Pruning and Quantizing MobileNet trained on DREAMT Dataset

CUAI Summer Conference DA 1팀

2025.07.06

발표자: 박준우

Team

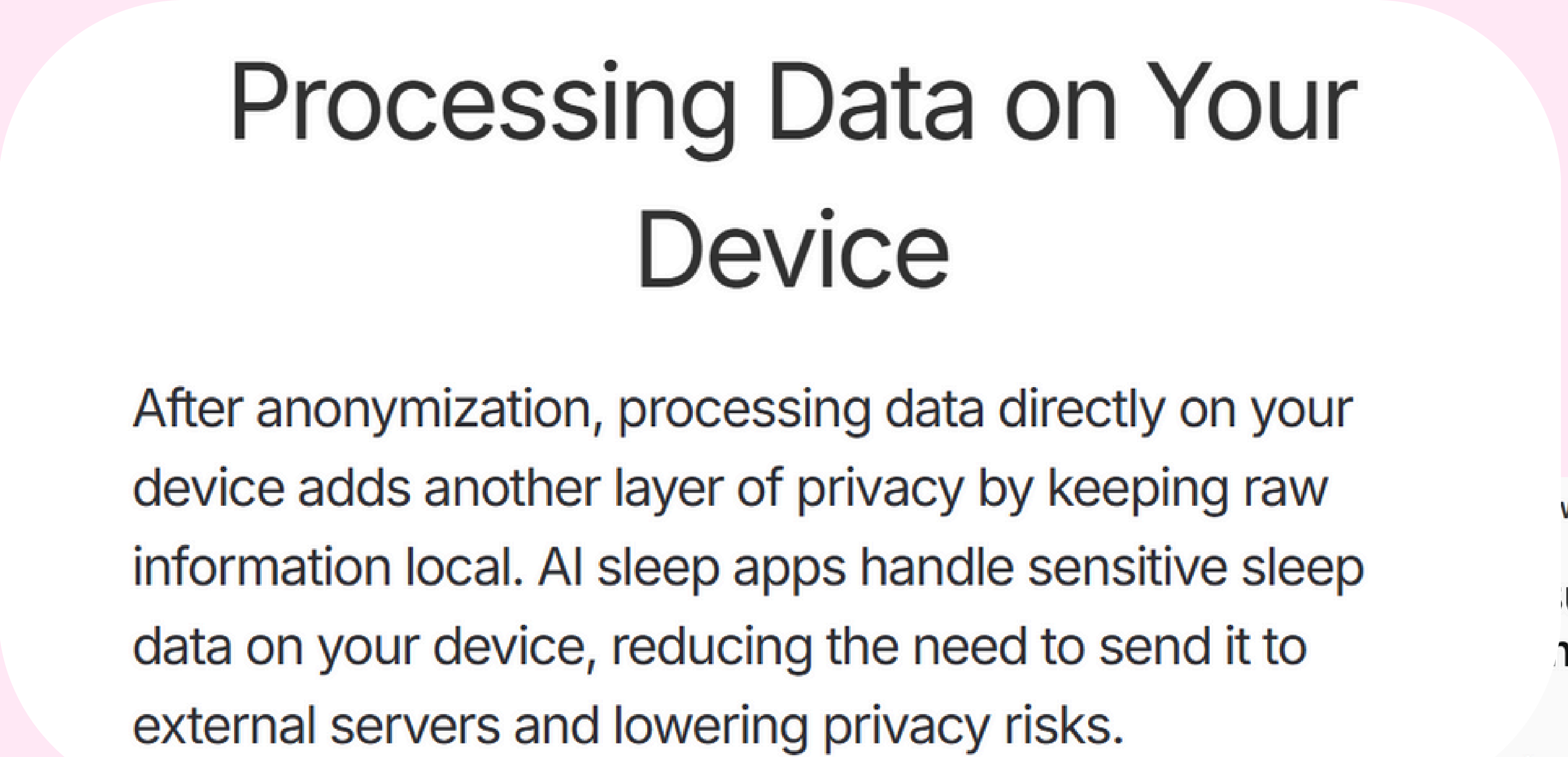
박준우

노기현

박정민

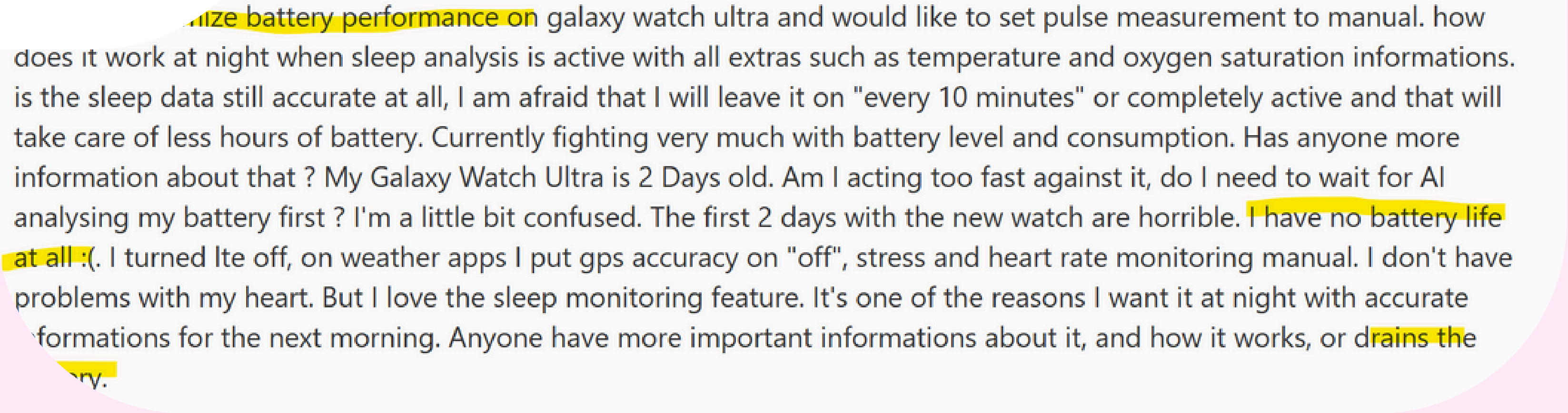
Why do we need to quantize / prune?

→ To achieve On-Device Inference



Processing Data on Your Device

After anonymization, processing data directly on your device adds another layer of privacy by keeping raw information local. AI sleep apps handle sensitive sleep data on your device, reducing the need to send it to external servers and lowering privacy risks.



Watch • 6 mo. ago

umption with pulse sensor + how it affects sleep
nt

nize battery performance on galaxy watch ultra and would like to set pulse measurement to manual. how does it work at night when sleep analysis is active with all extras such as temperature and oxygen saturation informations. is the sleep data still accurate at all, I am afraid that I will leave it on "every 10 minutes" or completely active and that will take care of less hours of battery. Currently fighting very much with battery level and consumption. Has anyone more information about that ? My Galaxy Watch Ultra is 2 Days old. Am I acting too fast against it, do I need to wait for AI analysing my battery first ? I'm a little bit confused. The first 2 days with the new watch are horrible. I have no battery life at all :(. I turned lte off, on weather apps I put gps accuracy on "off", stress and heart rate monitoring manual. I don't have problems with my heart. But I love the sleep monitoring feature. It's one of the reasons I want it at night with accurate formations for the next morning. Anyone have more important informations about it, and how it works, or drains the ry.

Why do we need to quantize / prune?

→ To achieve On-Device Inference

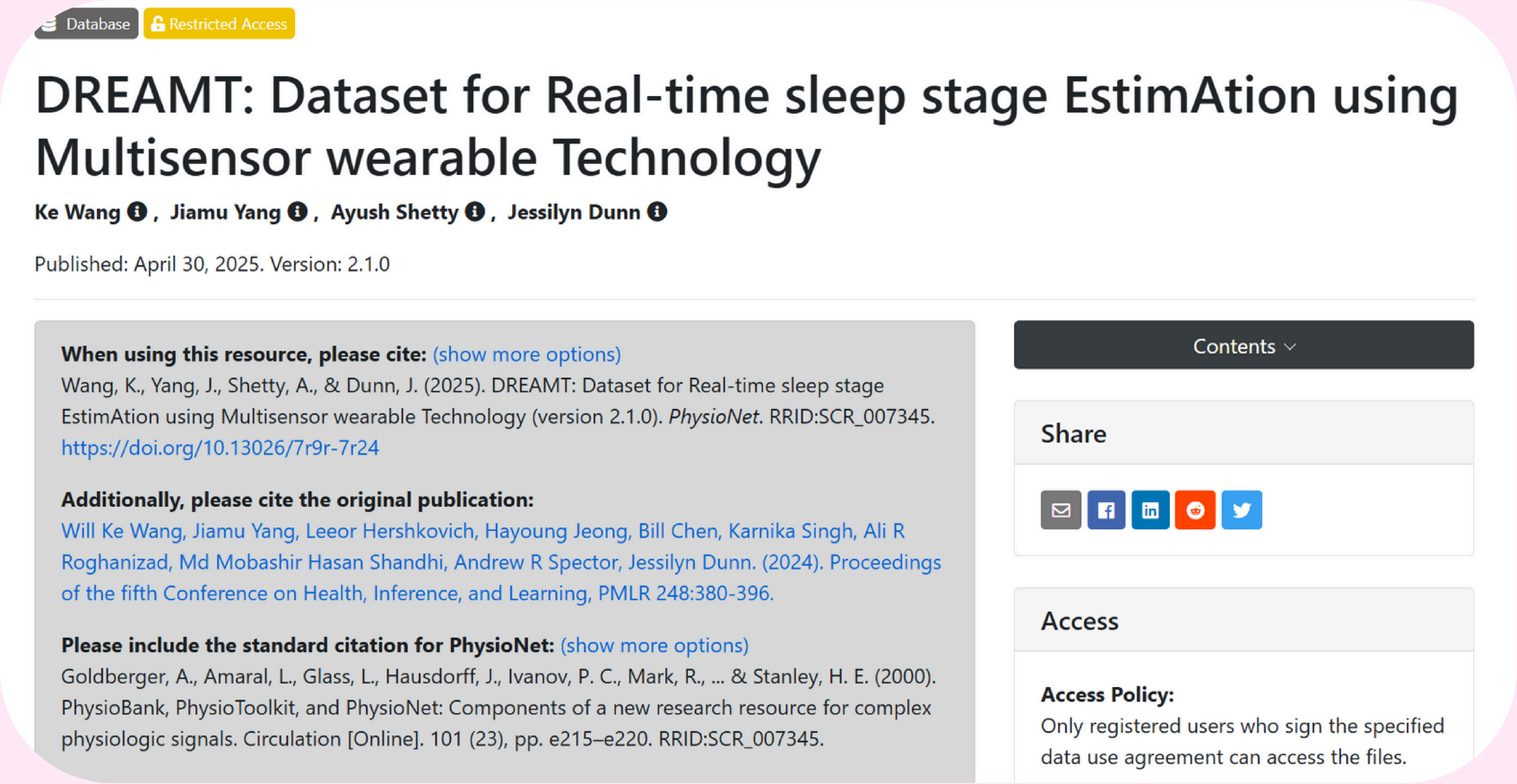
→Model that can actually run in wearable devices “locally” (e.g. Smart Watch)

Goal

- Minimal Accuracy Decline
- Reduced Model Size & Parameters
- Improved Memory & Latency

About Dataset

Dataset for Real-time sleep stage Estimation using Multisensor wearable Technology



DREAMT: Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology

Ke Wang ℹ, **Jiamu Yang** ℹ, **Ayush Shetty** ℹ, **Jessilyn Dunn** ℹ

Published: April 30, 2025. Version: 2.1.0

When using this resource, please cite: (show more options)
Wang, K., Yang, J., Shetty, A., & Dunn, J. (2025). DREAMT: Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology (version 2.1.0). *PhysioNet*. RRID:SCR_007345. https://doi.org/10.13026/7r9r-7r24

Additionally, please cite the original publication:
Will Ke Wang, Jiamu Yang, Leeor Hershkovich, Hayoung Jeong, Bill Chen, Karnika Singh, Ali R Roghanizad, Md Mobashir Hasan Shandhi, Andrew R Spector, Jessilyn Dunn. (2024). Proceedings of the fifth Conference on Health, Inference, and Learning, PMLR 248:380-396.

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.

Share

Access

Access Policy:
Only registered users who sign the specified data use agreement can access the files.

Baseline Model

MobileNet v2 adapted for 1D Time-Series

Calculate Acc / F1-score / Confusion Matrix

Pruning

torch.nn.utils.prune() 이용

→ weight sparsity (30%, 50%, 70%, ...)
→ Check the difference in Accuracy

Post-training dynamic quantization

torch.quantization.quantize_dynamic 이용

→ convert to int8

→ check the model size/Accuracy

Comparing Acc, latency, and size

- Baseline model
- Pruned
- Quantized
- Pruned + Quantized

감사합니다