Enhancing Medical Image Analysis with Multi-Task Learning Using Visual Transformers

Aleksandra Vatian[1][0000-0002-5483-716X], Ivan Tomilov[1][0000-0003-1886-2867], Mikhail Gritskikh[1][0000-0002-1361-6037], Natalia Dobrenko[1][0000-0001-6206-8033], Anton Kharytonov[1][0000-0002-8826-8583], Yuliya Valitova[1][0000-0001-5345-5461] and Natalia Gusarova[1][0000-0002-1361-6037]

[1]1ITMO University, 49 Kronverksky av., St. Petersburg, Russia
alexvatyan@gmail.com

Abstract. Medical image analysis often relies on models optimized for specific tasks, such as classification or detection. However, this single-task approach limits the utilization of shared features across tasks and becomes particularly inefficient when data availability is limited. In this paper, we investigate the potential of multi-task learning (MTL) with Visual Transformers to optimize both classification and reconstruction tasks, especially when training data is scarce. Using datasets like BRATS (binary classification of brain tumor presence on MRI slices) and ultrasound images of muscles (for identifying pathologies such as Myopathy, Myelopathy, and Polyneuropathy), we evaluate MTL against standalone and pre-training paradigms. Results indicate that MTL significantly enhances model performance, particularly in classification tasks, by leveraging shared representations and improving attention mechanisms. Our findings demonstrate that MTL mitigates the challenges of limited data availability by effectively transferring knowledge between tasks, making it a valuable strategy for medical imaging applications.

Keywords: Multi-task learning, Visual Transformers, Medical Imaging, MRI, Ultrasound, Brain Tumor Classification, Muscle Pathology Detection, Limited Data Learning.

1 Introduction

1.1 Background and Motivation

Medical imaging plays a pivotal role in clinical decision-making by providing non-invasive insights into pathological conditions [1]. In recent years, deep learning-based approaches have significantly improved performance in diagnostic tasks such as tumor classification, anomaly detection, and disease progression assessment. However, these models often rely on single-task learning (STL), which assumes large, well-annotated datasets specific to each task [2].

Multi-task learning (MTL) presents a promising alternative by jointly optimizing multiple objectives, leveraging shared representations to improve generalization and sample efficiency. By incorporating multiple tasks, MTL mitigates overfitting and provides improved robustness when training data is limited [5][6].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_16

Transformers, particularly Visual Transformers such as DINO, have gained attention for their ability to model long-range dependencies within images. Recent studies have demonstrated their superiority over CNN-based architectures in tasks such as segmentation and classification [9][10].

Unlike prior works that focus on multi-task optimization for medical imaging [2], this study incorporates MIM-based reconstruction [3] alongside classification, demonstrating how shared feature representations can improve diagnostic accuracy. By directly comparing MTL with STL and pretraining paradigms, we provide empirical evidence that MTL not only achieves higher classification accuracy but also reduces overfitting in data-limited settings.

Our study evaluates MTL's effectiveness across both small and large dataset sizes, revealing that MTL's advantages become more pronounced when training data is scarce.

1.2 Objectives

The primary objective of this study is to evaluate the effectiveness of MTL using Visual Transformers in medical imaging applications with limited data. Specifically, we aim to:

- Quantify the impact of MTL on classification task by comparing its performance against STL and pre-training approaches.
- Evaluate performance of MTL for small-sized datasets compared to large-sized datasets.
- Analyze attention map visualizations to assess the effectiveness of learned feature representations and their relevance to pathology-specific regions.

By addressing these objectives, we aim to contribute to the development of more efficient deep learning models for medical image analysis, particularly in resource-constrained scenarios where large-scale annotated datasets are unavailable.

2 Theoretical Foundations

2.1 Multi-Task Learning Framework

The concept of multi-task learning is rooted in the idea that jointly optimizing multiple objectives can improve model efficiency and performance. MTL is formally defined as an optimization problem where a model learns a set of related tasks simultaneously [4][8]. The total loss function for an MTL model is expressed as:

$$L = \sum_{i=1}^{T} \lambda_i L_i$$

Where T is the number of tasks, λ_i is the weight for task i, L_i is the task-specific loss.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_16

2.2 Loss Functions

Loss functions play a crucial role in optimizing deep learning models, especially in the context of multi-task learning (MTL), where different tasks often require distinct objective functions. In this study, we employ a combination of loss functions tailored to classification and masked image modeling (MIM) tasks.

MIM involves predicting masked pixel values from an input image, making regression-based loss functions a natural choice. One of the most used loss functions for this task is the Mean Squared Error (MSE), which is defined as:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

For the classification task, where the model predicts whether a given medical image belongs to a certain pathology class, we use the Binary Cross-Entropy (BCE) loss:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}[\, y_i * log(\hat{y}_i) + (1 - y_i) * log(1 - \hat{y}_i)]$$

2.3 Comparison of Pretraining and Multi-Task Learning Loss Functions

A key challenge in MTL is balancing different loss functions, as optimizing for one task may degrade performance on another. Traditional pretraining approaches optimize a model for a general task before fine-tuning it on the target task.

In pretraining, the optimization objective is:

$$\theta^{pretrain} = \arg min\theta L_{pretrain(\theta)}$$

where θ are the model parameters. During fine-tuning for the target task:

$$\theta^{fine-tune} = \arg min\theta L_{target(\theta)}$$

This sequential process can lead to catastrophic forgetting because the gradient $\nabla\theta L_{target}$ may overwrite the features learned from $L_{pretrain}$.

In MTL, the optimization is simultaneous:

$$\theta^{MTL} = \arg min\theta\big(\lambda_1 L_{pretrain(\theta)} + \lambda_2 L_{target(\theta)}\big)$$

where λ_1 and λ_2 balance the importance of the pretraining and target tasks. This prevents catastrophic forgetting because both tasks contribute to the shared gradient:

$$\nabla_\theta L_{MTL} = \lambda_1 \nabla\theta\, L_{pretrain} + \lambda_2 \nabla\theta\, L_{target}$$

Thus, features useful for both tasks are retained.

2.4 Analysis of Regular Training and Multi-Task Learning Loss Optimization

In MTL, the model learns a shared representation $h_{shared} = f(x;\, \theta_{shared})$, where h_{shared} is the shared feature representation learned by the model, x is the input, and θ_{shared} are the parameters of the shared layers with the objective:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_16

$$L_{MTL} = \sum_{i=1}^{T} \lambda_i L_i\left(h_{shared}, \theta_i^{task}\right)$$

Where θ_i^{task} are the task-specific parameters. While in MTL, the shared parameters θ_{shared} are updated using gradients from all tasks:

$$\nabla_{\theta_{shared}} L_{MTL} = \frac{\lambda_i}{n_i} \sum_{k=1}^{n_i} \nabla_{\theta_{shared}} \ell_{i(x_k, y_k)} + \frac{\lambda_j}{n_j} \sum_{j \neq i} \sum_{k=1}^{n_j} \nabla_{\theta_{shared}} \ell_{j(x_k, y_k)}$$

MTL increases the effective dataset size for the shared parameters.

$$n_{effective} = n_i + \sum_{j \neq i} \frac{\lambda_j}{\lambda_i} n_j$$

Tasks with larger effective dataset size reduce the overall variance, stabilizing the updates to shared parameters.

A key benefit of MTL is the reduction in variance of the parameter updates. The variance of the gradient updates for the shared parameters θ_{shared} is given by:

$$Var\left(\nabla_{\theta_{shared}} L_{MTL}\right) = \frac{\lambda_i^2}{n_i^2} \; Var\left(\nabla_{\theta_{shared}} \ell_i\right) + \sum_{j \neq i} \frac{\lambda_j^2}{n_j^2} Var\left(\nabla_{\theta_{shared}} \ell_j\right)$$

MTL leads to lower variance, provided that auxiliary tasks contribute sufficient training samples. This stabilizes gradient updates and prevents abrupt weight changes, improving model robustness.

In scenarios where the tasks are correlated, the total variance must also account for covariance terms:

$$Var(L_{MTL}) = \frac{\lambda_i^2}{n_i^2} Var(L_i) + \sum_{j \neq i} \frac{\lambda_j^2}{n_j^2} Var\left(L_j\right) + \sum_{i \neq j} \frac{\lambda_i \lambda_j}{n_i n_j} Cov\left(L_i, L_j\right)$$

If tasks are positively correlated ($Cov > 0$), total variance increases, meaning tasks reinforce each other's learning signals. In situation when tasks are negatively correlated ($Cov < 0$), total variance decreases, meaning tasks act as a regularizer.

3 Methodology

3.1 Dataset Selection and Preparation

In this study, we utilize two distinct medical imaging datasets: ultrasound images of muscles and MRI scans from the BRATS dataset. These datasets were chosen for their relevance in evaluating multi-task learning (MTL) performance across different imaging modalities and pathologies.

The ultrasound dataset consists of grayscale scans of muscle tissue, annotated with labels indicating the presence or absence of muscular diseases such as Myopathy, Myelopathy, and Polyneuropathy. Given the inherent variability in ultrasound imaging, this dataset poses challenges in feature extraction and classification, making it an ideal testbed for evaluating the effectiveness of a multi-task model.

The BRATS dataset, on the other hand, comprises MRI scans that are widely used in brain tumor research. The classification task in this study focuses on binary tumor

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_16

detection - distinguishing between MRI slices with and without tumor presence. The complexity of MRI images, combined with the necessity for precise localization of tumor regions, adds another layer of difficulty to the learning process.

Both datasets undergo preprocessing steps such as normalization, resizing, and augmentation to ensure consistency across training samples. Given the different imaging characteristics, specific transformation pipelines are applied to retain essential medical features while standardizing the input format for the model.

3.2 Model Design and Learning Strategy

The model architecture is based on a Visual Transformer (DINO), which serves as the core feature extractor. Unlike conventional CNN-based approaches, transformers enable long-range dependencies and capture global and local patterns, which are particularly beneficial for medical imaging applications. Two task-specific heads are incorporated into the model.

Masked Image Modeling (MIM) Head, responsible for reconstructing missing image patches, enabling the model to learn structural and contextual information from medical images [7].

Classification Head, designed to predict disease labels based on learned feature representations, enabling accurate identification of pathological conditions.

The integration of these task-specific heads facilitates knowledge sharing between classification and reconstruction tasks, leveraging the strengths of both to improve overall model generalization.

3.3 Training Approach and Optimization

To ensure robust learning, we adopt a multi-task training strategy where both classification and reconstruction losses are optimized simultaneously. The classification task employs Binary Cross-Entropy (BCE) Loss for tumor and muscle pathology detection. The reconstruction task utilizes Mean Squared Error (MSE) Loss.

Bellow you can see an example of MIM task evaluation on ultra-sound muscles images dataset.

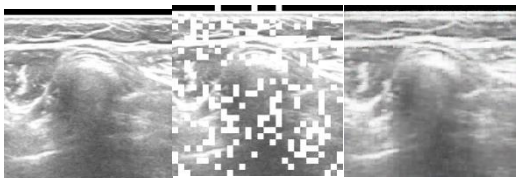

Fig. 1. Evaluation of MIM *task for ultra-sound image of muscle*.

3.4 Evaluation Metrics

To gauge the effectiveness of the proposed approach, a comprehensive evaluation framework is implemented. For classification performance, we rely on the Area Under the Receiver Operating Characteristic Curve (ROC AUC).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_16

For the reconstruction task, we evaluate model output using quantitative metrics such as MIM loss (MSE) and qualitative assessments of visual fidelity. By analyzing the reconstructed images, we can determine the model's ability to recover missing details and preserve structural consistency in medical imaging scenarios.

Through this methodology, we aim to demonstrate the benefits of multi-task learning in improving both classification accuracy and image reconstruction quality.

4 Experiments and Results

4.1 Evaluating Multi-Task Learning Performance

Our experiments were conducted using two distinct medical imaging datasets: ultrasound images of muscles and MRI scans from the BRATS dataset. The primary objective was to evaluate the effectiveness of Multi-Task Learning (MTL) compared to pre-training and single-task learning approaches.

Table 1. Ultra-sound images of muscles pathology classification.

Training strategy	ROC-AUC
STL	0.72
Pre-training	0.64
MTL	0.88

Table 2. MRI scans classification.

Training strategy	Sampling	ROC-AUC	ROC-AUC lower bound	ROC-AUC upper bound
STL	All data	0.71	0.69	0.72
MTL	All data	0.71	0.71	0.71
STL	Under sampled	0.70	0.67	0.73
MTL	Under sampled	0.71	0.68	0.74

The results demonstrated that MTL improves classification performance, particularly in scenarios where the training dataset is relatively small. A key observation was that reducing the number of training samples led to better results on the hold-out test set for MTL. This finding aligns with our theoretical understanding of MTL.

On the other hand, pre-training the model on a general task before fine-tuning it for classification resulted in decreased performance. These findings highlight the advantage of training classification and reconstruction objectives simultaneously within an MTL framework rather than relying on a sequential pre-training approach.

4.2 Interpreting Attention Mechanisms

To gain further insights into model behavior, we analyzed attention maps generated by the Visual Transformer backbone. In MTL-trained models, attention maps showed

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_16

increased focus on diagnostically relevant regions within medical images. This suggests that MTL enhances the model's ability to identify critical features, reinforcing the benefits of shared feature representations across tasks.

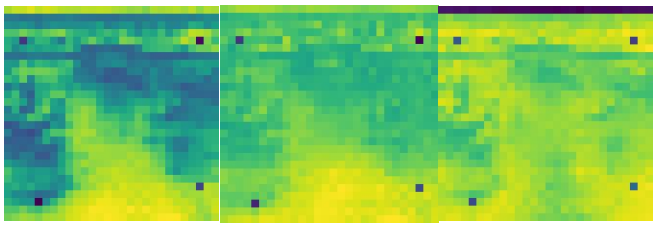

Fig. 2. Attention map visualization for MTL with ultra-sound images.

In contrast, models trained using pre-training or single-task learning exhibited more diffuse attention distributions, often highlighting irrelevant regions. These results further confirm the effectiveness of MTL in directing model focus toward the most informative parts of medical images, improving interpretability and diagnostic accuracy.

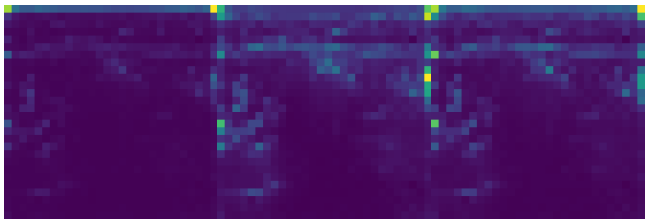

Fig. 3. Attention map visualization for STL with ultra-sound images.

Overall, our experimental findings provide strong empirical support for the theoretical advantages of MTL, demonstrating its ability to enhance generalization and improve attention-based representations in medical image analysis.

5 Conclusion and Future Work

This study investigated the impact of Multi-Task Learning (MTL) in medical image analysis using Visual Transformers, specifically in the classification of ultrasound muscle images and MRI-based brain tumor detection. Our findings reinforce the theoretical advantages of MTL, particularly in cases where the availability of labeled data is limited. Pre-training the model before fine-tuning led to a decline in performance. Additionally, attention map analysis revealed that MTL-trained models focused more effectively on diagnostically relevant regions, indicating an improvement in feature learning and interpretability.

Future research should explore more advanced techniques to refine the MTL framework. One potential direction is the investigation of adaptive task-weighting mechanisms, which could dynamically balance classification and reconstruction losses based

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_16

on learning progress. Further, expanding this study to multi-class and multi-label classification tasks would provide a more comprehensive understanding of MTL's advantages in diverse medical imaging settings. Another promising direction involves evaluating MTL on additional imaging modalities, such as CT or PET scans, to assess its broader applicability across different diagnostic domains.

By addressing these directions, future work can further solidify MTL as a viable solution for improving medical data analysis, advancing AI-driven diagnostic tools that are both efficient and clinically relevant.

Acknowledgments. This work was supported by Russian Science Foundation, Grant № 23-11-00346..

References

1. Yang, K., Dong, X., Tang, F., Ye, F., Chen, B., Liang, S., Zhang, Y., Xu, Y.: A Transformer-Based Multi-Task Deep Learning Model for Simultaneous T-Stage Identification and Segmentation of Nasopharyngeal Carcinoma. Frontiers in Oncology 14, 1377366 (2024)
2. Suigu Tang, Xiaoyuan Yu, Chak Fong Cheang, Yanyan Liang, Penghui Zhao, Hon Ho Yu, I Cheong Choi, Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images, Computers in Biology and Medicine, Volume 157, 2023, 106723.
3. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H. (2022). SimMIM: A Simple Framework for Masked Image Modeling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9653–9663
4. Zhou Y, Chen H, Li Y, Liu Q, Xu X, Wang S, et al. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. Med Image Anal. (2021).
5. Marquet, T., & Oswald, E. (2023). A Comparison of Multi-task Learning and Single-task Learning Approaches. *In* A. Das, S. Bhasin, & D. Jap (Eds.), *AIHWS 2023: Artificial Intelligence in Hardware Security* (pp. 123–138). Springer.
6. Xin, Z., Wu, C., Liu, D., Gu, C., Guo, J., Hua, J.: Enhancing CT Image Synthesis from Multi-Modal MRI Data Based on a Multi-Task Neural Network Framework. arXiv preprint arXiv:2312.08343 (2023)
7. Zhang Y, Li H, Du J, Qin J, Wang T, Chen Y, et al. 3D multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. IEEE Trans Med Imaging. (2021).
8. Haque, A., Imran, A.-Z., Wang, A., Terzopoulos, D.: Generalized Multi-Task Learning from Substantially Unlabeled Multi-Source Medical Image Data. arXiv preprint arXiv:2110.13185 (2021)
9. Tagnamas, J., Ramadan, H., Yahyaouy, A. et al. Multi-task approach based on combined CNN-transformer for efficient segmentation and classification of breast tumors in ultrasound images. Vis. Comput. Ind. Biomed. Art 7, 2 (2024)
10. Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D.: Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review. Medical Image Analysis 91, 103000 (2024)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_16