# Assignment 3 Report Speech Understanding

Submitted By:

Sonu Shreshtha (P24CS0006)

## 1. Write a review of the approved paper in your own words. Write the following points to write a review:

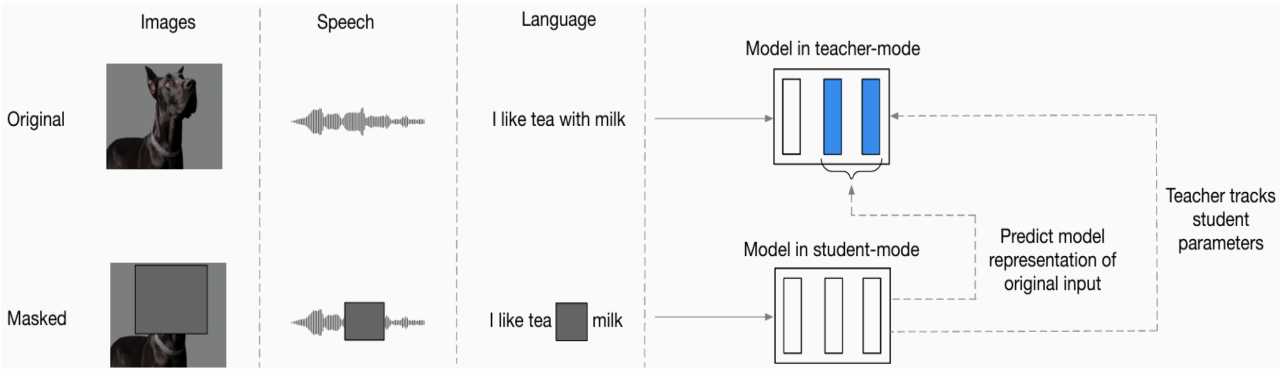
### 1. Title of the paper

data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language.

### 2. Summary of the paper

In this paper, the authors introduce data2vec, a novel single self-supervised learning method that can be effective for vision, speech and language. In contrast to previous SOTA approaches that use modality(text, vision & speech) specific objectives and targets, data2vec inspired from human learning uses the same learning method for all modalities. The core idea introduced in this paper, is to predict latent representations of the full input data based on a masked view of the input in a self-distillation setup using a standard Transformer architecture. Authors argued in this paper, instead of predicting modality-specific targets such as words, visual tokens or units of human speech which are local in nature, data2vec predicts contextualized latent representations that contain information from the entire input. To achieve a unified framework, authors proposed a teacher-student architecture, where the teacher network processes the full input, while the student network predicts these latent representations from a masked input, with teacher parameters being an exponentially moving average of student parameters. Authors performed several experiments across modalities and showed state-of-the-art performance across all three modalities(text, vision and speech) on major benchmarks.

3. Paste a Figure to represent the main architecture (or idea) of the paper (only one Figure from the paper).



The figure shows how data2vec follows the same learning process for different modalities. The model first produces representations of the original input example (teacher mode) which are then regressed by the same model based on a masked version of the input.

4. Strengths of the paper (Write only Technical Strengths)

a. One of the key technical strengths of the paper is that data2vec predicts contextualized latent representations by predicting rich latent representations that contain information from the entire input, unlike prior approaches limited to local targets.

b. Paper introduces self-distillation architecture where the same model is used in two modes, teacher mode with unmasked input and student mode with masked input.  Teacher weights are an exponentially moving average of student weights.

c. Very first time by taking inspiration from human learning a unified self supervised learning approach is proposed that works effectively across three different modalities(text, speech and vision).

## 5. Weaknesses of the paper (Write only Technical Weaknesses, do not write writing mistakes)

a. Proposed approach although unified the learning objective across the modalities but it still uses modality specific input encoders.

b. Proposed approach also uses different masking approaches for different modality(text, speech & vision) rather than a unified masking framework.

c. Although the paper proposed a unified self supervised learning approach for different modalities(text, speech & vision), the authors have not shown any experiment of actual multi-modal training or cross-modal transfer such as audio-visual speech recognition or cross-modal retrieval.

## 6. Minor Questions/Minor Weakness

a. Why have authors in this paper not compared the computational efficiency relative to modality specific approaches?

b. Why did authors in this paper use modality specific input encodes?

c. Why have authors not shown how sensitive is performance to the choice of masking strategy within each modality?

## 7a. Suppose you are an actual reviewer: Provide a few suggestions as a reviewer to the author to improve the weakness of the paper and how the paper’s research idea and claims (if any) can be more strengthened. (Write within 7-8 lines only)

As a reviewer I would like to recommend authors of the paper to explore a unified feature encoder architecture that could work across modalities(text, speech & vision) and a common masking strategy for all the different modalities to strengthen the core idea of the paper, a unified self supervised learning approach that works effectively across different modalities. Further I would recommend authors, they could have shown experimental results on actual multimodal learning using data2vec, with paired data from different modalities such as text and vision etc.

## 7b. What rating would you give to this paper? Provide the rating with proper justification in not more than 3 lines.

As a reviewer of the paper, I would rate this paper as Strong Accept with rating 8 out of 10. This paper presents a significant advancement in self-supervised learning by introducing first of its kind, a unified framework that achieves state-of-the-art results across different modalities. Although this work does not achieve a complete unified self supervised framework for different modalities in all aspects of the pipeline, the core contribution of using contextualized latent representations as targets is very nicely supported by extensive experimentation.

## II. Bonus Question:

Reproduce the results of the paper on any 2 datasets mentioned in the paper.

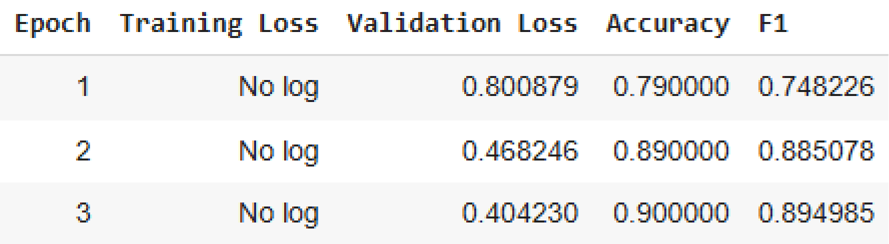
Kindly referred to Jupyter Notebook:
**AssignmentThree_ReproducingResults.ipynb**

I have used a  pre-trained model from Hugging Face to reproduce the paper results. For Text Modality, I have reproduced the paper result for the dataset SST and MNLI on GLEU benchmark task. For Vision Modality, I have reproduced the result of the ImageNet Dataset as shown in the paper.

Take any other dataset not used in the paper, and divide it into train and test split. Fine-tune the model with DoRA and show results on the test set of the chosen dataset, as well as on the 2 datasets you have used in the above part in II(i).

Kindly referred to Jupyter Notebook:
**AssignmentThree_FineTuningDora.ipynb**

I have selected **ag_news** as a new dataset to fine tune the data2vec Model for text modality using DoRA. I have shown the model performance significantly improved after the fine tuning.



| Epoch | Training Loss | Validation Loss | Accuracy | F1 |
|---|---|---|---|---|
| 1 | No log | 0.800879 | 0.790000 | 0.748226 |
| 2 | No log | 0.468246 | 0.890000 | 0.885078 |
| 3 | No log | 0.404230 | 0.900000 | 0.894985 |

# References

1. https://ai.meta.com/blog/ai-self-supervised-learning-data2vec/
2. https://arxiv.org/abs/2202.03555
3. https://huggingface.co/docs/transformers/en/model_doc/data2vec
4. https://proceedings.mlr.press/v162/baevski22a/baevski22a.pdf