ODELIA

Open Consortium for Decentralized Medical Artificial Intelligence

HORIZON-HLTH-2021-CARE-05-02

Deliverable D1.2

GUIDANCE DOCUMENT ON RUNNING SL IN CLINICAL ENVIRONMENTS

Lead beneficiary	TUD
Author(s)	Oliver Saldanha
Reviewer	Wouter Veldhuis (UMCU)
Dissemination level	PU
Type	R
Delivery date	19 December 2024 (revision)

ODELIA is funded by the European Union's Horizon Europe Framework under Grant Agreement 101057091

Funded by the European Union

TABLE OF CONTENTS

DISCLAIMER

SUMMARY

This progress report outlines several significant achievements in the ODELIA D1.2 phase.

INTRODUCTION

Swarm Learning, an innovative approach to machine learning in healthcare, expands upon the foundational principles of federated learning by incorporating a decentralizing element, thereby eliminating the requirement for a singular, authoritative entity. This strategy leverages the power of Artificial Intelligence (AI), the immediacy of edge computing, and the robust security framework provided by blockchain technology. In summary, Swarm Learning represents a forward-thinking, privacy-centric Machine Learning (ML) model that decentralizes the learning process. At its core, the Swarm Learning framework utilizes the computational resources located at or in close proximity to the distributed data points to execute the Machine Learning algorithms that are responsible for model training. The security paradigm of blockchain technology is employed to facilitate safe and secure peer learning and knowledge sharing. In the Swarm Learning paradigm, the training process of the ML models takes place at the edge of the network, where the data is most fresh and where timely, data-driven decisions are of paramount importance. This decentralization facilitates a unique information-sharing environment where only the distilled insights from the ML model training, not the raw data, are shared amongst collaborating ML peers. This not only bolsters data security but also enhances the privacy of the data involved, making Swarm Learning an optimal choice for sensitive environments like clinical settings.

HARDWARE REQUIREMENTS OF THE MACHINE

The following hardware requirement is recommended for the MRI Brest cancer tumor prediction:

- **RAM**: At least 32 GB of RAM is required, but ideally, 64 GB of RAM should be used for optimal performance.
- **CPU:** A minimum of 8 CPU cores is required, but it is recommended to have 16 CPU cores for better efficiency.
- **GPU:** An NVIDIA GPU with a minimum of 24 GB of RAM is required, but for improved performance, a GPU with 48 GB of RAM is recommended.
- **Storage:** The absolute minimum storage requirement is 4 TB, but for better data management and storage capacity, it is recommended to have 8 TB of storage.

By demonstrating that medical report generation can be accomplished with these lightweight hardware specifications, it highlights the feasibility and accessibility of the task.

SOFTWARE REQUIREMENTS OF THE MACHINE

Operating system: To successfully run the Swarm Learning Environment on the Linux-qualified on Ubuntu 20.04 Operating system, the following recommendations and compatibility information should be considered:

- Supported Ubuntu Versions: The Swarm Learning Environment has been tested and confirmed to work on the following Ubuntu versions: Ubuntu 20.04 LTS, Ubuntu 22.04.2 LTS, and Ubuntu 20.04.5 LTS.
- Avoid Experimental Releases: It is advised to avoid using experimental releases of Ubuntu beyond the LTS 20.04 version, as they may lead to unsuccessful operation of the swop node.

Container hosting platform: HPE Swarm Learning is optimized for Docker 20.10.5, ensuring compatibility with IPv4. It's advisable to run Docker as a non-root user for security, and configuring network proxy settings is made straightforward. These steps enhance the functionality and security of HPE Swarm Learning in collaborative machine learning setups.

Machine Learning Framework: We used PyTorch 1.5-based Machine Learning models implemented but also qualify with Keras 2.9.0 (TensorFlow 2 backend) and using Python3.

NETWORK REQUIREMENTS OF THE MACHINE

To establish a secure swarm learning environment in a hospital, several critical factors must be considered, particularly network and infrastructure settings. First and foremost, it is imperative that hospitals maintain a distinct network environment for swarm learning to thwart any unauthorized access. This separation is pivotal in reducing security risks, ensuring that the swarm learning infrastructure remains isolated from the hospital's clinical network. To fortify the security posture, the implementation of robust firewalls and intrusion detection systems is paramount, offering protection against potential cyber threats. Moreover, adopting a mesh-based VPN architecture guarantees the encryption of all data, reinforcing its safety and privacy.

- A minimum of one or a maximum four open TCP/IP ports in each node. All swarm nodes must be able to access the ports of every other node. For more information on port details that must be opened.
- Stable internet connectivity to download Swarm Learning package and Docker images.

DATA PRIVACY AND SECURITY

Data Anonymization: Patient data used for swarm learning should be anonymized and de-identified to comply with healthcare privacy regulations, such as HIPAA (in the United States) or GDPR (in Europe).
Access Control: Implement strict access controls and authentication mechanisms to ensure that only authorized personnel can access and use patient data for swarm learning.

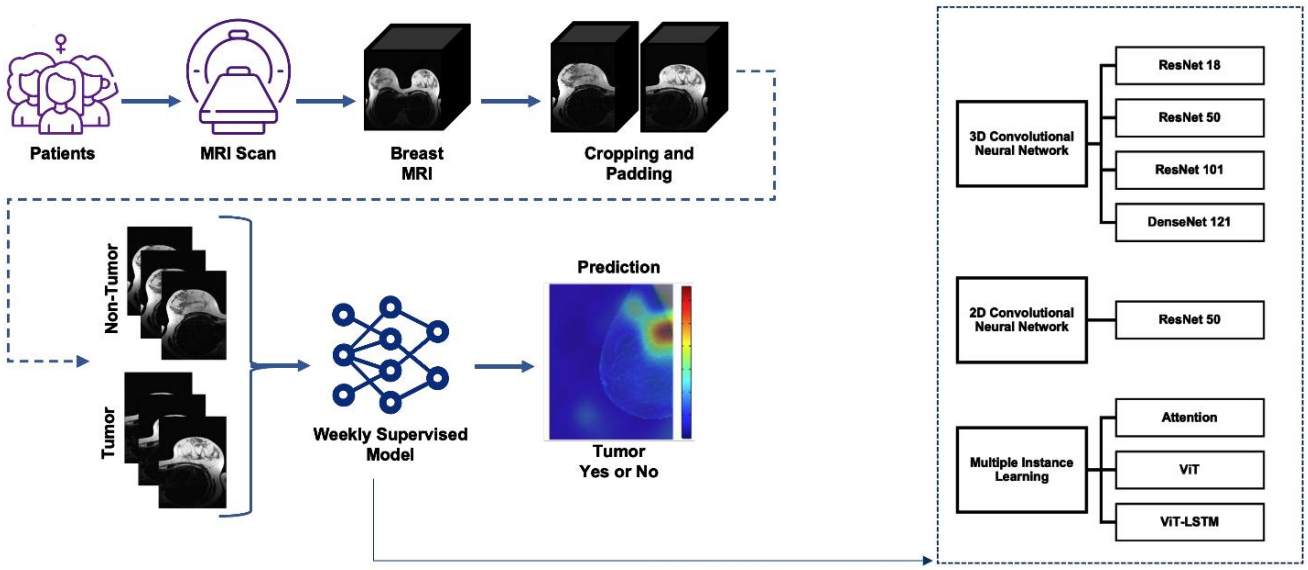
CONTACT AND ROLES

When setting up a swarm learning environment in a hospital, it's essential to identify and establish contacts for each site involved in the collaboration. Here's a breakdown of relevant contacts and their roles:

Research Contact: This contact is typically a healthcare professional or researcher responsible for coordinating research efforts at the hospital. They oversee data collection, data quality, and research objectives.
IT/Technical Contact: The IT/Technical contact is responsible for managing the technical infrastructure, network, and software required for swarm learning. They ensure that the IT systems are secure, scalable, and compliant with privacy regulations.
Security or Chief Information Security Officer (CISO): The CISO or security contact is responsible for ensuring the security of the hospital's digital assets, including patient data. They oversee cybersecurity measures, risk assessments, and incident response planning.
Data Protection Officer (DPO): In compliance with regulations like GDPR, a Data Protection Officer is responsible for ensuring that the hospital's data processing activities, including those related to swarm learning, comply with data protection laws. They oversee data privacy and compliance efforts.

DATA PREPROCESSING

The same preprocessing pipeline was applied to all datasets in this study. The preprocessing comprises two main steps. In the initial preprocessing step, the DICOM files are converted into NIFTI format, which facilitates the distinction and storage of images as pre-contrast and first post-contrast sequences. Following this, the difference between the first post-contrast and the pre-contrast images is calculated to produce subtraction images, known as sub-contrast sequences. After these initial transformations, individual cropping or padding is applied to the left and right breast volumes to suit the requirements of our model, which processes a single breast volume at a time. We are using intensity-based localization rather than manual segmentation or a separate AI algorithm for segmentation. We crop the height to 256 pixels, It calculates a threshold to find the foreground (presumably the breast area) and adjusts the crop dynamically to include this area, using a margin from the top. Each breast volume is then globally labeled for malignancy (yes/no) according to the classifications provided by the Duke, USZ, CAM, MHA, or UKA datasets. The images are subsequently resampled to achieve a uniform resolution of 256 x 256 x 32 voxels. By simplifying the problem of tumor detection into such a binary classification problem on the whole volume of a breast in an MRI image, we enable the problem to be analyzed with a range of weakly supervised prediction methods.



Data anonymization

In our approach to secure sensitive health care data, Digital Imaging and Communications in Medicine (DICOM) files are transformed into NIfTI (Neuroimaging Informatics Technology Initiative) files. By converting DICOMs to NIfTI format, DICOM tags, which may contain sensitive information, are inherently excluded from the resulting files.

Furthermore, folder names containing patient names are changed to remove any identifying information. Instead of patient names, folders are assigned pseudonyms—unique identifier names that can be matched to find the original patient if necessary. This matching table is stored separately in a secure location. Storage and processing take place exclusively on access-restricted computers to which only authorized personnel have access.

Additionally, the pseudonymized data itself remains within the institution, ensuring further control over access and use.

This method provides an effective solution for maintaining patient privacy while ensuring the usability of medical imaging data for our research within ODELIA.

Step by step Instruction to run swarm learning

After all the above requirement of data and setup are met the follow the following instructions:

1. Prepare the PC and Install Ubuntu on Windows

- Start by shrinking the Windows volume to make space for Ubuntu. Open System and Security > Windows Tools > Computer Management, then go to Disk Management. Right-click the Windows (C:) drive, choose Shrink Volume, specify the desired size (e.g., 500,000 MB), and complete the process.

- Reboot the PC, go to Advanced Startup Options, and select the USB device with the Ubuntu installer to boot from.
- Install Ubuntu by following the on-screen instructions provided by the Ubuntu installer.

2. Set Up Ubuntu with Essential Tools

- After Ubuntu is installed, set up some essential tools:
  - Install Visual Studio Code for code editing.
  - Install TeamViewer for remote access using the Linux version.
  - Use the Additional Drivers feature in Software Updater to select and install the NVIDIA 525 driver.
  - Install Git to manage repositories.
  - Set up SSH for secure remote connections by installing both client and server components.

3. Create a User Account and Set Up the Repository

- Create a user account named `swarm` and grant it administrator (sudo) privileges.
- Set up a dedicated directory on the system where Swarm Learning resources will be stored and grant open permissions to facilitate access.
- Clone the Swarm Learning repository from the GitHub link provided in your setup documentation.

4. Install and Verify the CUDA Environment

- After setting up the CUDA environment and installing NVIDIA drivers, verify the installation by checking for outputs that confirm the GPU is recognized by the system. This verification step ensures the environment is properly set up to handle the necessary computational loads.

5. Configure the Swarm Learning Environment

- Define specific environment variables for the system:
  - Set <sentinel_ip> to the IP address provided for your institution by the VPN server.
  - Set <host_index> to the name of your institution (for example, TUD, VHIO, etc.).
  - Choose a <workspace_name> based on the radiology model you'll be working with (e.g., odelia-breast-mri or marugoto_mri).
- Configure Data Path for HPE Swarm Setup:
  - Specify the data path in the HPE swarm setup by defining the environment variable <data_path>, which points to the location of your local dataset. This path must be set for each node to ensure data is accurately located and used during training.
  - Example: Set <data_path> to /mnt/data/Duke_MRI or similar, aligning it with each institution's local data storage setup. This ensures that data is properly "plugged in" to the swarm learning environment, allowing seamless access and consistency across training nodes.

6. Run Initial Automation Setup Scripts

- Initiate automation setup scripts to ensure all required software is installed and that the environment is configured according to the Swarm Learning setup requirements.
- The automation process involves three stages:
  - Initialization: This script checks for required software and sets up the necessary foundational environment.
  - Server Configuration: This step configures the server with Swarm Learning specifics, using the sentinel IP and institution details.
  - Final Setup: This completes the setup, preparing the Swarm Learning environment for training, specifying details such as the number of peers and training epochs if required.

7. Launch and Manage Swarm Learning Nodes

- Start the Swarm Learning nodes:
  - Run a Swarm Network (Sentinel) node to handle network communication and management.
  - Run a Swarm SWCI node, which initiates and manages the training tasks. Typically, only the sentinel host will initiate the SWCI node to manage training across nodes.
- To monitor training progress, check the latest logs directly in the system.
- To stop the Swarm Learning nodes, use a specific command that allows you to stop individual node types or stop all nodes at once.

8. Access Results and Training Logs

- After training, you can review the results, saved models, and detailed logs. These can be found within the directory structure under the workspace name you specified, where all data, logs, and model artifacts are stored for easy access.

CONCLUSION

These are the guidelines that must be adhered to when implementing swarm learning within a clinical setup. This guide outlines the major steps needed to set up and manage Swarm Learning for clinical applications, following the recommended setup procedures in the document.