TOMAAT: volumetric medical image analysis as a cloud service

Fausto Milletari
NVIDIA
fmilletari@nvidia.com

Johann Frei
Technische Universität München
johann.frei@tum.de

Seyed-Ahmad Ahmadi
German Center for Vertigo and Balance Disorders (DSGZ)
Ludwig-Maximilian Universität München
Ahmad.Ahmadi@med.uni-muenchen.de

1 Motivation

In the last few years, deep learning techniques have revolutionized medical image analysis for tasks such as segmentation or classification, with performances that sometimes surpassed those of humans. Despite the popularity and the performances of recent methods, it is currently difficult to access state-of-the-art deep learning algorithms designed for medical tasks in order to run comparison, integrate them in existing workflows and use them for experimentation. In this work, we propose TOMAAT[1], an open-source framework and architecture that enables (i) researchers to make their algorithms accessible by packaging their models and data processing pipelines into a prediction service with customized interface; (ii) technical and clinical users to access prediction services through a client that is straightforward to adopt and operate; (iii) researchers to add their algorithms to the public TOMAAT announcement service, which communicates the list of available solutions to end-users, who can select the method they need for specific predictions.

We provide an example client implementation for the popular 3D Slicer framework [1]. An important related contribution in this direction is represented by DeepInfer [2], which proposes to make deep learning models running in docker containers available through a purpose-built module available as an extension of 3D Slicer. Differently than our framework, algorithms made available by DeepInfer are executed on the local machine. We see several advantages to a cloud-based approach:

1) Facilitated use for clinical partners: Setting up a deep learning environment, even in a containerized form, is often prohibitively complex for non-experienced users. Instead, we propose to keep client solutions thin by handling complex setups on the server-side.

2) Comparability of methods: Not all methods are published open-source, re-implementations require considerable effort and are not always correct. TOMAAT is designed to facilitate public usage and comparison of volumetric image analysis methods. Researchers willing to share their method as a service can sign it up to the public TOMAAT announcement service.

3) Hardware out-sourcing: High-dimensional medical images often require powerful hardware (GPUs), which can be handled more efficiently on the server-side.

4) Maintenance and updates: Models and pre-/post-processing steps can be updated on the server-side, always providing users with the latest prediction performance without additional efforts.

5) Intranet clouds: Clinical data safety often requires local networks to be kept separate from the internet. All components of TOMAAT can be easily set up inside a closed-off network, allowing internal and secure sharing of models and data across research groups of the same institute.

[1] Homepage: http://tomaat.cloud

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

2 Method

The TOMAAT framework can be divided in server, client, and announcement service. Their relationship is illustrated in Figure 1. Here, we provide a brief outline of each component, for detailed explanations, we refer the reader to our online manuscript [2].

Server: The server component provides a prediction endpoint in the network, meaning that it can be accessed by clients to obtain predictions through a network connection (local or remote) using the interface provided by TOMAAT and the HTTP protocol. The prediction service may be hosted by a deep learning research lab, on a machine equipped with adequate hardware. Multi-GPU setups are supported by TOMAAT, with the server automatically thread-locking the next available GPU for the client request.
Utilities: TOMAAT provides useful server-side wrappers for pre-processing, inference and post-processing. Pre-processing procedures include e.g. transformations and resampling of arbitrary-sized volumes to match the format expected by the neural network. The inference component allows wrapping of any deep learning frameworks, with an already included example wrapper for arbitrary Tensorflow models. Post-processing functions can e.g. restore original volume dimensions before serving the prediction result to the client.
Interface communication: Depending on the task and on the approach used, the number and type of arguments required to run the forward step may vary. In order to support any typology of model and task, TOMAAT server is designed to allow researchers to define custom interfaces for their services, based on standard data elements. For example, a method that requires multiple input volumes as an input, communicates to the client its needs before the actual prediction request takes place.

Client: The client component allows straightforward usage of remote and local prediction endpoints for the end-user (e.g. clinicians). A client interface can be implemented in any language and integrated in any platform with limited overhead through the interface of TOMAAT. Clients aiming to be general purpose and aiming to interact with any prediction endpoint, must be capable of handling the element-based interface specified by each server. This can be done by creating a modular interface that support the presence of multiple elements. We have developed a general-purpose client module with these characteristics for 3D Slicer [1]. Through this module we allow users to interact with our framework. We allow clients to (i) realize direct POST requests to servers whose URL (local or remote) are already known, to (ii) obtain the list of currently available public servers from any announcement service (in the clinical intranet, or the public default: http://tomaat.cloud:8001/discover), (iii) obtain results from prediction endpoints.

Announcement service: An announcement service maintains an updated list of all the public servers that have been registered as valid prediction endpoints, for example providing a segmentation algorithm for a specific anatomy and modality. A server needs to announce itself through a POST request containing essential information about the prediction endpoint, such as its URL, description, modality, task, anatomy, and other information as per documentation. Crucially, in order to successfully announce a new prediction endpoint, a secret API-key, unique to each server, must be supplied with the POST request. In this way we limit the capability of announcing services publicly over the web only to trusted developers. The list of currently available services is retrieved by the client by a GET request to the announcement service. Information about currently available servers will then be returned to the client together with a unique identifier for each server, enabling further direct connection between client and server. The server communicates the required inputs for its service to the client, which dynamically builds a suitable UI for the end-user to perform the prediction.

3 Experimental evaluation

In order to demonstrate the capabilities of our framework, we tested three different prediction endpoints for image segmentation: left ventricle in 4D cardiac ultrasound with a 3D FCNN inspired by V-Net [3]; prostate in MRI with another V-Net based model [3]; brain tumor in multi-variate MRI with the model provided by Wang *et al.*[4]. The use cases have noteworthy differences, highlighting the convenience of the modularity and cloud-based nature of TOMAAT. We compare the performances of TOMAAT in terms of response latency time in three scenarios: when the server is available through

[2] https://arxiv.org/abs/1803.06784

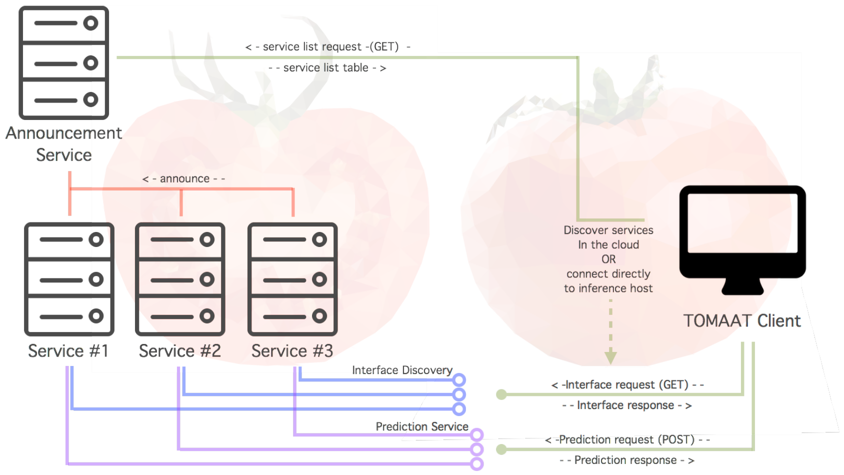
< - service list request -(GET) -
- - service list table - >
Announcement Service
< - announce - -
Service #1
Service #2
Service #3
Discover services In the cloud OR connect directly to inference host
TOMAAT Client
Interface Discovery
< -Interface request (GET) - -
- - Interface response - >
Prediction Service
< -Prediction request (POST) - -
- - Prediction response - >

Figure 1: Schematic representation of the architecture of the TOMAAT framework.

a different types of connection. For the first two methods, we also report the net GPU compute time (which does not include pre- and post- processing), as baseline. The results are shown in Table 1.

Table 1: TOMAAT latency for different scenarios.

Method	LAN	DSL	4G
3D V-Net Heart	3.93s (GPU: 1.25s)	6.21s (GPU: 1.24s)	8.70s (GPU: 1.25s)
3D V-Net Prostate	3.40s (GPU: 0.38s)	7.42s (GPU: 0.38s)	13.53s (GPU: 0.389s)
Brain tumor	20.12s	41.16s	82.49s

4 Conclusion

In this paper we have presented TOMAAT, an open-source framework to deploy and access deep learning models for 3D medical image analysis in an easy and straightforward manner. Our aim is to speed up and improve research in this field by giving technical and especially clinical end-users the possibility to experiment, compare and use state-of-the-art deep learning models for tomographic segmentation. Our framework allows to deploy models as cloud services accessible through a simple HTTP interface and add them to a public list of available servers.

References

[1] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

[2] Alireza Mehrtash, Mehran Pesteie, Jorden Hetherington, Peter A Behringer, Tina Kapur, William M Wells, Robert Rohling, Andriy Fedorov, and Purang Abolmaesumi. Deepinfer: open-source deep learning deployment toolkit for image-guided therapy. In *Medical Imaging 2017: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10135, page 101351K. International Society for Optics and Photonics, 2017.

[3] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[4] Guotai Wang, Wenqi Li, Sebastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. *arXiv preprint arXiv:1709.00382*, 2017.