 and 6. In both cases, the highest results arose from increasing the influence of the TimeSformer. In experiment 2, the TimeSformer features were multiplied by a scalar value of 1.5. This result was the best because it also came with the lowest loss. In Experiment 6, the AST features were multiplied by the scalar 0.75. The results from the experiments are found in the table below. 1. We have also included a figure detailing the loss 3. As the reader can see, our validation loss increases over the training procedure.

Experiment Number	Transformed Model	Type of Transform	Test Accuracy	Loss
1	TimeSformer	Sqrt()	11%	7.29
2	TimeSformer	pos_scalar()	**22%**	6.98
3	TimeSformer	neg_scalar()	11%	7.39
4	AST	Sqrt()	16%	7.30
5	AST	pos_scalar()	11%	7.66
6	AST	neg_scalar()	**22%**	7.52

Table 1: The test accuracy based on which model was changed before concatenation and which formula was applied.

4.2 Training Procedure

On the training process, given that we are merely fine-tuning the last three fully connected layers, we believe that 25 epochs is sufficient for convergence. We tried another 25 with a lower learning rate after the first 25, but found that convergence had already been achieved. Moreover, we used a Cross Entropy Loss function for training on all six models due to its intended use as a video classifier. An Adam's Optimizer was chosen based on its implementation in the TimeSformer model[2].

The training data was randomly sampled from VGGSound, a dataset which includes over 200,000 videos with hundreds of thousands of hours of audio. Once we had sampled roughly 5% of the dataset, we further broke the dataset down into valuation, training, and testing subdirectories. Benchmarking against MAViL, the original TimeSformer, and the AST, we received significantly lower results. However, we hypothesize the reason was due to the incredibly small dataset used. Moreover, the diversity of the dataset was an additional issue. Given that the model only sees approximately one example for each class, the low accuracy is not surprising.

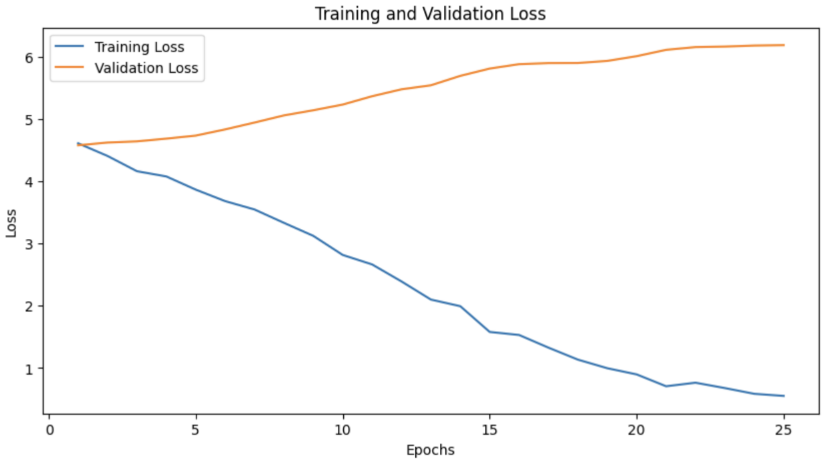

Figure 3: This figure was taken from Experiment 2 which had the lowest loss.

5 Discussion

This paper has proposed a novel approach to fusing transformers specializing in multiple modalities. Although it does not demonstrate SOTA performance or compare closely with other benchmarked models, further research with more time and computational research could expand this model's techniques and capabilities. Training on other datasets such as Kinetics-600, UCF-101, and the full VGGSound would undoubtedly raise accuracy and performance. This paper also maintains that models wishing to achieve or improve upon human results on tasks should be based on neuroscience. Our use of research proving that the brain weights visual data more heavily than audio allowed for the hypothesis that a neural net with greater emphasis on visual information would perform better. In that regard, despite low accuracy, our hypothesis was proven correct. In the two experiments which had the highest results, visual information was either amplified or audio information was impaired. In further research, we plan to train an improved model on more datasets and try other functions that greater heighten the visual features and further denigrate the audio. Although audio information is still necessary, the human brain dedicates merely 3%; in contrast, the visual cortex is almost 50%.[17]

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, pages 1–6. IEEE, 2020.

[5] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[6] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[7] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[9] Yu-Gang Jiang, Zuxuan Wu, Jinhui Tang, Zechao Li, Xiangyang Xue, and Shih-Fu Chang. Modeling multimodal clues in a hybrid deep learning framework for video classification. *IEEE Transactions on Multimedia*, 20(11):3137–3147, 2018.

[10] Qianyue Jin. Mixture of experts for image classification. 2021.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[12] You Li, Mingxin Liu, Wei Zhang, Sai Huang, Bao Zhang, Xingzhou Liu, and Qi Chen. Neurophysiological correlates of visual dominance: A lateralized readiness potential investigation. *Frontiers in Psychology*, 8:303, 2017.

[13] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[14] Yagya Raj Pandeya and Joonwhoan Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2):2887–2905, 2021.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[16] Charles Spence, Cesare Parise, and Y. C. Chen. The colavita visual dominance effect. In M. M. Murray and M. T. Wallace, editors, *The Neural Bases of Multisensory Processes*, chapter 27. CRC Press/Taylor & Francis, Boca Raton (FL), 2012.

[17] v56knVrwQgWk. What is the difference between the processing of visual information and auditory information? *Psychology & Neuroscience Stack Exchange*, 2021.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] Prateek Verma and Jonathan Berger. Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions. *arXiv preprint arXiv:2105.00335*, 2021.

[20] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[21] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Multilayer and multimodal fusion of deep neural networks for video classification. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 978–987, 2016.