WHITEPAPER

Foundation models in Aiforia

Learn how Aiforia incorporates foundation models into the semantic segmentation, object detection, and instance segmentation workflows in Aiforia® Create, Aiforia's versatile tool for developing AI models for image analysis in digital pathology.

Author:
Yrjö Häme, PhD, Director of Aiforia Platform

aiforia®
AI for image analysis

TABLE OF CONTENTS

1. Foundation models in digital pathology

Foundation models in image analysis refer to large, pre-trained models that can be applied across a wide range of tasks. These models are typically trained on vast datasets leveraging self-supervised training and vision transformer architectures. Foundation models serve as versatile building blocks that are fine-tuned or adapted for specific applications, such as object detection, segmentation, and classification while reducing the need for task-specific training.

Foundation models significantly reduce the need for large-scale labeled datasets and complex training pipelines for each new task. Pretraining on large datasets provides improved feature representations and results in an ability to generalize and adapt to a wide range of image analysis tasks. These benefits make foundation models more versatile and powerful in image analysis than traditional convolutional neural networks (CNNs) trained end-to-end.

The field of digital pathology has recently seen increasing interest in foundation models. Several publications have introduced foundation models specific to digital pathology, such as Atlas[1] and Virchow[2]. One of the most prominent applications of foundation models for digital pathology has been the generation of automatic descriptions from images, e.g., PathChat[3]. There are also general-purpose vision models available as open source that have value in digital pathology, for example models using DINOv2[4].

CNNs

- Consist of convolutional layers that excel at detecting local patterns and features in images
- Typically trained with random initialization for a specific task, and a CNN trained for one task may not perform well on another
- Require large, labeled datasets for supervised training
- Available for Aiforia® Create users

Foundation models

- Pre-trained models that serve as versatile building blocks, adaptable to a wide range of image analysis tasks
- Provide improved feature representations
- Reduce the need for task-specific training, large-scale labeled datasets, and complex training pipelines
- Available for Aiforia® Create users in 2025

2. Large foundation models require computational capacity

Powered by increased computational capacity and data availability, data sets used to train foundation models have been gradually growing, as have the foundation models' sizes. While a large model size typically improves the ability of the model to generalize for different tasks, it also comes with a high computational cost, both in training and deploying the model.

To ensure the scalability of the resulting task-specific models, it is necessary to consider the tradeoffs between model size and the need for broad generalizability in the application. While some applications require training sets with a high level of variability collected from different domains, increasing attention is being paid to optimizing foundation models for specific applications. This means better-curated training sets, which results in more cost-effective models.

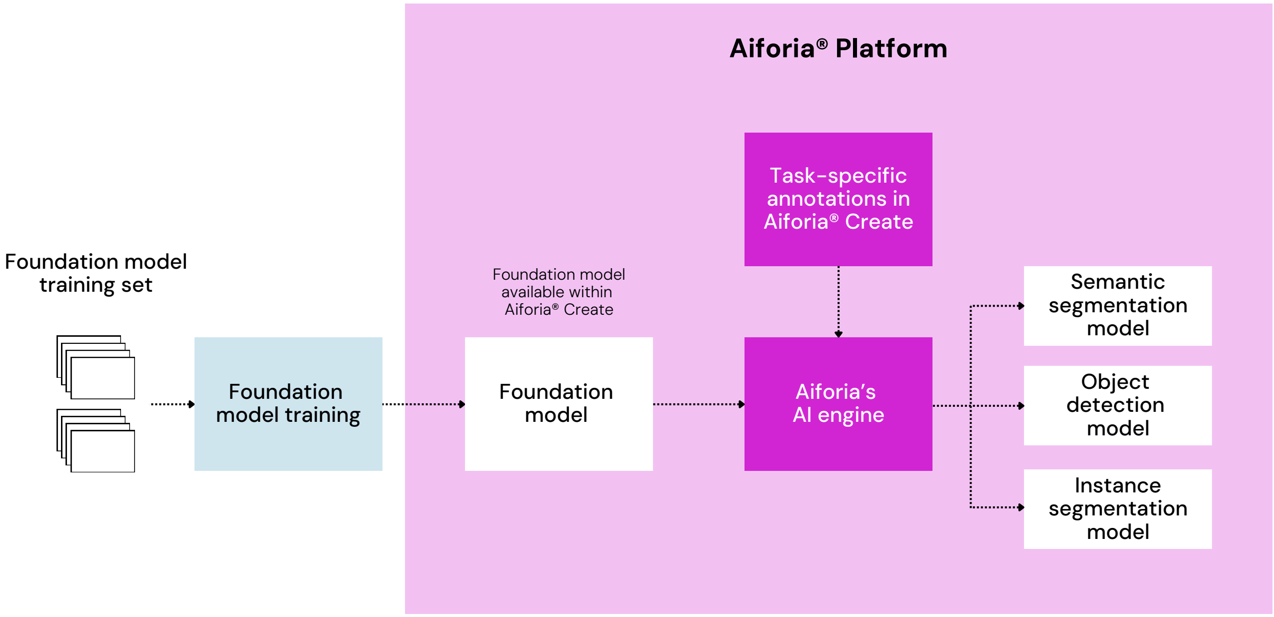
In pathology, foundation models are most useful in data-scarce settings and do not consistently outperform task-specific models when ample labeled training data is available. For clinical applications, thorough validation and task-specific training remain essential, suggesting a need for a hybrid approach that leverages the strengths of both foundation and end-to-end learning.

3. Foundation models in Aiforia

Building on existing functionality and competence, Aiforia incorporates foundation models into the semantic segmentation, object detection, and instance segmentation workflows available in Aiforia® Create. While many other applications leveraging foundation models rely only on slide-level labels, Aiforia is committed to enabling the most precise pixel-level tasks with the latest AI technology.

With the updated interface, the user can select between the available AI engine versions for each project separately. The list of engine versions includes the CNN-based engines and the new engine that enables training with foundation models. The project parameters switch automatically to ones applicable to the selected engine, and the AI engine selection is specific to each project.

The AI model training and evaluation workflow remains unchanged from the AI engine selection. The user can seamlessly switch between engine versions and leverage available annotations for existing projects. Creating new AI projects and annotations works exactly the same way as before.


Aiforia® Platform
Foundation model training set
Foundation model training
Foundation model available within Aiforia® Create
Foundation model
Task-specific annotations in Aiforia® Create
Aiforia's AI engine
Semantic segmentation model
Object detection model
Instance segmentation model

AI model training workflow in Aiforia® Create using Foundation models

4. Performance of a foundation model on an example use case

To demonstrate the value of training using foundation models in Aiforia, experiments were performed on an AI project with realistic, high-quality annotations. The project represents a semantic segmentation task for which Aiforia is typically used, namely **Gleason grading** with five classes: G3, G4, G5, Benign, and Background (representing everything not included in the other classes).

4.1 Data and study questions

The aim was to compare regular CNNs initialized randomly with models trained using foundation models to examine the following qualities:

1. Ability to generalize with different amounts of task-specific training annotations
2. Model accuracy with respect to the number of training iterations
3. Computational cost in inference

The number of available annotations was in the tens of thousands, while the area covered by the annotations typically represents only a small fraction of each annotated whole slide image. The set of images and the associated annotations were split using stratified sampling into two sets of equal size: training and testing.

Four different training set sizes were generated to enable testing with varying amounts of data. The full training set (representing 50% of the entire data set in the study) was further divided into sets of 50% (of the training set), 25%, and 10%, also using stratified sampling. All error metrics mentioned below are calculated only on the test set. The foundation model used for these experiments was trained with DINOv2[4], which is available as open source.

4.2 Generalization, accuracy, and training iterations

To demonstrate the ability to generalize, we examine how the number of task-specific annotations affects the training outcome. The F1 score (harmonic mean between precision and sensitivity) is used as the metric.

By providing different sizes of training sets, we see that training using the foundation model achieves **14%-26% better test results** than regular CNNs with random initialization for the same set of training annotations. This difference is larger when fewer annotations are used in training. Furthermore, training using the foundation model results in a higher F1 score when trained with 10% of annotations than regular CNNs with the full training set.

Observing the number of iterations required to train the model, the CNNs with random initialization start at a relatively low accuracy level and require 10,000-20,000 iterations to reach full accuracy. On the other hand, trainings using the foundation model have a better starting point, as expected, and are close to their full accuracy already after some thousands of training iterations, and the subsequent accuracy gains are modest.

4.3 Computational cost in inference

One of the most important factors to consider when designing a production-level workflow relying on foundation models is the computational cost incurred by the resulting task-specific AI model. While specific applications have their individual throughput requirements, in clinical digital pathology, a whole-slide analysis must typically be completed in a matter of minutes to keep the per-slide costs reasonable. In some workflows, the analysis is triggered on demand and may impose restrictions on the maximum accepted wait time. The combination of the large image size and high-resolution processing precludes the largest models.

The computational cost was examined for the test set analyses for both models. The average analysis times for the regular CNN model for Gleason grading were approximately 50%-65% of the corresponding execution time of the model trained using the foundation model.

This duration represents the inference task performed by the image analysis engine and excludes pre- and post-processing steps required for the entire workflow.

The relatively small difference in execution time between the two models is due largely to the modest size of the foundation model used in these experiments. A larger foundation model would significantly increase the analysis execution time.

5. Summary

Aiforia® Create provides users with the power of foundation models at their fingertips by incorporating them into semantic segmentation, object detection, and instance segmentation. Annotations in existing AI projects can be seamlessly used with the new foundation model engine. Our early experiments with a realistic semantic segmentation task indicate that foundation models will have a significant impact when building AI models in Aiforia® Create, particularly for difficult tasks.

Training using foundation models is expected to reduce the number of required training iterations and achieve better performance with less training data compared to regular CNNs trained with random initialization. While the improved performance comes with a longer duration in inference, the computational cost can be kept reasonable by selecting small foundation models, possibly ones that are optimized for the task at hand.

Aiforia® Create will launch foundation model access to users in 2025. For more information, please contact your Aiforia contact person or customer support.

References

1. Alber, M. et al. (2025). A novel pathology foundation model by Mayo Clinic, Charite, and Aignostics. arXiv preprint. https://doi.org/10.48550/arXiv.2501.05409

2. Vorontsov, E. et al. (2023). Virchow: A million-slide digital pathology foundation model. arXiv preprint. https://doi.org/10.48550/arXiv.2309.07778

3. Lu, M. Y. et al. (2024). A multimodal generative AI copilot for human pathology. Nature, 634(8033), 466-473. https://doi.org/10.1038/s41586-024-07618-3

4. Oquab, M. et al. (2023). Dinov2: Learning robust visual features without supervision. arXiv preprint. https://doi.org/10.48550/arXiv.2304.07193

Aiforia® Create – The power of AI in your hands

Learn more: aiforia.com/aiforia-create

aiforia®
AI for image analysis