DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DESIGN AND MANUFACTURING KANCHEEPURAM
CHENNAI - 600127

Synopsis Of

Enhancing the Transfer Performance of Pre-trained Models across Similar Domains & Tasks

A Thesis

To be submitted by

MERCY FAUSTINA J

For the award of the degree

Of

DOCTOR OF PHILOSOPHY

1 Abstract

Natural Language Processing (NLP) is an indispensable milestone to achieve artificial general intelligence. Even when humans find it hard to discern the elusive context from text information because of its ambiguous nature, deep neural networks which are a staple in artificial intelligence have exhibited remarkable performance exceeding humans' expectations. Despite playing a pivotal role in machine learning, deep neural networks come with a slew of downsides like computational burden, high training time, dearth of training data etc., which limit their potential. On the other hand, the conventional machine learning paradigm always encourages learning a task in isolation, but in practice, most of the NLP tasks share many commonalities among them. Therefore, as a way of exploiting this shared common knowledge and eliminating the dependency on a humongous amount of training data, transfer learning is introduced. Transfer learning techniques help to repurpose a large trained model from one task or domain to another task or domain. This elicits broad applicability of pre-trained models to various tasks rather than associating it to only one task.

Witnessing the enormous success of pre-trained models, many similar huge pre-trained models are developed in line with such interest. Developing a pre-trained model is often capped by hardware constraints and computational resources which poses a serious caveat for the research community. This has awoken interest among researchers to leverage the performance of existing pre-trained models rather than developing a new pre-trained model from scratch. In a similar vein, this dissertation further investigates how to enhance the transfer performance of pre-trained models, particularly for Natural Language Understanding (NLU) tasks. The contributions presented in this thesis can be broadly classified into two categories based on the transfer strategy: Fine-tuning (Parameter-transfer) and Feature-based (Feature-transfer). In addition to this, attempts were made both in the lines of domain adaptation and task adaptation.

Firstly, the thesis studies the problem of enhancing the transfer performance for single-sentence classification tasks like sentiment analysis. We propose three novel approaches involving Bidirectional Encoder Representations from Transformers (BERT): STATS-BERT, prompting BERT and an ensemble of BERT and Convolutional Neural Network (CNN) architecture. The former work follows a fine-tuning approach where the versatile BERT model is further trained with a tweaked masking objective named statistical masking in place of random masking. The intent behind proposing this objective is to mask only statistically significant words, thereby, encouraging the model to predict them to gain more insights into the context. The second approach is prompt-based finetuning, which conflates the advantages of both prompting and fine-tuning strategies. We explored the impact of prompt templates in instigating the model to generate the right answers through manually designed templates and presented rudimentary results. Thirdly, we propose an ensemble model involving BERT and CNN architecture variants, which is purely a feature-based approach. As CNNs are adept at aggregating local information, we explored various CNN variants to complement the BERT features. All the above approaches work towards the same goal of capturing the underlying semantics of the input sentence to deliver better transfer performance for single sentence classification tasks.

Secondly, we extend the problem of enhancing the transfer performance of pre-

trained models for sentence-pair tasks like sentence similarity. The challenge in dealing with sentence pair tasks is capturing the interaction and relevance across sentences. Though models like CNN are skillful in capturing intra-sentence details, they lack inter-sentence details. Moreover, native pre-trained representations from models like BERT are isotropic in nature, hence, produce sub-optimal results for sentence similarity tasks. To remedy this issue, many strategies like whitening, contrastive learning are introduced on top of the pre-trained model. We propose two ways to further enhance the contrastively learned BERT representations: [i] Fast Fourier Transform (FFT)-enhanced contrastive framework [ii] Stacked co-attention mechanism. The former approach is an exploitation of a mathematical transformation, the Fourier transform, which worked in favor of producing meaningful representations in the vector space. The latter work jointly attends to the words of the other sentence to produce co-dependent representations to favor sentence-similarity tasks.

While all the previous contributions focus on the line of task adaptation, the third contribution concentrates on adapting a model to handle real-world applications through domain adaptation. We carry out this research along two verticals: [i]COVID domain adaptation with CORD-19 text data and [ii]Artistic Style transfer with Tanjore art images. Through COVID domain adaptation, we study how the source domain can positively or negatively influence the target domain based on its relevance to the target domain. We identified that biomedical models exhibit positive transfer, whereas, a generic aly trained model shows deterimental effects on target domain. Hence, we developed a Question-Answering (QA) system which is also a sentence-interaction task by leveraging a biomedical model and dedicate it to COVID domain. Artistic style transfer is another dimension of transfer where the source and target domains differ by their artistic styles. In this case, a subset of input attributes is transferred via a trained neural model onto the output space. Amid Western art style transfer works, we introduce Tanjore art style adaptation into God images which is more challenging as Tanjore art images do not hold clear distinguishable features. We attempted this task of style transfer using various neural architectures and found cycleGAN to be the ideal architecture to transfer the input style despite preserving the content of the output image. We introduce custom defined metrics to measure both content preservation and success of style transfer.

In a nutshell, the thesis extensively studies the problem of leveraging the transfer performance via features or parameters of existing pre-trained models to serve a similar domain or task through various proposed methodologies as discussed.

2 Objectives

This thesis studies the below listed problems in detail and proposes novel approaches to remedy the problem.

- **Enhancing the semantic understanding of pre-trained models** Though pre-trained models exhibit remarkable improvement over a plethora of downstream tasks, the successor model variants proved that there is still scope for advancement in many facets such as refinement of training objectives, hyperparameter tuning etc. In this line, we make an attempt to develop a model that is capable of capturing the underlying semantics of the input text to enhance the performance

of single-sentence classification tasks.

- **Enhancing the raw representations of pre-trained models** Although, the enormous success of pre-trained models can hardly be overstated, pre-trained models operate well only when they are put under pretrain-finetune paradigm. The native representations produced by the pre-trained models are isotropic in nature and do not lend meaningful representations for tasks like sentence-similarity. Hence, we enhance the representations of pre-trained models post pre-training.

- **Domain Adaptive Training** Pre-trained models are generically trained without any inclination towards any specific domain. Therefore, pre-trained models are less capable to handle domain-specific jargons. To resolve this issue, it is necessary to introduce a domain adaptive training phase as we adapt the models for a specific task.

- **Artistic Style Transfer** A new dimension of transfer learning where a subset of input attributes is transferred to output space with the help of a neural model. We explore in the lines of visual style transfer where style of one image is transferred over content of another image while preserving the content information.

Overall, the thesis investigates the research question, "How to enhance the transfer performance further for a specific downstream task of our interest using both feature-based and fine-tuning approaches?"

3 Existing Gaps Which Were Bridged

After an extensive study on the existing literature work, some research gaps and areas that have scope for improvement have been identified. They are listed as follows:

- As far as transfer learning setup is concerned, pre-training is a highly exhaustive phase, hence, leveraging the existing pre-training models by introducing novel objectives help to unlock the full power of pre-trained models. Hence, we meticulously tweak the pre-training objectives having the existing pre-trained model as a warm startup.

- Another vertical in repurposing the pre-trained models is by handling the features of the model efficiently. Though fine-tuning pre-trained models is found to be beneficial, feature-based approaches still have scope for betterment. Hence, we leverage the semantic understanding of the model output features using post-processing techniques.

- Albeit, pre-trained models are exalted enough for its remarkable performance, the native representations produced by them are not readily usable for sentence-interaction tasks. This is because of the lack of sentence-level information captured through its pre-training objectives. Hence, it is a prudent choice to imbibe sentence-level information into pre-trained models. We propose novel co-attention mechanism and FFT-enhanced framework to serve the purpose.

- Another axis of using pre-trained models is to shift the model to a new domain from its base through domain adaptation. Using pre-trained models for real world applications is a compelling yet challenging task. Therefore, adapting models like BERT or style of an input to a popular domain of interest is an interesting research direction. We explore COVID domain adaptation and image style adaptation.

4 Most Important Contributions

The research objectives listed in Section 2 are tackled using our three-fold contributions. The overall contributions of the thesis can be categorized based on three aspects of transfer learning: [i]Adaptation Style [ii]Task type, [iii]Transfer type as depicted in Figure 1

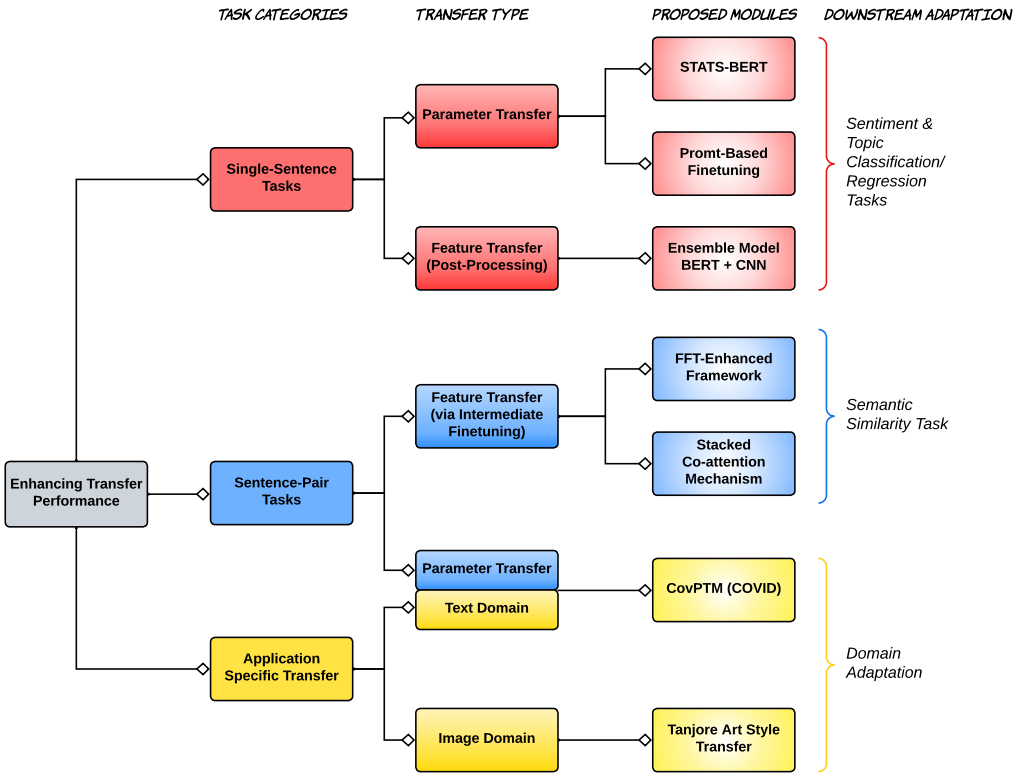


Figure 1: Work Organization Chart

4.1 Contribution 1: Enhancing Transfer Performance for Single-Sentence Tasks

The first contribution of this thesis focuses on enhancing the underlying semantics of the given text.

4.1.1 Parameter-Transfer

We propose two strategies under this transfer strategy: [i] Conventional fine-tuning and [ii] Prompt-based fine-tuning. As pre-training is a highly laborious task, we introduce both the approaches on top of the existing pre-trained model as a warm-startup.

STATS-BERT STATS-BERT is a statistically enhanced BERT Devlin *et al.* (2019) via fine-tuning approach. It focuses on refinement of masking strategy of BERT. We formulate a statistical masking objective to mask only significant words of the sentence involving statistical properties such as Term Frequency-Inverse Document Frequency (TF-IDF) and Entropy Montemurro and Zanette (2010).

TF-IDF is a product of two terms: TF and IDF as shown in Equation 1 that assigns weightage to every token reflecting the importance of a word specific to a document, given the large document corpus.

$$TF - IDF_{t,d,D} = TF_{t,d} * IDF_{t,D} \tag{1}$$

where t represents the term, d represents the document size and D number of documents in the corpus.

Shannon's entropy informationMontemurro and Zanette (2010) is the second measure used that relates word distribution in a document with its associated context and linguistic roles. This measure assigns higher value to words that follow non-uniform distribution than the words that are evenly distributed across the document. An input document of N tokens is split into P equal sized parts and entropy of word w, $H(J|w)$ is calculated as shown in Equation 2.

$$H(J|w) = -\sum_{j=1}^{P} \frac{n_j}{n} log_2 \frac{n_j}{n} \tag{2}$$

where n_j represents the frequency of the word w in part j of the given document and n represents overall word frequency. As shown in Equation 3, the information content in the document part is calculated as the difference between Shannon's mutual information of the given part of text and the shuffled version of it.

$$\Delta I_w(s) = p(w) * [\langle \hat{H}(J|w) \rangle - H(J|w)] \tag{3}$$

where $p(w)$ is the normalized frequency

We combine both statistical properties and present a unified measure for statistical masking objective as shown in Equation 4.

$$\Delta I_w(s) = TF - IDF * [\langle \hat{H}(J|w) \rangle - H(J|w)] \tag{4}$$

The proposed STATS-BERT model has been repurposed for downstream text classification tasks from various genres, particularly, for single sentence classification tasks. Results are tabulated in Table 1.

Table 1: STATS-BERT Results on Downstream Tasks

TASK Metric	MR Acc/F1	IMDB Acc/F1	TREC Acc	AGNews Acc	SST(Dev) P.corr/S.corr	SST(Test) P.corr/S.corr	SST-2 Acc
Random-BERT	73.9/75.5	91.9/92	96.3	94.1	80.5/80	81.6/81	91.8
STATS-BERT	74/75.3	91.7/91.8	96.9	94.4	80.4/80.1	82/81.3	91.7

As the statistical objective aligns with the topic and question classification task, STATS-BERT shows 0.6%, 0.3% improvement in accuracy than baseline model and shows on par performance over other tasks.

Prompt-Based Fine-tuning Though pretrain-finetune paradigm is ideally applicable across many tasks, it requires minimal task-specific training, it does not help data scarce settings. Hence, in the second work, we introduce prompt-based finetuning following the setup of AMuLAP Wang *et al.* (2022) and LMBFF Gao *et al.* (2021a) where minimal gradient updates are permitted via limited training instances to conflate the advantages of both approaches. Our prime focus is to carefully design prompts in such a way that the pre-trained language model is prompted to generate the most appropriate label words as the high probability tokens that represents the respective class. We conducted this experiment over RoBERTa Zhuang *et al.* (2021) model and explored various prompt templates and studied the impact of them on the downstream datasets such as SST2, QNLI, MRPC and found a noticeable improvement on SST2 and QNLI. We present only subset of SST2 results briefly. The manually curated prompt templates and the corresponding generated label words are presented in Table 2.

Table 2: Generated Label Words for SST2

Template	**Positive**	**Negative**
<S1>The movie is [MASK].	here excellent amazing great fantastic brilliant below superb terrific terrifying phenomenal incredible wonderful awesome outstanding beautiful	over terrible awful bad good horrible garbage trash finished done weird gone boring disgusting disappointing possible
<S1>All in all [MASK]	excellent great recommended impressive enjoyable positive brilliant fantastic awesome amazing outstanding, worthwhile satisfying scratch solid	good though disappointing work bad yes mediocre fun ... fine no... true this : equal

The results obtained via these prompt templates are tabulated in Table 3. As the SST-2 dataset contains movie reviews only, including the word "movie" gives better scores whereas adding generic terms like **"$\langle S1 \rangle$. All in all [MASK]"** gives detrimental effects and the label words of the negative class includes positive words like good, fine, yes etc.

Table 3: Results from RoBERTa by introducing the manually designed prompts on SST-2

Template	**SST-2**
Full Finetuning	95
Manual Label 0 shot	83.6
AMuLaP	93.2
Auto-L	93.5
AMuLaP + Auto-L	93.4
$\langle S1 \rangle$ The movie is [MASK].	**94.38**
$\langle S1 \rangle$ All in all [MASK]	89.22

4.1.2 Feature-Transfer: Ensemble Model

Parameter-transfer gives significant gains when compared to feature-transfer method but some of the downsides like inducing computational overhead, losing pre-trained knowledge, demanding task-specific data, extended training duration etc limits its usability. Hence, this module targets the same objective of enhancing the underlying semantics but through feature-transfer. An ensemble model is developed involving BERT and CNN architecture variants Wu *et al.* (2019). As CNNs are adept in accessing the global and local information, we uncovered multiple variants of CNN architecture like dilated Yu and Koltun (2016), depthwise CNN to leverage the semantic quotient of the representations of a few popular pretrained models like BERT, RoBERTa, DistilBERT for the downstream classification tasks.

Table 4: Results on BERT Model Features (BERT-CNN)

Models	#T_P(M)	Sentiment Classification				Topic Classification				Question Classification			
		SST-2		IMDB-2		AGNews		NewsGroup		TREC-6		TREC-47	
Baselines(Feature-Based)		Acc	F1	Acc	F1	Acc	Wt.F1	Acc	Wt.F1	Acc	Wt.F1	Acc	Wt.F1
BERT + Bi-LSTM	23.6	88.4	88.9	86.7	86.8	93.4	93.4	73.4	73.2	95.1	95	74.8	70.4
BERT + Bi-GRU	17.7	89.9	90.2	87.3	87.4	93.8	93.8	73.7	73.1	94.7	94.5	74.4	70.7
Proposed Static Version													
BERT + Conventional CNN	4.6	91.1	91.3	**88.1**	**88.1**	**94.3**	**94.3**	87.4	87.4	95.9	95.8	79.3	76.1
BERT + Dilated CNN	6.6	**92**	92.2	87.8	87.9	93.6	93.6	**89.9**	89.8	**96.6**	**96.5**	87.7	86.5
BERT + Dilated CNN (context window restricted)	5.9	91.9	**92.3**	87.9	87.9	93.7	93.7	89.8	**89.8**	96.4	96.4	**87.8**	**86.9**
BERT + Depthwise CNN	4.6	88.8	89.3	85.4	85.6	92.9	92.9	71.6	70.9	88.2	86.9	59.6	48.7
Fine-tuned BERT (Baseline)	109.4	92.5	92.6	93.8	93.8	94.7	94.7	93.1	93.1	96.7	96.6	92.3	92
Proposed Dynamic Version													
BERT + Conventional CNN	114	92.8	93.1	88.9	89	94.7	94.7	93	93.1	97.5	97.4	93.4	93.3
BERT + Dilated CNN (context window restricted)	115.3	92.4	92.5	88.7	88.7	94.7	94.7	93.3	93.3	97.5	97.4	93.8	93.8
BERT + Depthwise CNN	114	92.2	92.2	88.6	88.8	94.7	94.7	92.3	92.4	97.4	97.3	93.4	93.2

*The scores in bold are the best scores among the Feature-Based Approaches. Underlined scores are the overall best. #T_P(M) represents number of trainable parameters.

We treat the large language models as sentence encoders and use sum of last four layer representations as seeds of transfer which are of dimension d*L, where d represents hidden dimension(768) and L represents sequence length(128). This work can also be viewed as post processing of the features obtained from the successful pre-trained models. We experimented both static and dynamic versions where pre-trained model weights are frozen and unfrozen respectively. Out of all the CNN variants, we infer that the dilated CNN variant substantially outperforms the RNN counterparts and other CNN variants in 5 out of 6 downstream tasks. On the other hand, depthwise CNN,

another variant of CNN that is popular for efficient use of parameters shows detrimental effects. The experimental results are tabulated in Table 4.

4.2 Contribution 2: Enhancing Transfer Performance for Sentence-Pair Tasks

The second contribution focuses on leveraging the performance on sentence pair tasks like sentence-similarity and question-answering tasks.

4.2.1 Sentence Similarity Tasks

The proposed modules for sentence similarity task are feature-enhancement via intermediate fine-tuning. The motivation behind this feature enhancement post pre-training is that the native representations rendered by the pre-trained models are isotropic in nature and do not lend well for tasks that involve sentence-interaction. To mitigate the aforementioned issue, contrastive learning setup helps Gao *et al.* (2021*b*). Following this, we propose two enhancement modules on top of this contrastively trained BERT in a siamese network architecture that helps pre-trained models to deliver meaningful representations: [i] FFT-SimCSE framework [ii] Stacked Co-attention mechanism.

FFT-SimCSE Framework In this framework, inspired from FNet architecture Lee-Thorp *et al.* (2022), we introduce parameter free Fourier layers in a contrastive learning setup to highlight the context of the input, thus, producing meaningful representations. We retrofit three Fourier transform layers on top of contrastively trained BERT to enhance the semantic quotient of the representations. The intuition behind having multiple FFT layers is to declutter the input hierarchically and imbibe only the needed information in the sentence representation.

Algorithm 1 Stacked Co-Attention Mechanism

Input: A Sentence pair, $< S_1, S_2 >$
Output: A Pair of Co-Attended Representations, $< \hat{S}_1, \hat{S}_2 >$
$< S_1^t, S_2^t >\Leftarrow$ WordPiece Tokenizer(S_1, S_2)
$< \tilde{S}_1, \tilde{S}_2 >\Leftarrow$ Siamese BERT Encoder(S_1^t, S_2^t)
C $\Leftarrow$ Affinity Matrix Calculation
$\hat{S}_1$ = Stacked Coattention(C, S_1)
$\hat{S}_2$ = Stacked Coattention(C^T, S_2)
Stacked Co-Attention Module (C, S):
    $C_r \Leftarrow$ Row wise Normalization(C)
    $C_c \Leftarrow$ Column wise Normalization(C)
    Attented Representation, $S_A \Leftarrow$ S * C_r
    Stacked Attended Representation, $S_{AA} \Leftarrow S_A$ * C_c^T
    Attention_weights, A $\Leftarrow$ Sum_over_word_dimension(S_{AA})
    Co-Attended Representation, $\hat{S} \Leftarrow$ S * A

Table 5: Evaluation of Sentence Embeddings on STS Tasks. Reported Scores are spearman correlation values between gold labels and similarity scores.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
InferSent-GloVe	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
USE	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
$SBERT_{base}$	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$ConSERT_{base}joint$	70.53	79.96	74.85	81.45	76.72	78.82	77.53	77.12
$ConSERT_{base}sup-unsup$	73.51	84.86	77.44	83.11	77.98	81.80	74.29	79.00
$ConSERT_{base}joint-unsup$	74.07	83.93	77.05	83.66	78.76	81.36	76.77	79.37
$SimCSE-BERT_{base}$	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
FFT-SimCSE*	76.06	84.94	80.86	86.05	81.63	84.42	80.27	82.03
Stacked Co-Attention*	**77.64**	**86.02**	**82.2**	**87.78**	**82.87**	**85.90**	**81.50**	**83.41**

*On comparing to baselines, underlined scores are best scores of FFT-SimCSE and bold scores are best scores of Stacked Co-attention Mechanism.

Stacked Co-attention mechanism This approach enhances the representations of the pre-trained model in a contrastive framework by developing co-dependent representations between a given pair of sentences using attention-over-attention strategy Cui *et al.* (2017). The stacked co-attention mechanism that we propose is parallel in nature where computation of one co-attending feature is independent of the other sentence in the pair as shown in Algorithm 1.

Table 5 summarizes the performance scores obtained over the full test set of STS tasks. From the results, it is inferred that FFT-SimCSE framework and stacked co-attention mechanism outperforms all the considered baselines including the closest work, SimCSE Gao *et al.* (2021*b*) by 0.45% and 1.84% on average respectively.

4.2.2 Question-Answering Task

This work focuses on leveraging the performance of Question-Answering task in the line of domain adaptation.

COVID-19 Domain-Adapted Pre-Trained Model Though pretrained models enabled us to limit our dependency on training data and compute resources, they are generically trained models. A generic model is not a good fit for all domains as every domain holds a long list of domain-specific jargons that is incomprehensible for the generic pretrained model. This challenge is addressed by introducing an intermediate pre-training towards the target domain with target specific data called as domain adaptation. We choose CoVID-19 as our domain of interest and used CORD-19 Wang *et al.* (2020) dataset for target domain adaptation. We carry out the research in two parallel directions in order to understand the proportionality of relatedness between source domain and target domain with rate of transfer performance. Hence, we picked one dissimilar or distant relative which is BigBird Zaheer *et al.* (2020) trained on generic corpora that exhibited negative transfer as shown in 6a.

TASK	VS	CC	SST-2
BERT	62.5	82.0	72.4
BigBird	70.1	89.6	86.8
CR-Bird	62.1	76.8	61.7

(a) Accuracy scores of COVID-19 Adapted BigBird Model

Model	Exact Match	F1
BERT-base	24.3	43.2
ELECTRA	25.2	43.9
COVID-BioBERT	30.4	56.8
COVID-BioElectra	26.9	51.2
COVID-PubMedBERT	32.7	57.6
COVID-BioM-ALBERT	39.1	64.7

(b) Accuracy scores of COVID-19 Adapted Biomedical Models

Another source is from an adjacent domain or nearest neighbors which are biomedicals models like BioBERT Lee *et al.* (2020), BioELECTRA Kanakarajan *et al.* (2021) etc tht exhibited positive transfer as shown in 6b. Hence, we developed a question-answering model specifically dedicated for CoVID domain using CoV-QA dataset and recorded greatest improvement in the COVID-BioM-ALBERT model of more than 21% increase in F1 score than the considered baseline.

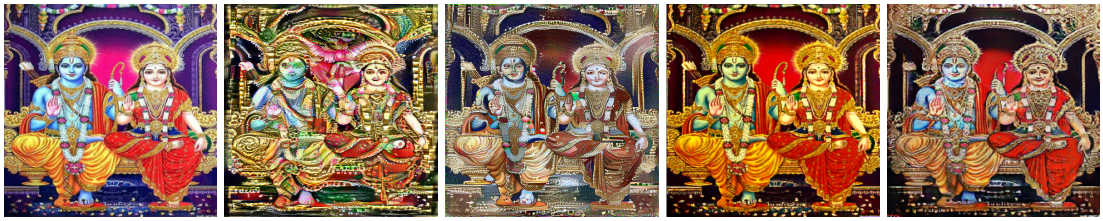
4.3 Contribution 3: Application-specific Domain Transfer

The third module takes an application in particular and shifts the input domain images to target domain which differs by its artistic style. Style transfer is transfer of visual appearance from source domain image to target domain image. In other words, style transfer task considers two input images: content image and style image where the derived style representations from the style image is transferred onto the content image without any information loss. The existing style transfer models mainly focus on style of western art forms. We study a more challenging problem of style transfer of tanjore art paintings which does not have clearly distinguishable features. Out of all the considered models like VGG, nn-Hallucinations, high resolution network, the CycleGAN model works best for this purpose as it ensures meaningful mapping across source and target domain images because of its cycle consistency loss. CycleGAN Zhu *et al.* (2017) optimally captures the style features of the Tanjore paintings and applies the same onto the output content image where it embosses the ornaments and highlights gold textures with perceived depth as shown in Figure 2. The style transfer outputs produced by all the models are displayed in Figure 3. As an attempt to further enhance the success of transfer, we follow it up with an enhancement module where the 256*256 output image from cycleGAN is scaled to a higher resolution of 1024*1024 using SRGAN network.

(a) Content (b) Style (c) Output

Figure 2: Style Transfer output from the cycleGAN model

While minimal effort is taken towards quantitatively validating the style transfer outputs, we evaluate our proposed model output with two metrics as follows: Structural



(a) Content Img. (b) VGG Model (c) nn-Hallucinations (d) Photorealistic (e) CycleGAN

Figure 3: Comparing outputs from different style transfer models with Tanjore painting as style.

Similarity Index (SSIM) and Effectiveness (E) score. SSIM measures content preservation and E scores measure success of style transfer as shown in Table 7 and Table 8 respectively.

Table 7: SSIM score of the cycleGAN style transferred image.

Image	Ghiasi et al	Gatys at al	Ours
Img	0.51	0.51	**0.71**

Table 8: E score of images before and after enhancement. Ei is the E score of the i^{th} layer of the pre-trained VGG neural network.

Image	E1	E2	E3	E4	E5
Img before enhancement	-1.43259	-1.59743	-2.86097	-4.02971	-4.21953
Img after enhancement	-1.06566	-0.82541	-2.23804	-3.71348	-4.17003

5 Conclusions

The emergence of transfer learning techniques in NLP propelled a new research discipline in deep learning leading to the evolution of enormous heavily trained large neural models. Rather than developing another large pre-trained model, the thesis studies the problem of enhancing the transfer performance of the existing pre-trained models particularly for NLU tasks that includes intra-sentence and inter-sentence tasks. As part of solving this objective, the dissertation focuses on bridging the gaps identified through literature survey by proposing novel strategies following both feature-transfer and parameter-transfer approaches. Overall, the work is carried out in two directions: task-specific adaptation and application domain-specific adaptation. The former approach aids in elevating the semantic understanding of the models and its features, thereby enhancing transfer performance for the downstream tasks in question. The latter approach is an exploration of other dimensions of transfer learning like domain adaptation and style transfer.

6 Organization of the Thesis

The proposed outline of the thesis is as follows:

(a) Chapter 1: Introduction

(b) Chapter 2: Literature Survey

(c) Chapter 3: Enhancing Transfer Performance through Fine-tuning Approaches.

(d) Chapter 4: Enhancing Transfer Performance through Feature-based Approaches

(e) Chapter 5: Enhancing Transfer Performance for Sentence-Pair Tasks.

(f) Chapter 6: Application Specific Domain Transfer.

(g) Chapter 7: Conclusion and Future Scope

7 List of Publications

I. REFEREED JOURNALS BASED ON THE THESIS

1. Mercy Faustina J, Akash V, Anmol Gupta, Divya V, Takasi Manoj, Sadagopan N, Sivaselvan B, A Study of Neural Artistic Style Transfer Models and Architectures for Indian Art Styles *Network: Computation in Neural Systems, Taylor and Francis*, vol.34, pp 285-305, (2023). (SCIE-IF:7.8)
   DOI: 10.1080/0954898X.2023.2252073.
2. Mercy Faustina J and B Sivaselvan, An Ensemble Neural Model using BERT and Dilated CNN to leverage semantic quotient for Text classification *The Journal of Supercomputing*, Springer (Under Review).
3. Mercy Faustina J and B Sivaselvan, A Novel FFT & Stacked Coattention Mechanism to enhance the semantics of Sentence Representations *Journal of Experimental and Theoritical Artificial Intelligence*, Taylor and Francis (Under Review).

II. PRESENTATIONS/PUBLICATIONS IN CONFERENCES BASED ON THE THESIS

1. Mercy Faustina J, Abhinand Rajagopal, N Kausik, Pranav Parameshwaran and B Sivaselvan, STATS-BERT: An Enhanced BERT model with statistical masking for improved transfer performance *9th International Conference on Pattern Recognition and Machine Intelligence (PReMI'21), ISI Kolkata*, vol 13102, Springer LNCS (2021).
2. Mercy Faustina J, Aparajith Raghuvir and Sivaselvan B CovPTM : COVID-19 PreTrained Model - Generic Model vs Biomedical Model Adaptation *20th IEEE India Council International Conference (INDICON)*, (2023).
3. Mercy Faustina J, Paleti Krishnasai, Sivaselvan B, A Prompt-Based Transfer Learning System for Text Classification *(Submitted to 29th Annual International Conference on Natural Language & Information Systems (NLDB 2024))*

References

1. **Cui, Y.**, **Z. Chen**, **S. Wei**, **S. Wang**, **T. Liu**, and **G. Hu** (2017). Attention-over-attention neural networks for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 593–602, doi:10.18653/v1/P17-1055.

2. **Devlin, J.**, **M.-W. Chang**, **K. Lee**, and **K. Toutanova** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, doi:10.18653/v1/N19-1423.

3. **Gao, T.**, **A. Fisch**, and **D. Chen** (2021*a*). Making pre-trained language models better few-shot learners. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, **1**, 3816–3830, doi:10.18653/v1/2021.acl-long.295.

4. **Gao, T.**, **X. Yao**, and **D. Chen** (2021*b*). SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910, doi:10.18653/v1/2021.emnlp-main.552.

5. **Kanakarajan, K. r.**, **B. Kundumani**, and **M. Sankarasubbu** (2021). BioELECTRA:pretrained biomedical text encoder using discriminators. *Proceedings of the 20th Workshop on Biomedical Language Processing*, 143–154, doi:10.18653/v1/2021.bionlp-1.16.

6. **Lee, J.**, **W. Yoon**, **S. Kim**, **D. Kim**, **S. Kim**, **C. H. So**, and **J. Kang** (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

7. **Lee-Thorp, J.**, **J. Ainslie**, **I. Eckstein**, and **S. Ontanon** (2022). FNet: Mixing tokens with Fourier transforms. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4296–4313, doi:10.18653/v1/2022.naacl-main.319.

8. **Montemurro, M.** and **D. Zanette** (2010). Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, **13**(2), 135–153.

9. **Wang, H.**, **C. Xu**, and **J. McAuley** (2022). Automatic multi-label prompting: Simple and interpretable few-shot classification. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5483–5492. URL https://aclanthology.org/2022.naacl-main.401.

10. **Wang, L. L.**, **K. Lo**, **Y. Chandrasekhar**, **R. Reas**, **J. Yang**, **D. Burdick**, **D. Eide**, **K. Funk**, **Y. Katsis**, **R. M. Kinney**, **Y. Li**, **Z. Liu**, **W. Merrill**, **P. Mooney**, **D. A. Murdick**, **D. Rishi**, **J. Sheehan**, **Z. Shen**, **B. Stilson**, **A. D. Wade**, **K. Wang**, **N. X. R. Wang**, **C. Wilhelm**, **B. Xie**, **D. M. Raymond**, **D. S. Weld**, **O. Etzioni**, and **S. Kohlmeier** (2020). CORD-19: The COVID-19 open research dataset. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

11. **Wu, F.**, **A. Fan**, **A. Baevski**, **Y. Dauphin**, and **M. Auli** (2019). Pay less attention with lightweight and dynamic convolutions. *International Conference on Learning Representations*.

12. **Yu, F.** and **V. Koltun** (2016). Multi-scale context aggregation by dilated convolutions. *International Conference on Learning Representations*.

13. **Zaheer, M.**, **G. Guruganesh**, **K. A. Dubey**, **J. Ainslie**, **C. Alberti**, **S. Ontanon**, **P. Pham**, **A. Ravula**, **Q. Wang**, **L. Yang**, *et al.* (2020). Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, **33**, 17283–17297.

14. **Zhu, J.-Y.**, **T. Park**, **P. Isola**, and **A. A. Efros** (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

15. **Zhuang, L.**, **L. Wayne**, **S. Ya**, and **Z. Jun** (2021). A robustly optimized BERT pre-training approach with post-training. *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 1218–1227.