August 2023

Explainable Server Cooling Schedule Prediction Using Machine Learned Model Conditioned on Multimodal Data

D Shin

Explainable Server Cooling Schedule Prediction Using Machine Learned Model Conditioned on Multimodal Data

ABSTRACT

Server cooling management frameworks that utilize neural networks are trained with a stream of multimodal sensor data, However, model predictions from such models lack explainability. This disclosure describes a fine-tuned model conditioned on multimodal sensor data to perform cooling schedule prediction. The model can also provide explainability by responding to natural language queries. The approach utilizes a transformer decoder architecture, with the model conditioned on multimodal sensor data from a natural server environment. The conditioning enables localizing the understanding of a language model to the specific context of use for server cooling scheduling decisions.

KEYWORDS

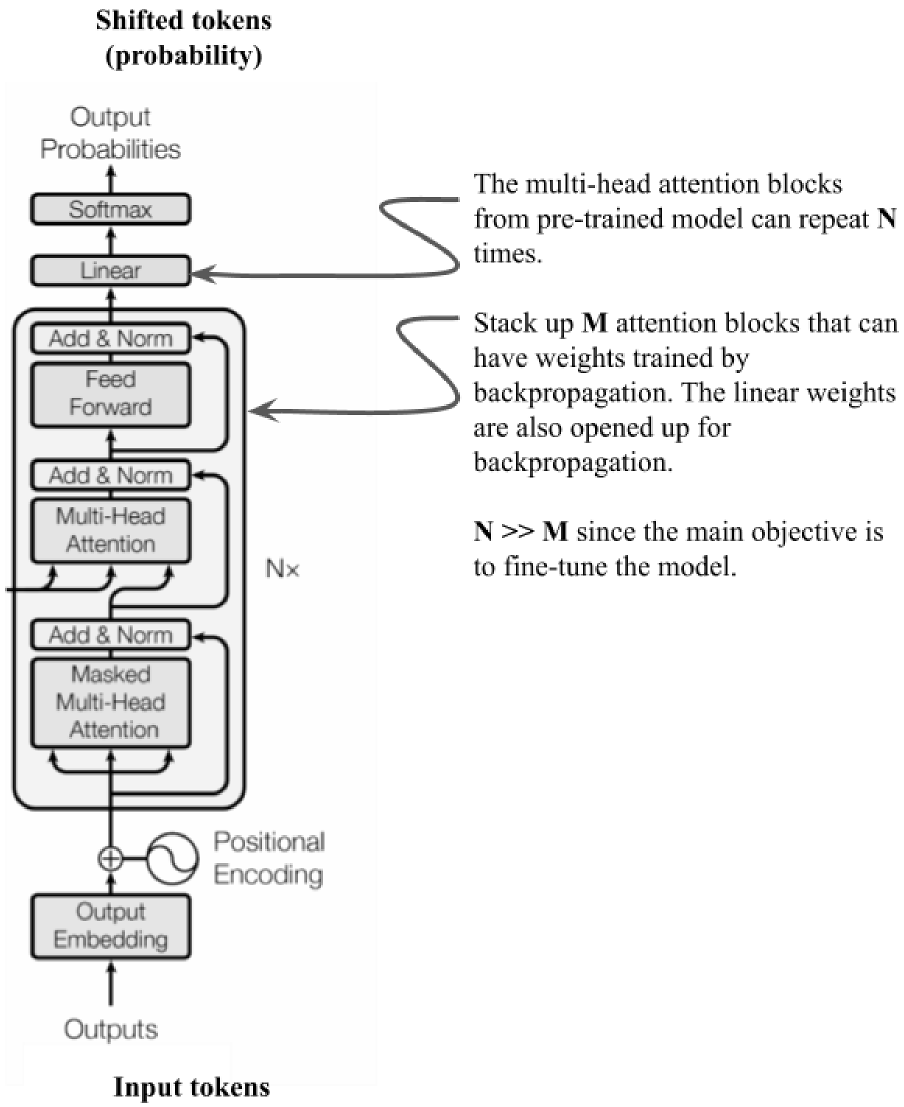
- Server cooling
- Multimodal data
- Cooling schedule
- Transformer decoder architecture
- Model conditioning
- Model explainability
- Localized understanding
- Large language model (LLM)

BACKGROUND

Efficient cooling of servers is an important task. There are some server cooling management frameworks that utilize neural networks. The general approach is to train the model with a stream of multimodal sensor data, e.g., spatially distributed in-room temperature readings from a server room, that have a correlation with future cooling schedules. The model is trained to perform this prediction. However, such models operate simply with a data-in-data-out principle and lack explainability. The cooling schedules predicted by the model do not include context about various schedule decisions. Also, there is no opportunity for a user to interact and probe the model recommendations and why they are optimal.

DESCRIPTION

This disclosure describes a language interface designed for cooling schedule management. The interface enables users to query and understand the design choices recommended by a model. The approach utilizes a transformer decoder architecture, conditioning it using custom multimodal sensor data available in a natural server environment. The conditioning is to localize the understanding of the language model to the specific context of use for server cooling scheduling decisions.



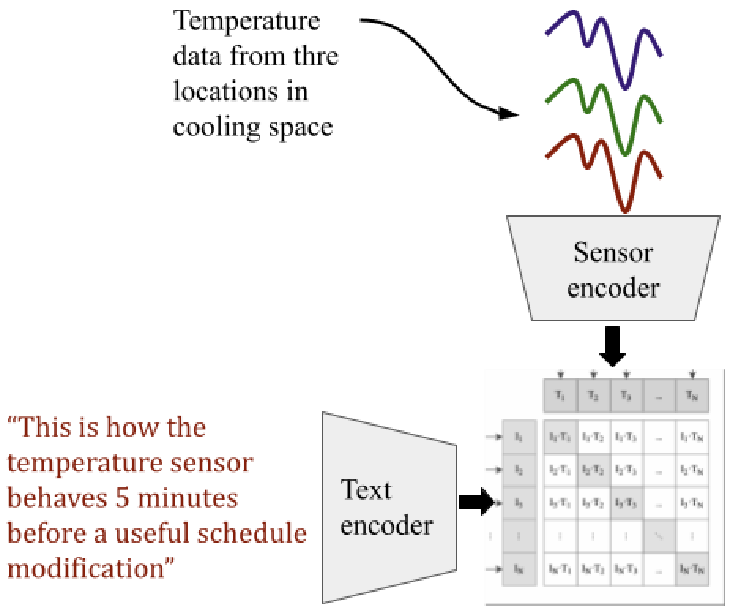
The multi-head attention blocks from pre-trained model can repeat **N** times.

Stack up **M** attention blocks that can have weights trained by backpropagation. The linear weights are also opened up for backpropagation.

N >> M since the main objective is to fine-tune the model.

Fig. 1: Fine-tuning pre-trained models

Per techniques of this disclosure, a pre-trained model is fine-tuned to the specific application (predicting cooling schedule). As illustrated in Fig. 1, attention blocks are stacked up, with weights trained by backpropagation. Linear weights are also opened up for backpropagation. Multi-head attention blocks from a pre-trained model can be repeated N times. Since the main objective is to fine-tune the model, N is chosen to be substantially larger than M.


Temperature data from thre locations in cooling space
Sensor encoder
"This is how the temperature sensor behaves 5 minutes before a useful schedule modification"
Text encoder

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Fig. 2: Learning bidirectional mapping between sensor data and text

As illustrated in Fig. 2, a text encoder can be utilized to encode natural language text. A sensor encoder can encode corresponding sensor data, e.g., temperature data from three locations in the space that is being cooled. A bidirectional mapping between sensor encoding and text encoding is learned by borrowing successful visual language model architectures. This allows tokenizing sensor data and injecting it into the transformer decoder at runtime.

Once the training of the models is done, during monitoring time, a user can simply chat with the cooling schedule management interface. The user query is tokenized and concatenated with the tokens obtained from sensor-language mappings. This enables the transformer model to have the full context of the space (as indicated by the sensor data) and the user need (as indicated by the query). The model can then generate a reliable and explainable answer.

For example, the user can ask probing questions such as "why does the cooling schedule management strategy look different between events A and B?" A response that provides the reasoning, e.g., “This is due to the observed temperature prior to schedule change events A and B.” In this manner, a fine-tuned model conditioned on multimodal sensor data can perform cooling schedule predictions and can also provide explainability by responding to natural language queries.

CONCLUSION

This disclosure describes a fine-tuned model conditioned on multimodal sensor data to perform cooling schedule prediction. The model can also provide explainability by responding to natural language queries. The approach utilizes a transformer decoder architecture, with the model conditioned on multimodal sensor data from a natural server environment. The conditioning enables localizing the understanding of a language model to the specific context of use for server cooling scheduling decisions.