On the Cross-Modality of Pre-Training

Diego Alysson and Leonardo Boulitreau

December 2021

Abstract

Transfer learning is a key concept in deep learning. It enables models to acquire knowledge from previously trained models, avoiding the exhaustive work to always develop them from scratch. Current state-of-the-art Natural Language Processing (NLP) models are almost entirely based on the pre-training-finetuning transfer approach. This work focuses on analysing how the models pre-trained in different domains can rapidly adapt to a different one on a downstream task. We perform a text classification experiment on the IMDB dataset with the BERT architecture loading weights from models pre-trained in speech and image, which are the Wav2Vec2.0 and the Vision Transformer (ViT), respectively. Also we experiment with finetuning directly the Wav2Vec2.0 to the text task, in order to remove any possible inductive bias for the text that could be present on BERT. Result show that the transfer only happened with from the Wav2Vec2.0, but it happened to be statistically as good as a Xavier Uniform Initialization.

1 Introduction

Transfer learning is a field of research consisting in a methodology of machine learning that focuses on transferring knowledge across domains. The concept might have initially came from educational psychology in a sense that transfer is considered as the result of generalization of experience [1]. Among the subareas in transfer learning, sequential transfer learning is the most promising in natural language processing (NLP). It refers to the method in which tasks are learned in sequence. There is first a pre-training phase in which representations are learned from a source domain or task, and then an adaptation (or finetunning) phase, in which the learned knowledge is applied to target task or domain [2]. This approach of pre-training in a data-rich dataset provides the model the ability to develop general-purpose abilities and knowledge that can be passed to a downstream task. It has been typically done on a supervised manner and more recently in an unsupervised way, providing many state-of-the-art results [3].

There is strong evidence that, in a general way, the intrinsic nature of the pre-training can be better than initializing a model from scratch. In [4], for example,

the authors analyse the performance of a pre-trained model for 26 different downstream text classification tasks. By using only an adapter module, which adds just a few parameters, near state-of-the-art performance is attended.

Beyond text, speech processing in the context of deep learning requires the understanding of both language and also acoustic content, such as phonemes, tones, words and semantic meanings [5]. Thus, it is reasonable to imagine that the pre-training in this domain would yield also some general text knowledge to be used in a downstream alike task.

Considering this approach, we aim to evaluate the hypothesis that instead of always developing new models, one can use a pre-trained (even if it is in a different domain than that from the desired target task) to obtain better performance, being that model already has some general knowledge. Furthermore, another possible use for these models could be an aggregation of knowledge: in each new pre-training realized of the model some general knowledge of the new domain is adhered by the model, making the representations richer along both domains.

Our contributions are the following:

- We evaluate the knowledge transfer between models with pre-training in different domains of the target task.
- We analyse the performance of a BERT [6]-like architecture in a text classification task with pre-training done both audio (Wav2Vec2.0) [7] and image (Vision Transformer(ViT)) [8] domains.
- We attempt performing a text classification task without any inductive bias regarding text by using the entire Wav2Vec2.0 architecture and pre-training.

This work is divided as follows: Section 2 contains some research done related to transfer learning in the NLP context; Section 3 details the theoretical background of the architectures used; Section 4 describes the experiments performed, the configurations, and the dataset used; Section 5 presents the outcome of the experiments; Section 6 contains a brief discussion concerning the results; Section 7 raises conclusions to sum up the overall transfer analysis and Section 8 presents some future work to be done.

2 Related Work

Cross-Domain Transfer Learning: Recent work has studied the knowledge transfer between models pre-trained with different-but-related domains. The authors of [9] train Long-Short-Term-Memory (LSTM) models on non-linguistic data and also in other languages and evaluate these with a fine-tuning on a downstream NLP task. Several domains are considered, such as MIDI music, JAVA code, artificial parentheses languages, and several languages, with the fine-tuning being in a Spanish task. The authors intended to measure how

much the structure awareness of a language model acts as an inductive bias to improve the performance when transferred from one language or symbolic system to another.

The work of [10] aims to analyse the growing evidence that pre-trained models can transfer knowledge for both new languages and also non-linguistic data. The authors explore how many of this transfer remains when word identity is lost (with scrambled domains). They experiment with BERT and a GloVe [11] initialized LSTM. These pre-trained models are evaluated on both normal and scrambled dataset of sequence classification and labeling tasks. The results show high rates of transferability only for the BERT model in the evaluated scrambled domains and only in the task of classification.

The authors of [12] used the BERT pre-trained in text to perform an Automatic Speech Recognition (ASR) downstream task. This was performed through two fine-tuning steps: the first consists in creating a language model using the transcriptions of the speech samples, and the second, to properly generate the desired ASR sequence, consists in using a decoding model to output the amount of words necessary. It receives the transcriptions for comparing the outputs, and also the acoustic embeddings, to be the inputs to each decoding step. These acoustic embeddings are obtained through an acoustic encoder, which converts the acoustic features into the desired embeddings through a small convolutional net.

There is also plenty of work considering transfer from the pre-training in one language to another in a downstream task: [13], [14] and [15].

Multimodal Pre-Training: There has been also extensive approaches in literature that aim to unify representations for both speech and text domains. These works are focused on the task of Spoken Language Understanding (SLU), which consists in receiving an input speech signal and then understanding its linguistic content and make predictions. The work of [16] consists in a semi-supervised framework that performs a joint pre-training of aligned text and speech modules (a BERT and a Transformer Encoder, respectively) and obtains a shared latent space for both domains. When fine-tuned, it improves the state-of-the-ate performance on the spoken SQuAD data set by more than 10%.

Recently, the authors in [17] unify two pre-trained models, the BERT for the text data, and the Wav2Vec2.0 for speech data with an aggregation module, an attention-based net. Their results outperform the current approaches in low-resource speech recognition, with a CER (Character Error Rate) of as low as 3.8% on the AISHELL-1 dataset [18].

Other papers that also perform speech-language pre-training in a similar way are [19] and [20]. These works show that the aggregation of speech and text in a common representation is possible. So, even with the format difference, a connection between text and speech data is obtained, thus paving the way for a knowledge construction based that can be obtained and transferred through both types of data.

3 Background

Among different architectures based on transformers, BERT, Wav2Vec2.0, and the Vision Transformer (ViT) were selected to carry out the transfer learning study. These were chosen due to their similarity in the use of a transformer (facilitating pre-trained weight transfer) and also because of their different pre-training data types, ranging from text processing, audio, and images, which is important for evaluating the intended cross-domain setup.

3.1 BERT

The BERT (Bidirectional Encoder Representations from Transformers) [6] language model was developed for the training of deep bi-directional models from unlabeled text. As shown, in literature, pre-training models are effective in the auto-supervised NLP context. However, the algorithms that implemented these techniques were unidirectional, which greatly limited the architecture, in addition to representing a sub-optimal solution. BERT corrects these restrictions by adding the language model mask (MLM) that randomly masks text tokens for the purpose of predicting them. This strategy allowed for a bi-directional context representation and proportionate state-of-the-art results in several NLP tasks.

BERT possesses two main steps: the pre-training, which consists in training the model with unlabeled data between different tasks, and the fine-tuning of parameters initialized from pre-training, now adjusted using labeled data. BERT's architecture, shown in Figure 1, is a bidirectional multi-layer Transformer encoder, implemented as originally described in [21]. Initially, two architectures of the BERT model were described and evaluated, the $BERT_{BASE}$ (L=12, H=768, A=12, Total Parameters=110M) and the $BERT_{LARGE}$ (L=24, H=1024, A=16 , Total Parameters=340M), where L is the number of transformer blocks, H the hidden size, and A the number of self-attention heads. As input to the model, BERT deals with either a single or a pair of sentences. The first special token, [CLS], is used for aggregation or classification tasks. For pairs of sequences, the token [SEP] is also used to separate them.

Two main pre-training techniques were reported for the model, the Masked language model (MLM) and Next Sentence Prediction (NSP), which was discarded after due to mediocre performance. The main reason for using the MLM technique is the possibility of bi-directional training, forcing the model to predict which words are missing in the input text and not the entire input. For the experiments reported in BERT, 15% of all tokens were masked. After, fine-tuning is performed in order to adapt the pre-trained model to a desired target task. Fine-tuning is described as having a much lower computational cost as compared to model's pre-training stage.

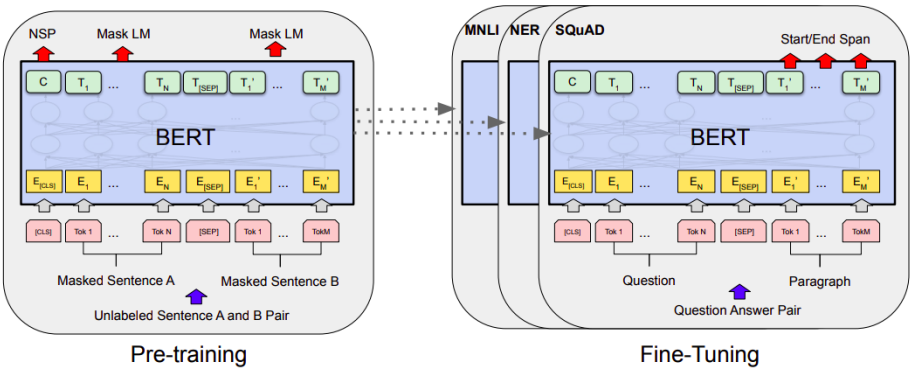

Figure 1: BERT Framework.

3.2 Wav2Vec2.0

To also take advantage of self-supervised learning, the Wav2Vec2.0 [7] framework presents an architecture capable of using this paradigm for training a network with raw audio data. The model aims to learn contextualized general audio representations for usage in other downstream tasks such as audio classification and automatic speech recognition.

The architecture begins with a feature encoding layer based on convolutional blocks: a convolutional layer is followed by a normalization layer and a GELU activation function. Then, these representations are handed over to transformer layers that can contextualize them. Instead of absolute positional embedding layers, convolutional layers capable of representing the relative positions were implemented. As part of this network training, representations calculated by the feature encoder are discretized, due to previous empirical work that presented good results. These can be chosen by the model in a differentiable manner through the Gumbel softmax function. The objective is calculated between the transformer output and the discrete representation with the goal being a combination of a contrastive and a diversity loss over the discrete codebooks. The representation of this framework is presented in Figure 2.

Pre-training the Wav2Vec2.0 is performed similarly to BERT, through the MLM, which in this case requires the model to identify the correct latent quantized audio representation. The pre-trained models are then fine-tuned for speech recognition or other tasks by adding a linear projection layer to the entire context network.

3.3 ViT

The Vision Transformer (ViT) [8] is a framework proposed by the Google Brain Team to learn general representations from images that can be then fine-tuned for specific downstream tasks. The authors consider a Transformer architecture

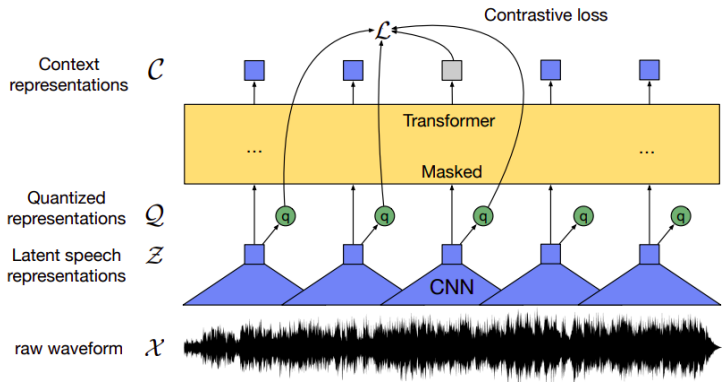

Figure 2: Wav2Vec2.0 Framework

with the fewest possibile modifications, due to its success in NLP applications.

The ViT architecture, shown in Figure 3, starts by dividing the input image in fixed-size patches that are then flattened and grouped to form the input sequence. This sequence is projected through a linear layer to the model's hidden dimension (for the base type it is $D = 768$), so that with the addition of common 1D positional embeddings, it can be received by the Transformer encoder (the same as in [21]). Similar to BERT, there is a learnable embedding on the transformer output to represent the overall image.

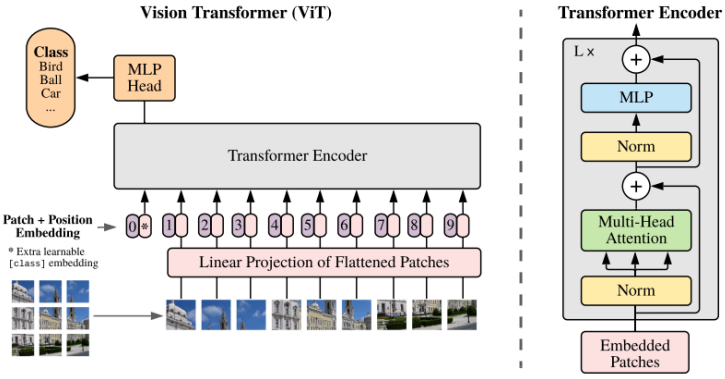

Figure 3: Vision Transformer (ViT) architecture

During supervised pre-training, a multi-layer Perceptron (MLP) with one hidden layer is used for classifying the output image representation in one of target classes. On the fine-tuning stage, this block is changed to a single linear layer. They acquire better than current state-of-the-art results (big ResNets) for several image classification tasks using large dataset. The authors also experiment

with self-supervised pre-training, using a masked patch prediction objective and consider the approach promising.

Another interesting observation is the fact that the ViT has way fewer inductive biases when compared to the standard CNN-based nets for images, which contain locality, translation equivariance and a 2D neighbourhood structure. Almost all layers of the ViT are completely global, with the exception of the MLP layers, that have translational equivariance and are local.

4 Methodology

To carry out a methodical study regarding the impact of transfer learning on networks previously trained in different domains, three different architectures were considered. Firstly the BERT, as the base model which will receive the knowledge transfer. This model is designed to handle text-type input. Two other networks, Wav2vec2.0 and ViT were chosen for their applications in contexts other than BERT, with Wav2vec2.0 being originally used for audio and ViT for images. The basic task for the experiments is the sentiment analysis, using the IMDB [22] dataset.

4.1 Dataset

In this work, the acl-IMDB dataset [22] was used as a base to evaluate and compare the different approaches and models. It is formed by a set of film analyses. On the collected data, each analysis received a score, which varies between 1 and 10. Analyzes with values below 5 are said to be negative and values above 6 are given as positive. The set contains 50000 analyses, which, in this work were divided into 20000 training samples, 5000 evaluation samples, and 25000 test samples.

The task considered is to classify the film reviews as being predominantly positive or negative. This dataset was selected due to the simplicity of both the data and the task that will be performed, being the focus of the work in comparing the model's results. A seed was set to ensure all models receive the same sample order in all training, validation and test sets and in all experiments realized.

4.2 Experiments

The code for all experiments is made available online at [1].

4.2.1 Baseline

Some scenarios were designed to quantify the importance of knowledge acquired by networks in different contexts. Initially, the baseline values were obtained

[1]Code for the experiments: `https://github.com/LeonardoBoulitreau/On-The-Cross-Modality-of-Pre-Training`.

using the pre-trained BERT model. Four different experiments were performed for this reference setup. Two experiments with training from scratch, where the BERT network is initialized with Xavier initialization [23] values: uniform and normal. The purpose of these experiments was to verify the learning capacity of a BERT architecture without any prior knowledge (no pre-training). The other two experiments were performed using BERT pre-trained with MLM ("bert-base-uncased"). With these initial configurations, we have the results referring to the use of the text BERT model in tasks of its domain. These are intended to serve as a reference and be compared with obtained through the other proposed experiments.

To do the classification of feelings in the IMDB base, all configurations considered in this work have a linear layer at the end, also Xavier-initialized, for the final classification between positive and negative.

4.2.2 Hierarchical Transfer

To understand the impact of knowledge contained in networks of different domains (or modes) gradually, different configurations were conceived. As the main knowledge extraction network, the Wav2Vec2.0 network ("facebook/wav2vec2-base") was used, transferring its weights to BERT ("bert-base-uncased"). Initially, only one of the transformer layers was transferred. This way, the impact of each cross-domain layer on the total knowledge of the network can be evaluated.

In this setup, three layers were then selected: the first, the intermediate (6th) and last layer, to evaluate hierarchically the contribution from the pre-training in audio to the text classification task. For each of these three layers, other experiments were carried out, totalizing six experiments, on which the selected layer was replaced by a Xavier-Initialized one. Thus, each considered layer was either substituted for other without prior knowledge or other by the analogous transformer layer of the Wav2Vec2.0 structure, with prior knowledge of audio.

4.2.3 Cross-Modal Transfer

After verifying the impact of knowledge transfer in different levels of layers, a considered "complete transfer of knowledge" from Wav2Vec2.0 to BERT was performed. For this, the standard embeddings layer was used, as performed in the original BERT and then all the transformer layers were loaded with weights coming from the Wav2Vec2.0 ("facebook/wav2vec2") model, pre-trained on the Librispeech [24] dataset.

In this experiment, we analyse the difference in performance between base and large models, and also conduct a freezing test, with the intention of compensating the initial information mismatch of the layers, similarly to [9]. For this, three different configurations were performed. Large BERT was used with the transformer layers initialized with Xavier; with weights transferred from Wav2Vec2.0 Large and finally with weights frozen transferred from Wav2Vec2.0 Large. In

this way, we can directly compare the results of the large and base model, and evaluate the freezing procedure.

Another approach proposed as Cross-Modal Transfer is with the use of ViT, in this way it is possible to compare and assess the differences between the use of different contexts, text, audio, and images, in addition to understanding whether contexts can be more correlated if compared to a second context. We take the best transfer case of the Wav2Vec2.0-based experiments, and we replicate to the ViT, to avoid reproducing not ideal experiments.

4.2.4 Semi-Cross-Modal Transfer

All experiments described so far were carried out with models that transferred their knowledge to BERT with only a pre-training in audio. However, in this test, we consider the Wav2Vec2.0-960h ("facebook/wav2vec2-base-960"), which is not only pre-trained, but also fine-tuned on 960 hours of the Librispeech dataset for an ASR task. Thus, the model has already experienced text entries in some way. The intention is to verify if this processing performed on the ASR task can somehow improve the performance of the text classification task by using the knowledge acquired.

To perform this comparison, the Wav2Vec2.0-960h weights were transferred to the BERT architecture. In a similar way to the experiment considering the model that did not have any contact with text, base and large models are evaluated. Also, we perform a freezing test to obtain the best freeze configuration: we experiment with freezing both the transformer and the embedding layers and only the transformer.

4.2.5 Removing the Inductive Biases

Finally some experimentation was done with the entire Wav2Vec2.0 architecture directly. Apart from entering directly with the textual tokens from the task, we considered experiments varying the textual input by performing L2 Normalization, Min-Max Normalization and Standard Normalization in order to approach the tokens sequence to a temporal signal. Also, some architectural changes were performed, such as changing the stride of the initial convolutional layers from [5,2,2,2,2,2,2], to [5,1,1,1,1,1,1] and [1,1,1,1,1,1,1] in order to control the information context window. Lastly, we experimented with freezing the transformer and also the transformer and embeddings during the finetuning process.

In all cases the experiment did not converge, being the result the same as randomness.

5 Results

To perform a fair comparison on the tests, all experiments were performed with the same conditions and hyperparameters, described in Table 1. All results are

considered on the test set. The curves are obtained on the validation set, to analyse their convergence and overfitting. For the models that used freezing techniques, the selected layers were frozen during all training epochs.

Name	Value
Optimizer	ADAMAX [25]
Learning Rate	0,0001
Batch Size	16
Epochs	11
Acc. Grad	5
Max Length	512

Table 1: Hyperparameters

The proposed baseline values can be seen in table 2. It can be observed that networks without any pre-training, which do not carry prior knowledge, had results comparable to randomness. On the other hand, the results obtained through the pre-trained models, with prior knowledge, and that were pre-trained in the same domain (from texts) consisted in values up to 93% of accuracy, with an F1 of 96% and a loss of 0.37. The statistical equivalence between the Xavier Uniform an the Xavier Normal initialization is also noted.

Model	Test Acc	Test F1	Test Loss
BERT-no PT w/ Xavier Uniform Init	0,50016	0,50006	0,69371
BERT-no PT w/ Xavier Normal Init	0,50016	0,50006	0,6937
BERT PT w/ CLS Xavier Uniform Init	0,93549	0,96560	0,35763
BERT PT w/ CLS Xavier Normal Init	**0,93646**	**0,96621**	**0,37283**

Table 2: Baseline Results

The results of the second experiment, referring to the partial transfer learning (single-layer) from the Wav2Vec2.0 network to BERT, are described in the table 3. It is observed that weights exchanged in the last layer of the transformer present values of up to 93% accuracy, F1 96% and loss 0.43. With the transfer of the intermediate layers (6th layer) the values obtained were up to 91% accuracy, 95% F1, and loss 0.46. Finally, concerning from the base layer of the transformer (1st layer), the maximum values obtained were with no transfer, but with Xavier initialization, consisting of 92% accuracy, an F1 of 0.96, and a loss of 0.37.

For the transfer of the whole Transformer's weights from the Wav2Vec2.0 network to the BERT, the results are described in the table 4. Among the different experiments performed, the best results were obtained through the base model, with no freezing, presenting results of up to 86% accuracy, 92% F1, and 0.39 loss. For large models, the best results were 63% of accuracy, an F1 of 71% and a test loss of 0,66137.

The experiments involving the pre-trained Wav2Vec2.0-960h have their results described in 5. In these models, the best results obtained were 86% accuracy,

Model	Test Acc	Test F1	Test Loss
BERT w/ 12th W2V2 Layer	**0,93457**	**0,96510**	**0,43287**
BERT w/ 1th Layer Xavier Init	0,92938	0,96221	0,36737
BERT w/ 12th Layer Xavier Init	0,91845	0,95602	0,39108
BERT w/ 6th Layer Xavier Init	0,91833	0,95020	0,45526
BERT w/ 6th W2V2 Layer	0,91505	0,95413	0,39099
BERT w/ 1th W2V2 Layer	0,85299	0,91665	0,57755

Table 3: Hierarchical Transfer Results

Model	Test Acc	Test F1	Test Loss
BERT w/ W2V2 Transformer	**0,86615**	**0,92551**	**0,38558**
BERT w/ Transformer Xavier Init	0,86559	0,92548	0,50063
BERT w/ W2V2 Frozen Transformer	0,85060	0,91642	0,36546
BERT w/ W2V2 Transformer (Larges)	0,50000	0,50000	0,69357
BERT w/ Transformer Xavier Init (Larges)	0.50000	0.50000	0.69322
BERT w/ W2V2 Frozen Transformer (Larges)	0,63624	0,71927	0,66137

Table 4: Cross-Modal Transfer Results

92% F1, and 0.44 loss.

Model	Test Acc	Test F1	Test Loss
BERT w/ W2V2-960h Transformer	**0,86952**	**0,92785**	**0,43809**
BERT w/ W2V2-960h Frozen Transformer	0,81566	0,89157	0,51889
BERT Frozen Embeds w/ W2V2-960h Frozen Transformer	0,51224	0,61885	0,69284

Table 5: Semi-Cross-Modal Transfer Results

Finally, the results considering the transference of the transformer's weights from the imaged-based ViT architecture to the BERT are shown in Table 6. The maximum values obtained were 50% accuracy, F1 50% and loss 0,69.

6 Discussion

As noted in the baseline results, networks that do not carry pre-training have results close to random. These results were expected since the amount of data and training time were not enough for the network to have the capacity to learn and converge. While networks with prior knowledge demonstrate a high capacity of transferring for the BERT architecture, to perform the fine-tuning of data. It is observed through Figure 4 that models without prior knowledge present a slight loss improvement, which might be statistically insignificant, and they do not converge to good results. The results of fine-tuning the pre-trained models in text quickly converged, with up to one epoch. Even though a slight overfitting occurred in the next epoch, the weights of the best loss were saved.

For the models proposed for exchanging single BERT layers by weights trans-

Model	Test Acc	Test F1	Test Loss
BERT w/ ViT Transformer	0,50016	0,500006	0,69325
BERT w/ Frozen ViT Transformer	0,49984	0,49977	0,69090

Table 6: ViT-BERT Transfer Results

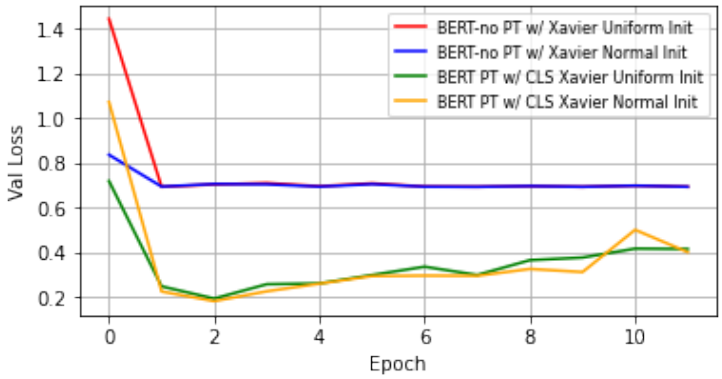

Figure 4: Convergence of Baseline

ferred from Wav2Vec2.0, it is observed that models that have layer exchanges closer to the end of the network, layers that have more refined representations, present results similar to "fully BERT models". With this, it can be considered that these layers do not have a high impact on the total learning of the network. Compared with the exchange of initial layers, in this case, layer 1, there is a great decay of the results, reaching a maximum of 85% accuracy, F1 91% and loss 0.58, for the Wav2Vec2.0 weights, demonstrating the importance of these layers for knowledge representation and transfer between domains. For intermediate layers, no major changes were observed in the results, when compared to the original model.

It was also compared with the transfer of Wav2Vec2.0 an initialization of selected layers with the Xavier initialization. With the use of this algorithm, regardless of the changed layer, it did not show large losses, which demonstrates that despite the loss of existing knowledge in the network from pre-training, it is smoothed with good initialization algorithms. It is also possible to see that, compared to a network initialized from scratch, there is a transfer of knowledge from Wav2Vec2.0 that allows the network to converge to a decent value, but this knowledge is equal to the exchange of these same layers by an optimized initialization algorithm. On the Figure 5, it can be seen that the models converge at similar times, except for BERT with the first layer exchanged by Wav2Vec2.0 weights which has a convergence delay, possibly due to the adjustment of the most important Transformer weights to the different domains.

Concerning the results obtained through the transference of the weights of all

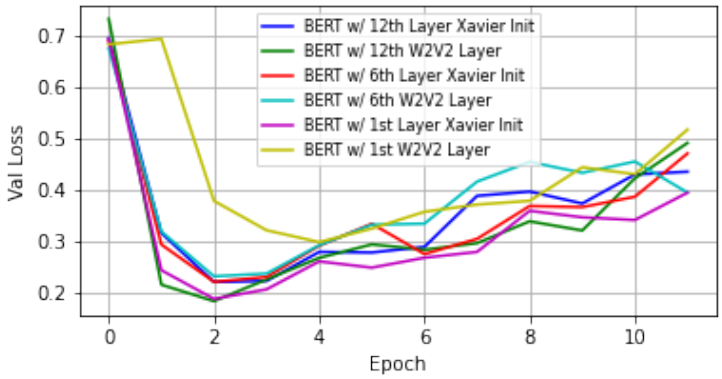

Figure 5: Convergence of Hierarchical Transfer

Wav2Vec2.0 Transformer's layers to the BERT's, similar results to those obtained in the exchanges for specific layers can be observed. There is a decrease in the results of the network when compared to the baseline. However, the network can converge, unlike models that start without any pre-defined weight. On the other hand, it can still be observed that values initialized in an optimized way, through the Xavier initialization have results statistically equivalent to those of Wav2Vec2.0.

Also, the transfer of models with transformer network freezing was also evaluated, which presented results close to non-frozen networks, demonstrating that a large part of the new network learning is in the embeddings section, while the weights of transformers can be practically re-used from the imported context. Additionally, one notable result is the faster convergence when freezing the transformer, since less parameters need to be changed. Large models were also studied but presented results were inferior to the base ones. This was expected since larger models need more parameters to be adjusted, and thus, possibly more epochs and data to converge, as seen in the Figure 6.

With resepct to the Wav2Vec2.0-960h, better results were expected due to the fine-tuning performed by this model, but the results obtained demonstrate that there are no gains concerning the standard model. Despite the fine-tuning in a task that has already observed text, there can not be noted better text-induced general representations.

Another result obtained in these tests is related to layer freezing. The freezing of embeddings and transformers of the hybrid network or only the transformers was compared. The results indicate that freezing the embeddings affected the performance in the task, and it is possible to verify that the new knowledge acquired in fine-tuning influences the pre-trained embeddings. When only the transformers are frozen, the network still manages to converge with good results. Through Figure 7 this is observed. As in the other experiments, the BERT

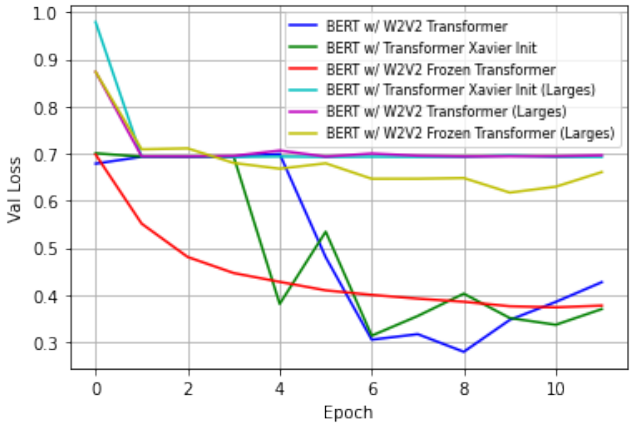

Figure 6: Convergence of Cross-Modal Transfer

embeddings block with the Wav2Vec2-960h transformer layers present a good convergence, followed by the BERT embeddings model and frozen transformer layers from Wav2Vec2-960h, which use prior audio and text knowledge but still train the values of embeddings and lastly, the model with both embeddings and transformers frozen, which did not converge.

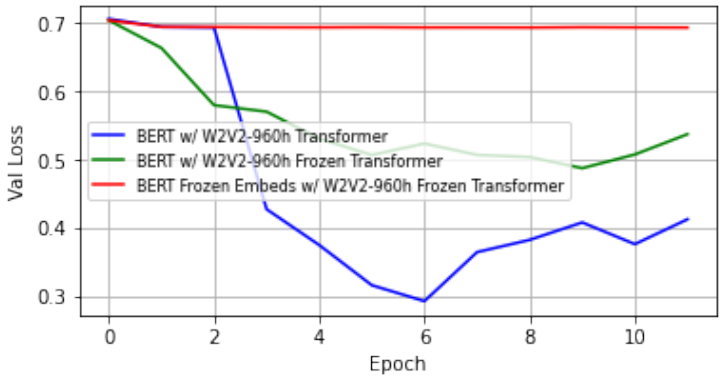

Figure 7: Convergence of Semi-Cross-Modal Transfer

To compare the results obtained through the transfer of knowledge between different domains, the ViT framework was also used, with knowledge transfer from images. Unlike the weights of Wav2Vec2.0, the knowledge acquired from ViT did not contribute to the text task, not converging the final result, in the two different experiments performed, with fully trainable ViT or with transformers freezing, which showed good results in Wav2Vec2.0. As can be seen in Figure 8. This indicates that the transferred knowledge depends on the original knowledge

area, demonstrating that some areas can be more correlated than others, such as audio to text.

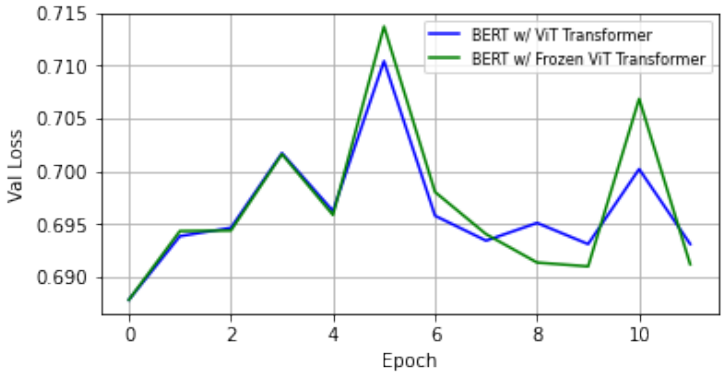

Figure 8: ViT Experiment

7 Conclusion

In this work, a methodical study on transfer learning between different transformer architectures pre-trained on several domains was carried out. Three domains were considered, text, through the BERT architecture, audio with the Wav2Vec2.0 architecture, and image, through the Vision Transformer (ViT). Among these domains, different scenarios were proposed, amid them, the comparative study between untrained networks, networks with in-domain pre-training, and networks finetuned through knowledge transfer. We also experimented with layer freezing, different number of layers transferred, different model sizes.

The results show the feasibility of knowledge transfer and demonstrate the advantages of this technique. However, it was also observed that optimized startup settings are as advantageous as transferring correlated areas. Also, we noted the advantage of using smaller base models due the faster convergence when compared to the large ones, which is possibly good for rapid adaptation. On a few-shot task for example, one might consider loading weights from a similar domain and use a smaller architecture, instead of starting from scratch. Lastly, it was also found that some domains may have a more intricate relationship with each other than others, in this case speech and text, when compared to the image domain.

8 Future Work

It is expected to expand the experiments accomplished with the ViT architecture to understand the reason of its worse performance when compared to the

pre-training in audio. In the same context, a study focusing on analysing the similarity of domains considering the pre-training-fine-tuning setup is also of interest.

Additionally, it is also necessary to analyze and compare the transfer of knowledge between other areas. Audio and image to text was studied in this work, but still open studies such as text to audio or image related tasks.

Finally, more sophisticated types of transfer that go beyond simply loading weights might be considered. Either some work in a similar way to section 4.2.5 of the use of an encoder to transform acoustic features on embeddings that are more recognisable to a text pre-trained net, for example, can be performed.

References

[1] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685, 2019.

[2] Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[4] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019.

[5] Hang Li, Wenbiao Ding, Yu Kang, Tianqiao Liu, Zhongqin Wu, and Zitao Liu. CTAL: Pre-training cross-modal transformer for audio-and-language representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3966–3977, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[8] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[9] Isabel Papadimitriou and Dan Jurafsky. Learning music helps you read: Using transfer to study linguistic structure in language models. *arXiv preprint arXiv:2004.14601*, 2020.

[10] Zhengxuan Wu, Nelson F. Liu, and Christopher Potts. Identifying the limits of cross-domain knowledge transfer for pretrained models. *CoRR*, abs/2104.08410, 2021.

[11] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[12] Wen-Chin Huang, Chia-Hua Wu, Shang-Bao Luo, Kuan-Yu Chen, Hsin-Min Wang, and Tomoki Toda. Speech recognition by simply fine-tuning bert. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7343–7347, 2021.

[13] Ke M. Tran. From english to foreign languages: Transferring pre-trained language models. *CoRR*, abs/2002.07306, 2020.

[14] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: an empirical study. *CoRR*, abs/1912.07840, 2019.

[15] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019.

[16] Yu-An Chung, Chenguang Zhu, and Michael Zeng. SPLAT: Speech-language joint pre-training for spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1897–1907, Online, June 2021. Association for Computational Linguistics.

[17] Guolin Zheng, Yubei Xiao, Ke Gong, Pan Zhou, Xiaodan Liang, and Liang Lin. Wav-bert: Cooperative acoustic and linguistic representation learning for low-resource speech recognition, 2021.

[18] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline.

In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5, 2017.

[19] Cheng-I Lai, Yung-Sung Chuang, Hung-Yi Lee, Shang-Wen Li, and James Glass. Semi-supervised spoken language understanding via self-supervised speech and language model pretraining, 2020.

[20] Yao Qian, Ximo Bian, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng. Speech-language pre-training for end-to-end spoken language understanding, 2021.

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

[22] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[23] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[24] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.