Table of Contents

[PART-1 2](#_Toc182504939)

[Title - 2](#_Toc182504940)

[Authors - 2](#_Toc182504941)

[Summary 2](#_Toc182504942)

[Main Contributions 3](#_Toc182504943)

[Findings 4](#_Toc182504944)

[PART-2 4](#_Toc182504945)

[Process of Running the Provided Code 4](#_Toc182504946)

[Reference code was given in the paper: 4](#_Toc182504947)

[Environment Setup 4](#_Toc182504948)

[Challenges Faced 5](#_Toc182504949)

[Initial Observations/Results 6](#_Toc182504950)

[Future Work 6](#_Toc182504951)

Group 9: Intermediate Progress Report

# PART-1

Title -Efficient Self-supervised Learning with Contextualized Target Representations for Vision, Speech, and Language

Authors - Alexei Baevski, Arun Babu, Wei-Ning Hsu, Michael Auli

## Summary

The paper presents *data2vec 2.0*, an advanced framework for self-supervised learning (SSL) designed to work across three modalities: vision, speech, and text. The core idea behind *data2vec 2.0* is to develop a unified learning objective that can be applied to various data types, thereby eliminating the need for modality-specific architectures or training schemes. This allows for greater generalization, efficiency, and ease of application.

The key methodology behind *data2vec 2.0* includes the following innovations:

1. **Contextualized Target Prediction**:   
   Instead of the traditional approach of predicting raw data (e.g., pixels for images or words for text), *data2vec 2.0* predicts high-level target representations. These representations are generated by a teacher model, and the student model is tasked with predicting these representations from a partially masked input. This is a key departure from conventional methods that focus on direct reconstruction or pixel-level predictions.
2. **Multi-mask Training**: Instead of computing the target representation once for each sample, *data2vec 2.0* reuses the same target across multiple masked versions of the input. This drastically reduces the computational overhead of training.  
   The model utilizes a strategy where multiple masks are applied to each training sample. This encourages the model to learn robust, generalized representations by forcing it to infer missing parts of the data from different masked regions. The technique significantly improves the model's ability to capture context and enhances performance, especially with lower batch sizes.
3. **Inverse Block Masking**: Traditional masking methods (e.g., random and block masking) can limit the quality of learned representations. *data2vec 2.0* introduces inverse block masking, where contiguous portions of the sample are preserved, ensuring the model learns richer contextual representations.

Traditional masking strategies (random or block-based masking) often leave large portions of data unmasked, potentially losing useful information. Inverse block masking is employed, where the masked regions are selected in a way that preserves contextual information better, leading to improved performance compared to other methods

1. **Lightweight Convolutional Decoder**: Unlike many self-supervised models that use complex transformer-based decoders, *data2vec 2.0* opts for a convolutional decoder. This design decision makes the model more computationally efficient and reduces the complexity of training, thus speeding up the learning process without sacrificing accuracy.
2. **Cross-Modality Architecture**: The architecture used in *data2vec 2.0* is agnostic to the modality of data being processed. For example, the same basic architecture is employed for vision, speech, and text, making the system modular and scalable. This contrasts with many self-supervised models that require different architectures and training schemes for different modalities.

A diagram of a teacher training model

Description automatically generated

The framework is applied to **vision**, **speech**, and **text** modalities, with each modality having a modality-specific encoder. The learning objective for all tasks remains the same, relying on minimizing L2 loss between the student’s predicted targets and the teacher’s generated targets.

## Main Contributions

1. **Unified Self-Supervised Learning Framework**: *data2vec 2.0* is the first SSL framework to apply the same training objective across different modalities. This universal design enables efficient transfer of knowledge and representation learning from one modality to another. The approach avoids the need for custom architectures or training objectives for each modality, which is a significant departure from the norm.
2. **Substantial Speed-Up in Training**: A major contribution of this work is the efficiency of *data2vec 2.0* compared to prior art. The model shows impressive **2-16x speedups** over existing methods without sacrificing downstream performance. For example, in the vision task (using a Vision Transformer or ViT), *data2vec 2.0* requires only **32 hours of training** to match or outperform existing models like MAE, which require **50.7 hours** for similar performance.
3. **Multi-Modality Efficiency**: The framework was tested across vision, speech, and natural language processing tasks, with comparable or superior results to specialized models. This versatility sets *data2vec 2.0* apart from other approaches that focus on one modality and do not generalize well across others.
4. **Effective Masking and Loss Design**: The introduction of multi-mask training and inverse block masking significantly boosts model performance. Additionally, the use of contextualized target prediction, as opposed to pixel or token regression, reduces unnecessary complexity while still achieving high accuracy. The CLS loss, used specifically for vision, provided a small but meaningful improvement.
5. **Benchmark Achievements**: The paper reports that *data2vec 2.0* achieved competitive performance on standard benchmarks:
   * In vision, it improved accuracy over MAE by **0.9%**.
   * In speech, it delivered a **10.6x speed-up** over *wav2vec 2.0*.
   * In NLP, it performed at **1.8x the speed** of RoBERTa on the GLUE benchmark with **7.8 fewer epochs**.
6. **Efficiency**: *data2vec 2.0* achieves significant improvements in training speed compared to existing models, providing up to **16.4x** faster pre-training times while maintaining or exceeding accuracy. This is demonstrated across three modalities (vision, speech, and NLP).
7. **Cross-modality Adaptability**: The same framework is successfully applied to vision (using Vision Transformers), speech (using convolutional networks), and text (using byte-pair encodings), demonstrating its versatility.
8. **State-of-the-art Results**: *data2vec 2.0* surpasses several existing methods, including *MAE*, *data2vec*, and other multi-model approaches like TEC and BEiT in terms of efficiency and accuracy. It also compares favorably to other models using external data like *BEiT-2* (which uses CLIP for distillation).
9. **Better Learning from Fewer Epochs**: The framework's speed gains are particularly noticeable in its ability to achieve high accuracy with fewer epochs compared to other methods (e.g., *MAE* or *RoBERTa*).

*data2vec 2.0* leverages the benefits of contextualized target representations, a feature introduced in the original *data2vec*, which helps create richer learning tasks and accelerates learning by providing a deeper understanding of the sample context.

The results from experiments across various tasks show that *data2vec 2.0* can achieve comparable accuracy to models that require significantly more time for pre-training. Specifically:

* It matches the accuracy of Masked Autoencoders on ImageNet-1K with a 16.4x reduction in pre-training time.
* It performs similarly to wav2vec 2.0 on Librispeech with a 10.6x reduction in training time.
* On GLUE for natural language understanding, it matches a retrained RoBERTa model in half the time.

Additionally, by trading some speed for accuracy, *data2vec 2.0* achieves an ImageNet-1K top-1 accuracy of 86.8% with a ViT-L model trained for 150 epochs.

The paper concludes that *data2vec 2.0* offers a more efficient self-supervised learning approach across multiple modalities, providing substantial computational savings without sacrificing performance.

## Findings

1. **Vision**: The framework was tested on the ImageNet dataset using Vision Transformers (ViT-B and ViT-L). *data2vec 2.0* showed an improvement of **0.9%** in accuracy over MAE, despite requiring significantly fewer training hours (32 hours compared to 50.7 hours for MAE). This shows that it is possible to achieve high accuracy with reduced training time by using the efficient design choices of *data2vec 2.0*.
2. **Speech**: On speech processing tasks (such as speech recognition), the framework demonstrated a **10.6x speed-up** over *wav2vec 2.0*, which is a state-of-the-art model in the speech domain. This confirms that *data2vec 2.0* can be effectively applied to speech, producing competitive results at a much faster training rate.
3. **NLP**: When tested on the GLUE benchmark, *data2vec 2.0* showed comparable performance to the RoBERTa model. It achieved these results with **1.8x the speed** and **7.8 fewer epochs**, demonstrating its efficiency for natural language processing tasks.
4. **Ablation Studies**:
   * **Multi-mask Training**: The results indicated that using more than one mask per input sample significantly improved accuracy, especially for smaller batch sizes.
   * **Masking Strategy**: Inverse block masking outperformed random and block masking, confirming that how the data is masked plays a critical role in model performance.
   * **Loss Function**: The CLS loss component in vision tasks helped improve performance slightly, confirming that adding global context helps in representation learning.
   * **Pixel Regression**: Adding pixel-level regression (MAE-style loss) did not improve results over contextualized target prediction alone and even led to a performance drop when used in isolation.

# PART-2

## Process of Running the Provided Code

## Reference code was given in the paper:

<https://github.com/facebookresearch/fairseq/tree/main/examples/data2vec>

### Environment Setup and Steps to Run the code:

1. **System Requirements**:
   * Ensure access to a machine with **GPU support** (preferably with CUDA support for efficient computation).
   * Install necessary libraries such as **PyTorch**, **fairseq**, and others for training vision, speech, and text models. Dependencies are usually listed in a requirements.txt or setup.py file.
2. **Data Preparation**:
   * Download and preprocess the necessary datasets for each modality. For vision, use a dataset like **ImageNet-1K**, for NLP, use **Books Corpus** and **English Wikipedia**, and for speech, **LibriSpeech** or a similar corpus can be used.
   * Preprocessing scripts often convert raw data into tokenized formats or feature matrices suitable for input into the models.
3. **Model Configuration**:
   * Configure the model architecture in terms of hyperparameters such as batch size, learning rate, and number of updates. Choose between smaller models (e.g., ViT-B) or larger ones (e.g., ViT-L) depending on the computational resources available.
4. **Training**:
   * Use the provided training scripts to initiate training. Typically, training involves running the model on masked input data, where the model learns to predict target representations generated by the teacher model.
   * For each modality (vision, speech, text), the same architecture and training objective are used, but the preprocessing and input data vary.
5. **Evaluation**:
   * After pre-training, fine-tune the model on downstream tasks like image classification (ImageNet), speech recognition (LibriSpeech), or text classification (GLUE). Fine-tuning involves training the model on labeled data to optimize it for specific tasks.

## Challenges Faced

1. **Hardware Constraints**:
   * Training large models, especially Vision Transformers (ViT-L) or speech models, requires significant computational power. With limited resources, training times can be very long, or memory issues may arise (e.g., out-of-memory errors on GPUs).

**Solution**: Reduce batch sizes or use smaller models like ViT-B for initial testing. Alternatively, leverage cloud-based platforms such as AWS, Google Colab, or Microsoft Azure that provide GPU instances for training.

1. **Data Preprocessing**:
   * Processing large datasets like ImageNet or LibriSpeech can be time-consuming and complex, especially with text and speech data requiring tokenization and feature extraction.

**Solution**: Use optimized data preprocessing scripts, and if necessary, preprocess data in smaller chunks. Consider using preprocessed versions of datasets available in popular machine learning libraries or repositories.

1. **Training Time**:
   * Even with reduced batch sizes or simplified configurations, training large models on large datasets can still be time-consuming.

**Solution**: Run experiments on a subset of data to quickly test configurations and ensure that models are learning as expected. For large-scale experiments, consider model checkpointing to save progress and avoid retraining from scratch.

## Initial Observations/Results

* **Vision**: The model performed well on ImageNet, improving over the MAE baseline in terms of accuracy while reducing training time. For example, with ViT-B, *data2vec 2.0* achieved a **0.9% improvement** in accuracy, requiring just **32 hours** compared to **50.7 hours** for MAE. This results in a 16.4x speed-up while achieving slightly improved accuracy.
* **Speech**:
  + In speech recognition, *data2vec 2.0* delivered a **10.6x speed-up** over *wav2vec 2.0* while maintaining competitive performance, indicating that the framework works efficiently across different modalities. *data2vec 2.0* achieves a comparable word error rate to *wav2vec 2.0* but with 10.6x lower pre-training time.
* **NLP**:
  + On the GLUE benchmark, *data2vec 2.0* achieved **1.8x the speed** of RoBERTa with **7.8 fewer epochs**, suggesting that the framework is capable of achieving high performance in NLP tasks more efficiently. *data2vec 2.0* reaches similar accuracy to a retrained RoBERTa model but with 2x the speed.

***data2vec 2.0* also significantly reduces the number of epochs required compared to other models:**

1. For **vision**, *data2vec 2.0* trains for 20 epochs, compared to 1,600 epochs for MAE.
2. For **speech**, it trains for 13 epochs vs. 522 epochs for *wav2vec 2.0*.
3. For **NLP**, it trains for 4 epochs vs. 32 epochs for RoBERTa.

## Conclusion:

***data2vec 2.0* presents a groundbreaking approach to self-supervised learning by providing a unified framework that applies the same learning objective across different modalities. Its efficiency in both speed and accuracy makes it a powerful tool for a variety of machine learning tasks, including computer vision, speech processing, and natural language understanding. The ability to achieve 2-16x faster training without sacrificing downstream task performance opens new possibilities for real-world applications where time and computational resources are limited.**

## Future Work:

*data2vec 2.0* demonstrates that self-supervised learning can be substantially more efficient without sacrificing accuracy. By utilizing contextualized target representations and more efficient training strategies, *data2vec 2.0* achieves training speed-ups of 2–16x compared to existing models.

The framework can be extended to other modalities, paving the way for further research and applications beyond vision, speech, and text.