A Study of Multimodal Architectures for improving non-English Classification

Team Fusion Force : Alberto Debes Shreyas Verma Saksham Arora Suraj Shourie

Abstract

Natural language processing (NLP) innovations have completely changed how scholars and practitioners approach important societal issues. Modern approaches for text recognition and classification applications are now based on large language models. However, the development of advanced computational techniques and resources is disproportionately concentrated on the English language, sidelining most of the languages used in the world. While prior work has improved multilingual and monolingual models, we investigate the potential of multimodal machine learning to close the performance gap between English and non-English languages by incorporating the features captured in images. We focus on crisis-information dataset [1] *and show that adding image modality to non-English (Hindi in our case) text embeddings does improve upon the downstream classification task.*

1. Introduction

1.1. Aim

Large pre-trained Natural Language Processing models like BERT [1] are frequently employed in research and practice for classification tasks to address social issues [6] [7]. However, these models tend to under-perform for other languages apart from English, which may restrict the insights they can derive. Researchers have looked to tackle this issue by using multimodal learning to access information from images [8]. Through a comparative analysis of different text and image networks, we seek to understand which architectures work best. We also seek to understand how much images can improve the discrepancy in performance between English and non-English text classification models.

Research has shown that text-only classification models consistently have higher performance with the English language inputs over non-English language across both monolingual and multilingual models, different classification tasks, and metrics [8]. This has provided motivation for testing novel approaches that can improve this disparity in performance [5]. Researchers have succeeded in lessening the performance gap by leveraging a multimodal approach that combines image and text classification models. They found that performance improves considerably when including images as a modality compared to text-only classification models [8]. We are intrigued by the information that images can add to a classification system in conjunction with textual embeddings. Our hypothesis is that architectures like Inception-v3 and ResNet-18, which use the novel technique of skip connections, should increase the performance further and provide us with more robust embeddings from both these modalities. Also, state-of-the-art vision architectures such as Vision Transformers would also produce embeddings that are able to better encapsulate image information and add value to the multimodal setting. Our goal in this project is to present a comprehensive comparative analysis of the various architectures and provide a trade-off between model complexity and incremental improvement in performance.

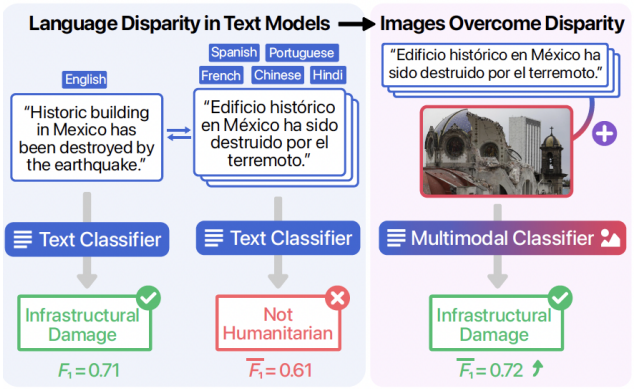

Figure 1. Overview figure. We use multimodal (image + text) learning to overcome the language disparity that exists between English and non-English languages. The figure illustrates an example of a social media post that is correctly classified in English but misclassified in Spanish. Including the corresponding image leads to correct classification in Spanish as well as other non-English languages.

[1] https://crisisnlp.qcri.org/crisismmd

Figure 2. Illustrative examples from considered multimodal dataset. We consider the crisis humanitarianism dataset (number of classes: 5; infrastructure and utility damage: 10%, rescue volunteering or donation effort:14%, affected individuals: 1%, other relevant information:22%, not humanitarian:53%),

1.2. Dataset

We use the CrisisMMD dataset [2] for classification as this task saw the largest benefit from adding images to Hindi text models in [8]. This shows promise that our hypothesis can be validated more distinctly on this dataset for a non-English language classification task. This dataset consists of over 10k tweets, each with text and an image regarding a humanitarian crisis (e.g. hurricane Harvey, California wildfires, etc). This dataset is originally available in English language only. Given the lack of non-English multimodal datasets and for the sake of performing the analysis across different langauges on the same dataset, these tweets have been translated to Hindi using the MarianNMT langauge translation model [8]. This model is an industrial-grade system that also powers the Microsoft Translator and hence for the sake of maintaining simplistic assumptions and focussing on the research question at hand, we assume that the translations are accurate. Each entry is also labeled in terms of how much damage the crisis caused, ranging from "little/no damage" to severe damage. The data originally has 8 labels which have been clubbed into 5 broad categories. Furthermore, we focus on how the disparity in performance changes between Hindi and English language models as a result of our experimentation. Our reasoning being that Hindi language models have observed the largest improvement in performance from images as observed in [8].

[2]https://crisisnlp.qcri.org/crisismmd

We split the dataset into three using a 75%-15%-15% split to create the training, validation and test dataset. This gives approximately 6000 images and text for training, 1000 each for validation and testing.

2. Approach

To test our hypothesis that for classification task, the results of language only model can be improved using visual data as well, we first see the standalone results of both text and visual models and finally compare that to a multi-modal model.

Image-Only Model: To explore the impact of images on this classification task, we trained a Deep Residual Network (ResNet-18) to compute image embeddings. ResNet models have been shown to be better at propagating information through networks for image classification tasks thanks to their deep architecture. They leverage multiple layers while still having lower complexity than their counterparts (i.e. VGG networks)[3]. We hypothesized that using a deep residual networks will produce richer embeddings that can reduce the performance gap between English and non-English in language models.

We also fine-tune an Inception V3 architecture, which has been originally pre-trained on the ImageNet data. Inception V3 suggests stacking "wide" Inception modules in the architecture rather than simply "going deep" with the layers. It shows how different convolution filters when placed in parallel can encapsulate different parts of an image/its intermediate layer embedding and thus help capture information from various perspectives through these differently shaped filters. The reason why we selected Inception v3 along with ResNet for our analysis is because both the architectures stem for different ideologies and perform quite good on classification tasks. This helps us compare the results of two deep Vision architectures to check if there is a difference in the robustness of the embeddings produced by both, for the downstream classification task.

As images in our datasets have various dimensions, we apply a standard image pre-processing pipeline so that they can fit the pre-trained ResNet-18 model's input requirement. We test the output of our ResNet-18 model on the test data using accuracy, F1-score, confusion matrix etcetra. These are shown in the Results section 4.We also save the last layer of the ResNet architecture and use them as input embeddings for our multi-modal model later. We do the same for the Inception V3 architecture where we apply replace the penultimate final as well as auxiliary layers of the architecture with new fully-connected layers, to be fine-tuned on our dataset. The input images need to be resized to a 224x224 and 229x229 dimension for ResNet and Inception-V3 respectively.

Text-Only Model: We choose two pre-trained BERT models for their simplicity and ease of use and also how

effective they have proved in creating robust textual embeddings. First one DistilmBERT (distilbert-base-multilingual-cased on HuggingFace) to classify the English text and HindiBERT for Hindi (Doiron 2020). We fine-tune the pre-trained model on the training dataset. The process of fine-tuning a language model involves taking a pre-trained language model and replacing the "pre-training head" of the model with a randomly initialized "classification head" [8]. Similar to the image-only model described above, the results for the text-only model for both English and Hindi languages are shown in the results section 4. We also save the last layer of the BERT models as "embeddings" to be used in the multi-modal model.

Multimodal Model: We implement a multimodal or fusion classifier that combines the embeddings of both text and image data to perform classification based on the joint modeling of both input modalities. We use the embeddings created by the last layer of both text-only and image-only model. The two embeddings will be on different feature space based on what model is used to create them, as for example ResNet-18 compress images to a feature vector of dimension 512 and DistillBert compress data to 768 feature vector for each token. We simply concatenate these embeddings (though another approach would be to convert them a common feature space and then combine them) and pass them through a fully connected neural network. We tried different architectures but used something similar to the authors of the paper in [8], with 3 hidden layers and ReLu activation (details in Experimental Settings section 4).

The fusion layer can be created in various different ways. We can add a fully-connected layer with the text embeddings as input, to bring them to the same dimensionality as the image embeddings. Then, a summation or averaging operation could be done to create a fusion embedding, to be passed in the multimodal architecture ahead. But this would have come at the cost of increased computational complexity leading to increase in training time and space. This would also need to be made robust by backpropagating the final loss gradients to the input text embeddings, to ensure that both modality embeddings end up in the same hyperspace for homogeneity and intuitiveness. Thus, due to compute restrictions, we went ahead and simply concatenated the embeddings of the two different modalities for now.

We use the same evaluation metrics to evaluate the multi-modal classifiers as we did for the text-only and image-only models. They are present in the results section 4.

3. Experimental Settings

We used a 70-30% split for training and test data respectively. We used Distill BERT for multi-class classification and Hindi-BERT models from HuggingFace to create text embeddings. We also use ResNet-18 and Inception v3 to create visual embeddings. All our experiments were run on the COC ICE server using 2 RTX6000 GPUs. The training time for the image only models (25 epochs) is $\sim$63 minutes while it is $\sim$75 minutes for the text only models. We early stop the training process when the loss stabilizes, usually around 15 epochs. For our multi-modal model, the architecture comprises of an input layer (1200 neurons), 3 hidden layers (512, 128, 32 neurons), and an output layer (5 neurons = number of classes in the dataset). Dropout layers were added between each hidden layer to avoid overfitting on the training data. We used Adam optimizer and used ReLU activation. Our codes can be found at this Github Link.

4. Results

The results of our experiments are summarized in the table 1.

Just to see the confusion matrix for one model, we can refer to the image-only model in figure 4 for the performance of the ResNet-18 architecture. We do not show the confusion matrix for Inception architecture here as it is very similar to that of ResNet, something that is also very apparent from the metrics results table.

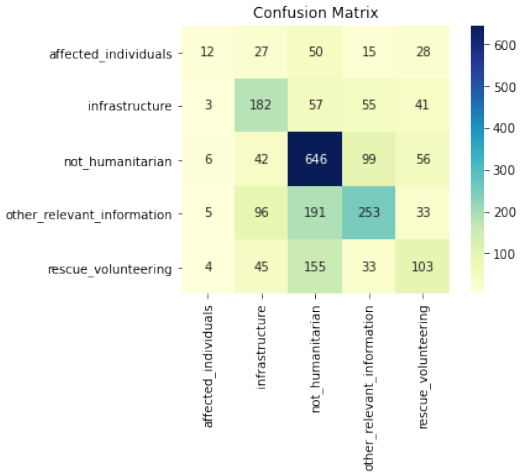

Figure 3. Confusion Matrix for ResNet-18 model

As can be seen on the table 1, the text-only models works best for English language than non-English language (Hindi in our experiment) across all metrics (F_1 score, recall, precision, and accuracy).

While the image-only model has been show to be a good classifier when the images contains key features for the classification task such as for emotion detection, in our experiment to classify humanitarian disasters, it performed the worst in terms of accuracy and other metrics. It is still better than random (given 5 classes). Multi-modal learning boosts classification performance for all the languages considerably with the inclusion of images as an additional modality

	Crisis-MMD Dataset			
	F_1	Precision	Recall	Accuracy
Text-Only Model				
English - DisilBERT	0.73	0.73	0.72	0.80
Hindi - HindiBERT	0.65	0.68	0.64	0.70
Image-Only Model				
Inception v3	0.42	0.42	0.41	0.49
ResNet-18	0.43	0.48	0.43	0.53
Multi-Modal Model				
English + Image	0.76	0.74	0.78	0.83
Hindi + Image	0.71	0.70	0.73	0.80

Table 1. Results Table

when compared against the performance of corresponding text-only classification models.

The main result is that multi-modal learning helps in bridging the gap between the classification performance of English vs. Hindi. As we can observe in the results table above, there is a drastic uplift in the Accuracy for the Hindi multimodal model when compared to the text only model [70% to 80 %]. While this uplift is not very apparent in the English multimodal model as the text only version seems to be saturating at optimal performance itself. This helps us confirm our hypothesis that there is certainly a bias of the English language while pretraining Language models.We also confirm that adding images to these text only models, in a multimodal setting, definitely helps in improving the performance of the downstream classification task.

We also see that using a more advanced architecture like Inception v3 or ResNet-18 provides marginally better embeddings as compared to the VGGnet used in [8]. The original paper fine-tunes the netwrok for 50 epochs to achieve an accuracy of $\sim$50% but we are able to achieve marginally higher accuracies in just 25 epochs, indicating the potential of these networks.

5. Relevant Works and Future Considerations

Our primary motivation of our project comes from [8] which discusses using multimodal data to overcome language disparities in text classification performance. In this paper, we observed the value addition that VGG-based embeddings made to the existing text-based multiclass classification system across three datasets - Crisis Humanitarianism, Fake News and Emotion Detection.

Building on the same, we explored other architectures that may create richer image embeddings. We aimed to replace the VGG-based embeddings with a RestNet architecture [3], since ResNet has shown it can outperform other models considerably on the ImageNet dataset. Building on the VGG approach, it uses shortcut connections that help in training very deep neural networks.

We note that the use of Vision Transformers [2] may improve the multimodal classification tasks with a focus on scalability and performance. Vision Transformers approach image embedding creation in a different way and are shown to outperform ResNet on the ImageNet dataset. Therefore, there are potential benefits from integrating these image models in the future. Additionally, we consider that pretrained multimodal models might also produce good results as has been seen in research and from the comparison of techniques in the following paper[4]

We have focused on enhancing the image embeddings in our work here, but a fair case can also be made for the text embeddings, and the potential in enhancing them under a predetermined computational budget. Language models like mT5 can help enhance these embeddings as well, given they are encoder-decoder architectures, which tend to perform better on downstream tasks in comparison to encoder-only architectures like BERT.

6. Work Division

Each team member contributed equally. The detailed contributions are in table: 2.

Student Name	Contributed Aspects	Details
Suraj and Alberto	Text Only model and Multi-modal model	Used different language models (BERT and MT5) for both English and Hindi language. After that used embeddings from both models to ru multi-modal model
Shreyas and Saksham	Image Only Model	Trained the ResNet-18 and Inception v3 models to create image embeddings and generate accuracy numbers for that

Table 2. Contributions of team members.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] 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. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Yikuan Li, Hanyin Wang, and Yuan Luo. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In

2020 IEEE international conference on bioinformatics and biomedicine (BIBM), pages 1999–2004. IEEE, 2020.

[5] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018.

[6] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47, 2019.

[7] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *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)*, pages 380–385, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Gaurav Verma, Rohit Mujumdar, Zijie J. Wang, Munmun De Choudhury, and Srijan Kumar. Overcoming language disparity in online content classification with multimodal learning, 2022.