Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Grósz, Tamás; Virkkunen, Anja; Porjazovski, Dejan; Kurimo, Mikko

Discovering Relevant Sub-spaces of BERT, Wav2Vec 2.0, ELECTRA and ViT Embeddings for Humor and Mimicked Emotion Recognition with Integrated Gradients

Published in:
MuSe '23: Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation

DOI:
10.1145/3606039.3613102

Published: 02/11/2023

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Grósz, T., Virkkunen, A., Porjazovski, D., & Kurimo, M. (2023). Discovering Relevant Sub-spaces of BERT, Wav2Vec 2.0, ELECTRA and ViT Embeddings for Humor and Mimicked Emotion Recognition with Integrated Gradients. In *MuSe '23: Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation* (pp. 27-34). ACM. https://doi.org/10.1145/3606039.3613102

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Discovering Relevant Sub-spaces of BERT, Wav2Vec 2.0, ELECTRA and ViT Embeddings for Humor and Mimicked Emotion Recognition with Integrated Gradients

Tamás Grósz
tamas.grosz@aalto.fi
Aalto University
Espoo, Finland

Anja Virkkunen
anja.virkkunen@aalto.fi
Aalto University
Espoo, Finland

Dejan Porjazovski
dejan.porjazovski@aalto.fi
Aalto University
Espoo, Finland

Mikko Kurimo
mikko.kurimo@aalto.fi
Aalto University
Espoo, Finland

ABSTRACT

Large-scale, pre-trained models revolutionized the field of sentiment analysis and enabled multimodal systems to be quickly developed. In this paper, we address two challenges posed by the **Mu**ltimodal **Se**ntiment Analysis (MuSe) 2023 competition by focusing on automatically detecting cross-cultural humor and predicting three continuous emotion targets from user-generated videos. Multiple methods in the literature already demonstrate the importance of embedded features generated by popular pre-trained neural solutions. Based on their success, we can assume that the embedded space consists of several sub-spaces relevant to different tasks. Our aim is to automatically identify the task-specific sub-spaces of various embeddings by interpreting the baseline neural models. Once the relevant dimensions are located, we train a new model using only those features, which leads to similar or slightly better results with a considerably smaller and faster model. The best Humor Detection model using only the relevant sub-space of audio embeddings contained approximately 54% fewer parameters than the one processing the whole encoded vector, required 48% less time to be trained and even outperformed the larger model. Our empirical results validate that, indeed, only a portion of the embedding space is needed to achieve good performance. Our solution could be considered a novel form of knowledge distillation, which enables new ways of transferring knowledge from one model into another.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Knowledge representation and reasoning**; *Natural language processing*; *Computer vision.*

KEYWORDS

XAI, Model Interpretation, Feature Selection, Multimodal Sentiment Analysis, Affective Computing

This work is licensed under a Creative Commons Attribution International 4.0 License.

MuSe '23, October 29, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0270-9/23/10.
https://doi.org/10.1145/3606039.3613102

ACM Reference Format:
Tamás Grósz, Anja Virkkunen, Dejan Porjazovski, and Mikko Kurimo. 2023. Discovering Relevant Sub-spaces of BERT, Wav2Vec 2.0, ELECTRA and ViT Embeddings for Humor and Mimicked Emotion Recognition with Integrated Gradients. In *Proceedings of the 4th Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation (MuSe '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3606039.3613102

1 INTRODUCTION

In recent years, large self-supervised models have become extremely popular as they demonstrated outstanding accuracies on numerous tasks [18, 21, 33]. Their strength lies in the unsupervised pre-training step, during which they process large quantities of unlabelled data and learn to extract meaningful latent features. These so-called neural embeddings are then used as features for another model or are directly connected to a new output layer, enabling the supervised fine-tuning of the whole model. Models belonging to the BERT family are considered state-of-the-art in text processing [11], wav2vec 2.0 [3] is one of the most popular multilingual solutions in case of audio input, while Vision Transformers (ViT) [12] and facial action units FAUs [13, 34] are regularly used to process videos.

Like in many other fields, the pre-trained solutions were quickly adapted for sentiment analysis and showed promising results [1, 6, 15, 25]. This is reinforced by the fact that all baseline solutions in the Multimodal Sentiment Analysis Challenge (MuSe) 2023 are built upon a uni-modal pre-trained network [2, 8]. The competition introduces multiple interesting tasks, and here we focus on the MuSe-Mimic and the MuSe-Humor sub-challenges.

The MuSe-Mimic utilizes a large-scale multimodal database consisting of over 18,000 video recordings (approximately 20 hours in total) made by users mimicking three emotions: Approval, Disappointment and Uncertainty. The task is to predict the annotated continuous emotional values using the video, audio and transcript. Further details about the dataset can be found in [8].

In the MuSe-Humor sub-challenge, competitors are asked to build a cross-cultural humor detector by developing models using 10 hours of recordings from 10 different German football coaches. The training and development data consists of recordings from the Passau Spontaneous Football Coach Humor (Passau-SFCH) corpus [9]. The detectors are tested with 6.5 hours of recordings

of English Premier League press conferences held by 6 different coaches. The main challenge of this task is the cross-cultural and cross-lingual nature of the sets, as models are trained with German videos are tested with English recordings.

Despite the immense popularity of pre-trained models, we know very little about the embedding vector spaces produced by them. Most current research focuses on either identifying the best pre-trained model for a given task [15] or the optimal way of fine-tuning them [26]. In [5], authors investigate self-supervised ViT features, reporting the emergence of explicit information about the semantic segmentation of an image, which is not observable as clearly with supervised ViTs. In the text domain, some effort has already been dedicated to investigating BERT-based word embeddings. In [31], authors focus on the locations of various words and their relative positions. Unfortunately, the latent embedding vector is considered an atomic unit by all studies, and its subatomic components are ignored.

Based on the observation that the same pre-trained embedding spaces can be used for numerous tasks, in this study, we hypothesize that the latent vectors consist of several sub-spaces relevant to different tasks. In order to identify such sub-spaces, we investigate the baseline models by utilizing a local interpretability method called Integrated Gradients (IG) [27] to calculate the attributions of each individual dimension. Once the input attributions are estimated, we can select a subvector of the original embeddings by filtering the dimensions that had the highest contribution towards the output of the baseline model. Lastly, we validate that the relevant part of the latent space is filtered by training another model (having the same architecture as the baseline one), but using only the selected features. Our empirical results demonstrate that using this technique, we get comparable results on the Mimic task, and even observe minor improvements on the Humor sub-challenge, by removing irrelevant input features.

Besides the model interpretation, we also address the cross-cultural nature of the Humor task by employing a translation-based monolingual solution. The superiority of monolingual solutions compared to multilingual ones has been reported in several papers already [28, 30]. Motivated by this, we propose a pipeline solution by first translating the German text into English and then using a monolingual BERT. Our empirical results showcase that the translation-base monolingual model outperforms the baseline model using a multilingual BERT.

Lastly, we also demonstrate the great benefits of fine-tuning a wav2vec 2.0 on the Mimic task with an appropriate loss function compared to only using it as a static pre-trained feature encoder.

2 BASELINE MODELS

Our solution heavily relies on the baseline models introduced in [8], so we explain them briefly in this section. All approaches can be categorized as Encoder-Decoder models [7], where the encoder extracts meaningful latent features from audio, video or text while the Decoder component learns to predict the annotated labels (emotion values or the humor tags).

The competition organizers extensively studied 8 different encoders [8]; here, we only focus on the best ones per modality. From the baseline results, presented in [8], wav2vec 2.0 [3] is clearly the best solution in the case of audio input for both tasks. This model has already demonstrated excellent results on various paralinguistic tasks [15], which motivated us to investigate it further.

In the text domain, two encoders performed exceptionally well, a multilingual BERT [11] in the case of Humor and ELECTRA [10] for the Mimic sub-challenge. Both models employ Transformers as their main component and are pre-trained on large quantities of text data. In our experiments, we interpreted both models to determine the task-relevant sub-spaces of their embeddings.

Looking at the video modality, we can see that two encoders perform equally. On the humor detection task, ViT [12] is the clear winner, but on mimicked emotion prediction, facial action units (FAU) [13] proved to be slightly better. While ViT is quite similar to other models, consisting of a large pre-trained Transformer, FAU is an outlier in this sense. The FAU Encoder is pre-trained to automatically estimate the activation of 20 different facial muscles via SVC models provided by the Py-Feat[1] library. This also means that FAU dimensions are easily interpretable, which we will leverage in our experiments.

On the Decoder side, a simple architecture is used, which consists of a GRU-RNN followed by two feed-forward layers. The GRU-RNN is meant to deal with the sequential nature of the encoder outputs, and only the final hidden representation is passed to the feed-forward layers. For the Humor task a binary cross-entropy loss function is used to optimize the Decoder, while for the Mimic task, Mean Squared Error (MSE) loss is utilized. Further technical details about the Encoders and Decoders can be found in the baseline paper [8] or in the released codes[2].

3 FIND THE TASK-RELEVANT SUB-SPACE OF THE EMBEDDINGS

Feature selection is a well-established and researched area consisting of many solutions with different advantages and disadvantages [17]. The main goal of all solutions is to separate the significant features from the irrelevant ones. Reducing the input to only the relevant dimensions also decreases the computational overload and sometimes improves the predictive performance of the models.

Our primary motivation is the so-called forward sequential selection (FSS) algorithm, which iteratively selects the most informative feature and adds it to the set of optimal features. FSS relies on the fact that a new model can be trained in each iteration using the already selected features and the potential candidates. Typically, simple models like Support Vector Machine (SVM) are used during the selection process to avoid the excessive computational costs of training a vast amount of models to find the optimal sub-space of the input. Naturally, the features discovered by FSS are most relevant to SVM models, highlighting the importance of model-specific feature selection compared to model-agonistic solutions, which aim to determine generally relevant inputs.

Motivated by FSS, our goal was to develop a model-specific feature selection for Deep Neural Networks. Unfortunately, a direct adaptation of FSS to DNNs is not feasible due to the considerable training time of our Decoders compared to simple SVMs. Luckily,

[1] https://py-feat.org/
[2] https://github.com/EIHW/MuSe-2023

modern tools of Explainable Artificial intelligence (XAI) [29] offered us a way to peak into the baseline Decoders. Specifically, we employed a post-hoc model interpretation technique to separate the feature space into necessary and irrelevant sub-spaces.

3.1 Interpreting Neural Models via Integrated Gradients

Integrated Gradients (IG) [27] offers a simple and reliable way of interpreting already trained DNNs without altering the model, thus ensuring that the interpretation does not hinder its performance. IG, like many other alternatives, is a gradient-based solution, and it generates the so-called input attributions, which estimate the contributions of each individual input towards the final output of the model. This also indicates that IG is an instance-based interpretation technique, meaning that it requires a dataset to produce the interpretations for human observers.

To calculate the input attributions with IG, we need to have some input data x, and the corresponding baselines of inputs x', which in our experiments are set to be vectors containing zeros. Next, we define $\gamma : [0, 1] \rightarrow R^N$ as a smooth path between the baseline and the actual input ($\gamma(0) = x'$ and $\gamma(1) = x$). Then the input attributions of model Θ are defined as

$$IG(x, \Theta, \gamma)_i = \int_{\alpha=0}^{1} \frac{\partial \Theta(\gamma(\alpha))}{\partial \gamma(\alpha)_i} \frac{\partial \gamma(\alpha)_i}{\partial \alpha} \, d\alpha, \tag{1}$$

where $\frac{\partial \Theta(x)}{\partial x_i}$ is the gradient of the models output along the i^{th} dimension of the input x. In practice, calculating the exact value of the integral is intractable, but estimating it via a summation can be done efficiently based on the Riemman approximation of the integral;

$$IG(x, \Theta)_i = \frac{(x_i - x_i')}{m} * \sum_{k=1}^{m} \frac{\partial \Theta(x' + \frac{k}{m} * (x - x'))}{\partial x_i}, \tag{2}$$

where we approximate the path between x' and x at m discrete positions. For more technical details about IG, we refer the interested reader to [27] and [19].

We should note that current interpretation methods are generally considered fragile [14], i.e. easily influenced by adversarial samples and perturbations. In [20], the authors argue the importance of having a concrete definition of interpretation before evaluating the faithfulness of an interpretation. In this work, we chose to test the reliability of interpretations via probing [4], i.e., training a new Decoder using only the most influential features according to the interpretation.

3.2 Filtering relevant features using input attributions

Selecting relevant features based on an already trained model is relatively underrepresented in the literature. The previous existing works mostly focus on selecting inputs based on the input weights of models [22]. In contrast, we argue that input attributions calculated by IG are applicable to locating optimal input sub-spaces. We are only aware of one work where something similar was attempted; in [16], the authors use the gradients of an ensemble's output to select a subset of relevant spectral features. Unlike [16], we use a more sophisticated solution to estimate the importance of features and investigate the distribution of attributions in more detail.

Once the input attributions are estimated, we need to aggregate the individual attribution vectors and decide which features are relevant. The aggregation can be easily done by calculating the mean and standard deviations for each input dimension. Normally, only the mean values are used, but to ensure that features that are only influential in a few cases do not get ignored, we also investigated the deviations of the attributions. Lastly, the irrelevant features can be filtered out by a simple threshold applied to the aggregated statistics.

4 TASK SPECIFIC SOLUTIONS

In this section, we explain some additional task-specific techniques that we employed during the competition in order to achieve the best results.

4.1 Translation-based BERT solution for the Humor challenge

In the Humor task, the main challenge was posed by the cross-lingual nature of the data (German training and development sets, English test data). While the baseline solution addresses it by using a multilingual BERT (mBERT) as an encoder, we wanted to explore a translation-based alternative. In the literature, there are conflicting experimental results concerning the superiority of multilingual and monoligual BERTs. In [32], authors report that even low-resource languages are well covered by mBERT and monolingual results are inferior to the multilingual ones. An opposing view can be found in [28, 30], where monolingual models are shown to be superior to mBERT.

Motivated by these observations and the fact that transcripts of press conferences held by football coaches are probably not well represented in any BERT model, we propose a translation-based pipeline. In the first step, we selected a well-performing German-to-English neural translation model[3] [23]. Next, we extracted the sentence embeddings of a monolingual BERT trained only on English data in the same way as explained in [8]. Lastly, we trained the Decoders using the same hyperparameters and architectures as in the case of mBERT, only changing the features to the monolingual BERT embedding. These experiments demonstrated that despite the occasional mistakes of the translation system, humor can be translated, and monolingual embeddings are better for this task.

4.2 Fine-tuning wav2vec 2.0 using correlation loss

Observing the superior results provided by the wav2vec 2.0 models prompted us to investigate them further. In other articles, it is reported that fine-tuning the whole model, or at least the Transformer part leads to better performance than simply using the wav2vec 2.0 as a feature encoder [15]. The disadvantage of fine-tuning is the increased computational cost, even if we have access to modern GPUs. Due to this fact, we only experimented with fine-tuning on

[3] https://huggingface.co/facebook/wmt19-de-en

the Mimic task, where wav2vec 2.0 baseline solution proved to be by far the best.

As an additional adjustment, we changed the loss function and maximized the batch-level Pearson correlation;

$$PearsonLoss(y, \hat{y}) = 1 - \frac{\sum_i (y_i - \overline{y_i}) * (\hat{y}_i - \overline{\hat{y}_i})}{\sqrt{\sum_i (y_i - \overline{y_i})^2 * (\hat{y}_i - \overline{\hat{y}_i})^2}} \quad (3)$$

where y contains the ground truth values for the batch and $\hat{y}$ are the predictions made by the model. To ensure the stability of the loss, we had to use a relatively large batch size of 32 which turned out to be big enough to ensure that the model converges without running out of memory. Due to computational restrictions, we only managed to fine-tune one model, which we could not evaluate during the competition phase, but only in the post-evaluation phase.

5 EXPERIMENTAL SETUP

To reproduce the official baselines, we used the source codes released by the organizers[4]. The post-hoc interpretation of the baseline Decoders was performed with the use of the Captum [19] toolkit. As baseline inputs for the IG method to be used in Equation 2, we employed vectors filled with zeros. To estimate the input attributions, 50 steps ($m = 50$ in Equation 2) were taken by the approximation method. Lastly, we always multiplied the gradients with the inputs to ensure that the sign and strength of the input are taken into account.

For the Humor task, only a single output neuron is used, thus IG is estimated via its gradient w.r.t. the inputs. In the Mimic task, three outputs are available (one for each emotion) so first, the IG attributions are calculated for each output separately. In the aggregation step, we average the attributions of each input towards all three outputs before thresholding. After the relevant features are selected, new Decoders are trained using the same architecture and hyperparameters as the baseline models, only adjusting the first layer of the GRU-RNN to accommodate the new, reduced input.

In the Humor Detection challenge, the official evaluation metric is the Area Under the ROC Curve (AUC), which evaluates the goodness of the predicted probabilities of different time segments containing humorous remarks. In the Mimic task, continuous values for three emotions (approval, disappointment and uncertainty) need to be estimated. To account for this, the organizers choose to use the average Pearson's correlation coefficient across the three targets.

6 RESULTS

6.1 Humor

While reproducing the official baseline Decoders we observed insignificant changes compared to the results reported in [8] mainly due to the change of computational resources (different GPU). Once all the Decoders were ready, we selected the best ones per modality. In the case of the Humor task, the chosen models were trained on wav2vec 2.0, BERT and ViT embeddings.

Using IG, as explained in section 3, we estimated the importance of each dimension of the embedding spaces. Figure 1 depicts the mean attributions calculated by using the training data. For better visualization, we re-ordered the dimension based on their average attribution value. The first interesting observation was the sparsity of the wav2vec 2.0 embeddings; apparently, more than half of the input vectors do not contribute towards the Decoder's output. In the case of ViT and BERT, almost all input dimensions had some contribution, but an exponential distribution of importance can still be observed.

Next, to ensure we retain vital, rare features, which have high attribution only in a few cases, we investigated the distributions of the attribution values. Figure 2 confirms a strong correlation between the average and the deviation of attributions, implying

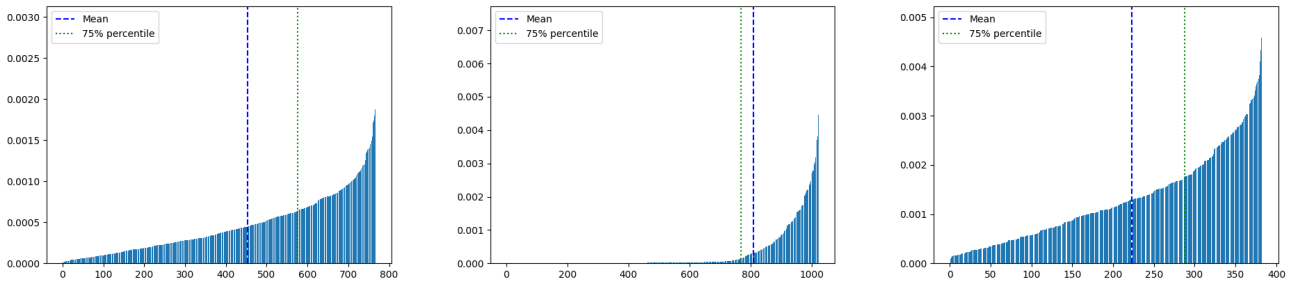
[4]https://github.com/EIHW/MuSe-2023

Table 1: Feature selection results on the development set of the Humor task using different thresholds. In each row, the first number is the best AUC-Score, together with the mean and standard deviations across the 5 seeds in parenthesis.

Features	Threshold	Num. Feats	Dev. AUC
Wav2Vec2.0	mean	216/1024	.8321 (.8281 ± .0027)
	75%	257/1024	.8343 (.8297 ± .0031)
ViT	mean	162/384	.7697 (.7577 ± .0088)
	75%	97/384	.7457 (.7261 ± .0135)
BERT	mean	288/768	.8065 (.7836 ± .0216)
	75%	193/768	.7951 (.7772 ± .0158)

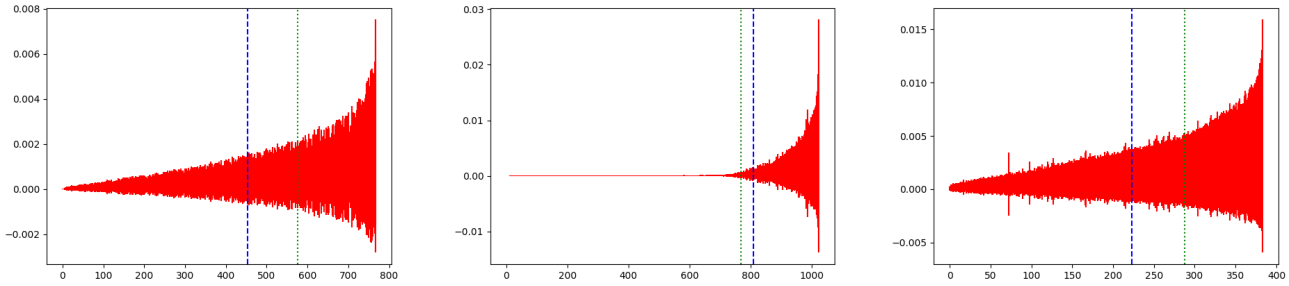
Table 2: Experimental results on the Humor task. Similar to [8] each line refers to experiments conducted with 5 fixed seeds and the first number is the best AUC-Score among them, together with the mean and standard deviations across the 5 seeds in parenthesis. Note, we only had a limited amount of submissions so the non-baseline systems have only one value in the test column indicating the test performance of the best system selected based on the development results.

	[AUC]	
Features	Development	Test
Audio		
[1] Wav2Vec2.0 [8]	.8435 (.8332 ± .0082)	.7940 (.7929 ± .0113)
[2]+ rel. subsp.	.8321 (.8281 ± .0027)	**.8074**
Video		
[1]ViT [8]	.8277 (.7890 ± .0257)	**.7457** (.7478 ± .0093)
[2]+ rel. subsp.	.7697 (.7577 ± .0088)	–
Text		
[1]mBERT [8]	.8105 (.7635 ± .0717)	.7572 (.7108 ± .0830)
[2]+ rel. subsp.	.8065 (.7836 ± .0216)	–
[3]BERT-en	.8274 (.8218 ± .0032)	**.7803**
[4]+ rel. subsp.	.8260 (.7837 ± .0216)	–
Late Fusion		
A^1+T^1 [8]	.8791 (.8600 ± .0218)	.8218 (.8067 ± .0149)
A^2+T^3	.8853	.8381
A^1+V^1+T^1 [8]	.8759 (.8504 ± .0209)	.8310 (.8244 ± .0168)
A^2+V^1+T^3	.8853	**.8420**



(a) BERT (b) W2V2 (c) ViT

Figure 1: Sorted average attributions of each dimensions in the embeddings spaces of the Humor task.



(a) BERT (b) W2V2 (c) ViT

Figure 2: The standard deviations of the sorted attributions of each dimensions in the embeddings spaces of the Humor task.

that thresholding based on the average is a safe option. Perhaps one exception is in the case of ViT, where one feature with low average attribution had a high deviation, but preliminary experiments revealed that adding that extra feature to the selected ones did not offer improvements.

Next, we investigated two possible thresholds for the selection of relevant dimensions. First, we calculated the average attribution of all dimensions and used it to filter out irrelevant parts of the input. As an alternative, we investigated the 75% percentile of the attributions as a decision criterion. For wav2vec 2.0 the latter choice led to a larger feature set, but the results were similar to those we got by using the mean, while for the other embeddings, the mean value proved to be a better threshold, see the results in Table 1.

Based on the results of Table 1, we opted to use the mean attribution as the final threshold. In Table 2, we compare the Decoders trained using only the relevant sub-space (*rel. subsp.*) of the embeddings to the ones utilizing the whole space. In most cases, reducing the input vectors leads to a minor drop in performance on the development set, except for ViT, where we observed considerable degradation. Looking back at Figure 1, we hypothesize that the relatively even attribution distribution of ViT compared to the others signals that the discarded dimensions still contained valuable information.

Based on the development results, we choose to evaluate the Decoder trained on the selected wav2vec 2.0 features on the test set. This model yielded .8074 AUC, which is considerably better than the .7940 AUC-score achieved by the Decoder trained on the whole embedding space. This result indicates that removing the irrelevant components of the embedded vectors improves the robustness of the model and enables better generalization. Another interesting observation is that using only the relevant sub-space also increases training stability in most cases, as the standard deviations of models trained on the reduced input are considerably lower than the baseline ones.

Next, we built the translation pipeline and evaluated the monolingual BERT-en embeddings. Interestingly, this solution outperformed the multilingual model on the development and test sets. This indicates that despite the occasional translation errors, the monolingual embeddings are still better suited for this task.

In the final experiments, we investigated how combined systems would perform. Similar to the baseline article [8], we employed a late fusion technique to merge the outputs of various models. The best results were achieved by fusing the reduced wav2vec 2.0 model with the monolingual BERT and ViT, reaching a .8420 AUC-score, a significantly better performance compared to the A+V+T solutions developed by the organizers.

Table 3: Experimental results on the Mimic task. Similar to[8] each line refers to experiments conducted with 5 fixed seeds and the first number is the highest correlation among them, together with the mean and standard deviations across the 5 seeds in parenthesis. Note, we only had a limited amount of submissions so the non-baseline systems have only one value in the test column indicating test performance of the best system selected based on the development results.

	[Mean ρ]	
Features	Development	Test
Audio		
[1]Wav2Vec2.0 [8]	.4317 (.4290 ± .0020)	.4296 (.4330 ± .0029)
[2] + rel. subsp.	.4311 (.4283 ± .0027)	.4256
[3] Fine-tuned[1]	.4824	***.4644***
[4] + rel. subsp.	.4793	*.4573*
Video		
[1] FAU [8]	.1280 (.1241 ± .0032)	**.1337** (.1319 ± .0019)
[2] + rel. subsp.	.1168 (.1123 ± .0042)	.1140
Text		
[1] ELECTRA [8]	.4079 (.4027 ± .0028)	**.3855** (.3902 ± .0037)
[2] + rel. subsp.	.3972 (.3945 ± .0024)	.3746
Late Fusion		
A^1+T^1 [8]	.4718 (.4695 ± .0022)	**.4679** (.4657 ± .0025)
A^2+T^2	.4706	.4578
A^1+V^1+T^1 [8]	.4789 (.4761 ± .0024)	**.4727** (.4711 ± .0023)
A^2+V^2+T^2	.4770	.4623

6.2 Mimic

For the Mimic task, we conducted the same steps as for the Humor challenge. The aggregated input attribution statistics are depicted in Figures 3 and 4. Similar to the other task, we can see that a considerable portion of the wav2vec 2.0 embedding vector is not utilized, and the distribution of feature importance is again exponential for all systems.

While testing different thresholds for feature selection, we once again concluded that the mean value of the feature attributions is optimal. The experimental results are summarized in Table 3. Unlike previously, here, the reduced input did not offer improvements in terms of performance. Overall, on this task, we can observe the trade-off effect of reducing the input to 25–42% of the original size, thus reducing the Decoder's size considerably, resulting in a relative performance degradation of only 1–15%.

Next, we experimented with the fine-tuning of the whole wav2vec 2.0 model[5] using the Pearson correlation loss, which required considerable time and computational resources. We can see that on the development set, it produced the best results, outperforming even the best multimodal baseline ensembles. Additionally, we also fine-tuned a variant in which we only connected the relevant Transformer outputs to the classification output. This second model achieved slightly worse correlation scores than the complete model but still outperformed all the baseline solutions.

[5]https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

Unfortunately, we could only evaluate this model in the Post-Challenge evaluation phase due to the long time required to find the optimal hyperparameters (mainly the batch size) and the long training process. Consequently, the test results reported are not eligible in the context of the competition but still highlight the superiority of a fine-tuned model compared to solutions only using it as a feature encoder.

As a final step, we also took advantage of the fact that the FAU features are by nature interpretable and visualized the aggregated attribution values using the py-feat toolkit. As Figure 5 depicts, the baseline Decoders trained to recognize approval, disappointment and uncertainty, value the feature dimension associated with the upper lid raiser muscle the most. Additionally, the features linked to muscles controlling the lip are also highly relevant. The main implication of these observations is that we can use Artificial Intelligence to gain new insights about our sentiment analysis data and task, similar to other fields of science [24].

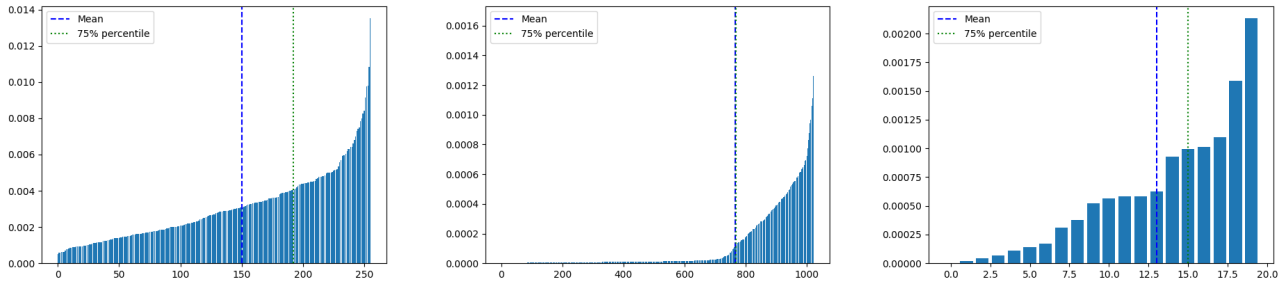
7 CONCLUSIONS

State-of-the-art solutions in many areas, including Sentiment Analysis, are becoming extremely reliant on self-supervised pre-trained models. The large neural models process large quantities of unlabelled data in their initial training phase and are proven to be extremely good feature encoders for numerous tasks. While most effort is dedicated to finding the optimal pre-trained model for a given task, very few works attempt to provide a deeper understanding of what kind of information is encoded in their embeddings. This work investigates multiple popular embedding systems and employs Integrated Gradients, a model interpretation technique, to identify task-specific sub-spaces of various encoders. We employ this new approach to better understand and improve the solutions developed for the MuSe 2023 competition, specifically for the MuSe-Humor and MuSe-Mimic tasks.

Our empirical results indicate that only a small portion of the embedding space is actually relevant for cross-cultural Humor Detection and estimating the intensity of three emotions. We demonstrate that interpreting the initial DNNs trained on the whole embeddings and then filtering their input based on input attribution values can lead to comparable or even better results, while reducing the input size of the Decoder considerably, leading to faster and smaller models.

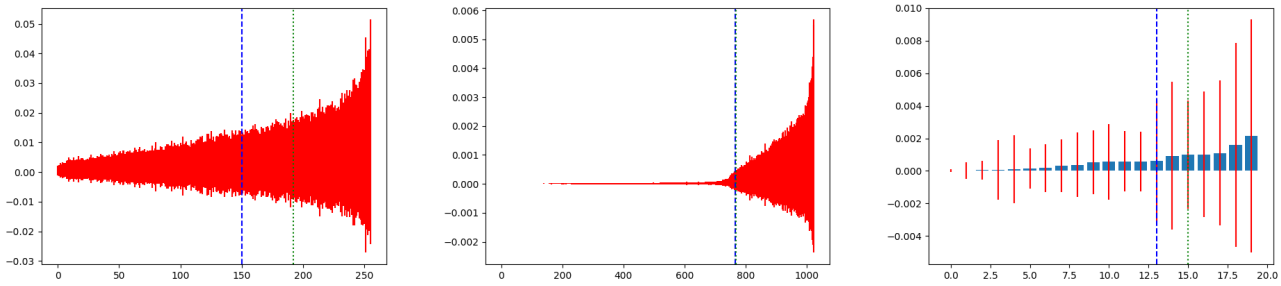
While our work focuses on interpreting only the Decoder components of the various systems, in the future, we aspire to investigate the feature encoders too. Now that we have established which parts of their output are relevant to each task, we can estimate which parts of the raw input (words, sounds and images) contribute most to these dimensions. We hypothesize that with some human annotation effort, it is possible to assign high-level, human-interpretable concepts to sub-spaces of the latent vectors produced by commonly used pre-trained models.

Lastly, we would like to point out that our proposed method can be viewed as a new type of knowledge distillation. In contrast to the traditional way of training the student to mimic the teacher, we utilize the teacher to select the relevant inputs for the student. Additionally, by investigating the input attributions, we could gain



(a) ELECTRA (b) W2V2 (c) FAU

Figure 3: Sorted average attributions of each dimensions in the embeddings spaces of the Mimic data.



(a) ELECTRA (b) W2V2 (c) FAU

Figure 4: The standard deviations of the sorted attributions for each dimensions in the embeddings spaces of the Mimic data.

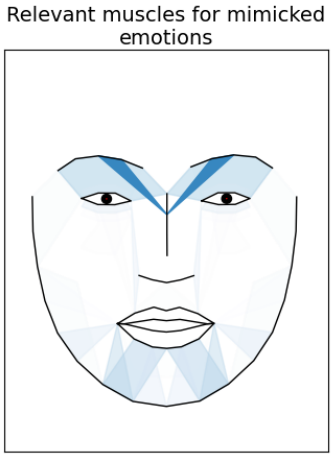


Figure 5: Visualization of the input attributions of FAU features for mimicked emotion detection.

new insights into the data and how DNNs accomplish Humor and Emotion Detection.

ACKNOWLEDGMENTS

The computational resources to perform the experiments were provided by Aalto ScienceIT. The authors are grateful for the founding received from the Academy of Finland project 345790 in ICT 2023 programme's project "Understanding speech and scene with ears and eyes".

REFERENCES

[1] Shivaji Alaparthi and Manit Mishra. 2021. BERT: A sentiment analysis odyssey. *Journal of Marketing Analytics* 9, 2 (2021), 118–126.

[2] Shahin Amiriparian, Lukas Christ, Andreas König, Eva-Maria Messner, Alan Cowen, Erik Cambria, and Björn W. Schuller. 2023. MuSe 2023 Challenge: Multimodal Prediction of Mimicked Emotions, Cross-Cultural Humour, and Personalised Recognition of Affects. In *Proceedings of the 31st ACM International Conference on Multimedia (MM'23), October 29-November 2, 2023, Ottawa, Canada.* Association for Computing Machinery, Ottawa, Canada. to appear.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[4] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48, 1 (March 2022), 207–219. https:

//doi.org/10.1162/coli_a_00422

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[6] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. 2022. ViTFER: Facial Emotion Recognition with Vision Transformers. *Applied System Innovation* 5, 4 (2022). https://doi.org/10.3390/asi5040080

[7] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[8] Lukas Christ, Shahin Amiriparian, Alice Baird, Alexander Kathan, Niklas Müller, Steffen Klug, Chris Gagne, Panagiotis Tzirakis, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. 2023. The MuSe 2023 Multimodal Sentiment Analysis Challenge: Mimicked Emotions, Cross-Cultural Humour, and Personalisation. In *MuSe'23: Proceedings of the 4th Multimodal Sentiment Analysis Workshop and Challenge*. Association for Computing Machinery. co-located with ACM Multimedia 2022, to appear.

[9] Lukas Christ, Shahin Amiriparian, Alexander Kathan, Niklas Müller, Andreas König, and Björn W. Schuller. 2022. Multimodal Prediction of Spontaneous Humour: A Novel Dataset and First Results. arXiv:2209.14272 [cs.LG]

[10] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*. https://openreview.net/pdf?id=r1xMH1BtvB

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[13] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[14] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681

[15] Tamás Grósz, Dejan Porjazovski, Yaroslav Getman, Sudarsana Kadiri, and Mikko Kurimo. 2022. Wav2vec2-Based Paralinguistic Systems to Recognise Vocalised Emotions and Stuttering. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 7026–7029. https://doi.org/10.1145/3503161.3551572

[16] Tamás Grósz, Mittul Singh, Sudarsana Reddy Kadiri, Hemant Kathania, and Mikko Kurimo. 2022. End-to-end Ensemble-based Feature Selection for Paralinguistics Tasks. arXiv:2210.15978 [eess.AS]

[17] Utkarsh Mahadeo Khaire and R. Dhanalakshmi. 2022. Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences* 34, 4 (2022), 1060–1073. https://doi.org/10.1016/j.jksuci.2019.06.012

[18] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. *ACM Comput. Surv.* 54, 10s, Article 200 (sep 2022), 41 pages. https://doi.org/10.1145/3505244

[19] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. arXiv:2009.07896 [cs.LG]

[20] Ninghao Liu, Yunsong Meng, Xia Hu, Tie Wang, and Bo Long. 2020. Are Interpretations Fairly Evaluated? A Definition Driven Pipeline for Post-Hoc Interpretability. *ArXiv* abs/2009.07494 (2020).

[21] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Björn W. Schuller. 2022. Audio self-supervised learning: A survey. *Patterns* 3, 12 (2022), 100616. https://doi.org/10.1016/j.patter.2022.100616

[22] Ali Mirzaei, Vahid Pourahmadi, Mehran Soltani, and Hamid Sheikhzadeh. 2020. Deep feature selection using a teacher-student network. *Neurocomputing* 383 (2020), 396–408. https://doi.org/10.1016/j.neucom.2019.12.017

[23] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, 314–319. https://doi.org/10.18653/v1/W19-5333

[24] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. 2023. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics* 24, 2 (2023), 125–137.

[25] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Proc. Interspeech 2021*. 3400–3404. https://doi.org/10.21437/Interspeech.2021-703

[26] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification?. In *Chinese Computational Linguistics*, Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu (Eds.). Springer International Publishing, Cham, 194–206.

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 3319–3328.

[28] Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. Mono vs Multilingual BERT For Hate Speech Detection And Text Classification: A Case Study In Marathi. In *Artificial Neural Networks in Pattern Recognition: 10th IAPR TC3 Workshop, ANNPR 2022, Dubai, United Arab Emirates, November 24–26, 2022, Proceedings* (Dubai, United Arab Emirates). Springer-Verlag, Berlin, Heidelberg, 121–128. https://doi.org/10.1007/978-3-031-20650-4_10

[29] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

[30] Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. arXiv:1912.07076 [cs.CL]

[31] Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *ArXiv* abs/1909.10430 (2019).

[32] Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT?. In *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Online, 120–130. https://doi.org/10.18653/v1/2020.repl4nlp-1.16

[33] Patrick Xia, Shijie Wu, and Benjamin Van Durme. 2020. Which *BERT? A Survey Organizing Contextualized Encoders. In *Conference on Empirical Methods in Natural Language Processing*.

[34] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. 2019. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer* 36 (2019), 1067–1093.