일반논문 (Regular Paper)
방송공학회논문지 제31권 제1호, 2026년 1월 (JBE Vol.31, No.1, January 2026)
https://doi.org/10.5909/JBE.2026.31.1.69
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

MaskSmooth: 대규모 언어 모델 사후 양자화를 위한 채널 선택적 스무딩 기법

Tahir Khalil[a)], 배 성 호[a)‡]

MaskSmooth: Channel Selective Smoothing for Robust PTQ of LLMs

Tahir Khalil[a)] and Sung-Ho Bae[a)‡]

요 약

대규모 언어 모델(LLM)은 인공지능 분야의 여러 영역을 획기적으로 발전시킨 혁신적인 기술로 부상하였다. 그러나 LLM은 막대한 연산량과 메모리 요구량으로 인해 자원이 제한된 환경에서 배포하는 데 큰 어려움을 초래한다. 최근 학습 후 양자화(PTQ)는 성능 유지를 통해 LLM을 압축하는 효과적인 접근 방식으로 주목받고 있다. 본 연구에서는 선택적 채널 평활화를 통해 양자화된 모델의 견고성을 향상시키도록 설계된 프레임워크인 MaskSmooth를 소개한다. 모든 채널을 필요 여부와 관계없이 균일하게 평활화하는 기존 방식과 달리, MaskSmooth는 유의미한 이상치 활성화를 보이는 채널만을 식별하고 집중적으로 처리한다. 마스킹 메커니즘을 적용하여 관련 채널을 분리함으로써, MaskSmooth는 정상적인 표현을 불필요하게 변경하지 않고도 더욱 효율적인 평활화를 달성한다. 다양한 LLM 아키텍처에 대한 실험 결과는 MaskSmooth가 양자화 견고성을 일관되게 향상시키고 주요 벤치마크에서 모델 정확도를 유지함을 보여준다.

Abstract

Large Language Models (LLMs) have emerged as a revolutionary technology, significantly advancing numerous domains across artificial intelligence. However, their extensive computational and memory requirements present major challenges for deployment in resource-constrained environments. Post-Training Quantization (PTQ) has recently gained substantial attention as an effective approach for compressing LLMs while maintaining performance. In this work, we introduce MaskSmooth, a framework designed to enhance the robustness of quantized models through selective channel smoothing. Unlike existing methods that uniformly smooth all channels regardless of their need, MaskSmooth identifies and targets only those channels exhibiting significant outlier activations. By applying a masking mechanism to isolate relevant channels, our approach achieves more efficient smoothing without unnecessary alterations to well-behaved representations. Experimental results across a variety of LLM architectures demonstrate that MaskSmooth consistently improves quantization robustness and preserves model accuracy on mainstream benchmarks.

Keyword : Large Language Models, Post Training Quantization, Model Compression

Ⅰ. Introduction

Large Language Models (LLMs) have demonstrated impressive performance across a multitude of natural language processing tasks, ranging from text generation[1] to reasoning and understanding[2]. Much of this capability arises from the large-scale training of these models, where both the amount of data and the available computational resources have grown significantly. Despite their remarkable performance, running and serving LLMs remains computationally intensive, often requiring substantial infrastructure. For reasons of privacy, latency, and accessibility, enabling LLMs to operate in resource-constrained environments has become a critical research challenge[3].

This challenge is compounded by the sheer size of modern LLMs, which contain billions of parameters and rely on extensive matrix operations for inference. To address these limitations, several model compression techniques have been developed, among which quantization[4] has emerged as the most prominent. Quantization reduces the memory footprint of model parameters by representing them in low-precision formats (e.g., INT8) instead of full precision (FP16 or FP32), thereby accelerating computations and reducing storage requirements. Among quantization techniques, Post-Training Quantization (PTQ)[5] has started showing promise within LLMs. PTQ requires only a full-precision pretrained model and a small calibration dataset, making it feasible to quantize large models on limited compute resources. In contrast, Quantization-Aware Training (QAT)[4] demands access to the full dataset and retraining of the quantized model, which can be impractical for large LLMs due to resource and data constraints.

In this work, we introduce MaskSmooth, an efficient PTQ framework for LLMs. MaskSmooth is motivated by the observation that certain activation channels consistently exhibit high values across tokens, making them harder to quantize compared to weights, which typically do not display such outliers. Prior work[6] has focused on mitigating this difficulty by scaling all channels, effectively shifting quantization challenges from activations to weights. However, we observe that uniformly scaling all channels is often unnecessary. Instead, MaskSmooth selectively identifies channels containing significant outlier activations and applies targeted smoothing, while masking the remaining channels. This selective approach improves quantization robustness while reducing unnecessary transformations, resulting in more efficient and accurate low-precision LLMs.

Ⅱ. Background

Quantization is a widely used model compression technique[4] that reduces the precision of neural network parameters and activations, enabling efficient storage and faster computation. In a typical setting, full-precision parameters (FP32 or FP16) are mapped to low-precision integer formats, such as INT8 or INT4. This transformation allows both memory savings and hardware acceleration for integer arithmetic, which is often more efficient than floating-point computation on modern accelerators.

Mathematically, a uniform linear quantization maps a real-valued tensor $x \in R$ to a quantized tensor $q \in Z$ using a scale factor $s > 0$ and an optional zero-point z:

$$q = round\left(\frac{x}{s}\right) \quad (1)$$

$$q = round\left(\frac{x}{s} + z\right) \quad (2)$$

a) 경희대학교 컴퓨터공학과(Department of Computer Science and Engineering, Kyung Hee University)
‡ Corresponding Author : 배성호(Sung-Ho Bae)
E-mail: shbae@khu.ac.kr
Tel: +82-31-201-2593
ORCID: https://orcid.org/0000-0003-2677-3186

※ This research was supported by the three MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support programs (IITP-2023-RS-2023-00258649), (IITP-2024-RS-2023-00259004) and (IITP-2024-RS-2024-00438239), supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)
・ Manuscript October 10, 2025; Revised December 12, 2025; Accepted December 15, 2025.

Eq (1) and Eq (2) refer to symmetric and asymmetric quantization respectively. In symmetric quantization, the zero-point is fixed at zero, and the range of quantized values is symmetric around zero. The scale is computed based on the maximum absolute value of the tensor:

$$s = \frac{max(|x|)}{2^{b-1} - 1} \tag{3}$$

where b is the number of bits used for quantization. Each quantized integer can be mapped back to a floating-point approximation of the original tensor via:

$$x_{quantized} \approx s \cdot q \tag{4}$$

Symmetric quantization is particularly appealing for LLMs due to its simplicity and efficiency. Since the zero-point is zero, matrix multiplications can be performed directly in integer arithmetic without additional offset adjustments. Moreover, it preserves the relative scale of positive and negative values, which is often critical for maintaining model accuracy in attention mechanisms and feed-forward networks.

Ⅲ. Related Works

Model compression has become a critical area of research to enable deployment of deep neural networks on resource-constrained devices. One of the earliest comprehensive approaches[7] combined pruning, trained quantization, and Huffman coding to significantly reduce model size while maintaining accuracy. Subsequent research focused specifically on quantization as a means to reduce both memory footprint and computational cost. Some techniques[4] utilized arithmetic operations which allow neural networks to operate with low-precision parameters, achieving substantial speedups on hardware accelerators optimized for integer computations. More recent methods[8] have explored adaptive rounding strategies which improve the fidelity of quantization by minimizing rounding errors and preserving model accuracy without retraining.

Post-Training Quantization (PTQ)[5] has recently gained prominence as a practical approach for compressing Large Language Models (LLMs) without requiring full retraining. Early works, such as GPTQ[9] and ZeroQuant[10], focused on efficiently quantizing weights to low-bit precision while maintaining accuracy on generative and instruction-following tasks. These methods demonstrated that careful weight quantization, sometimes combined with approximate rounding or error compensation, could yield substantial memory and compute savings even for billion-parameter models. Building on this, SmoothQuant[6] introduced a strategy to mitigate quantization errors in activations by shifting part of the scaling from activations to weights. This approach reduces the impact of outlier activations, which are particularly problematic in transformer-based LLMs. Similarly, AWQ[11] and OWQ[12] identify channels or values that deviate significantly from the majority distribution and apply targeted quantization techniques to improve robustness. Other notable PTQ frameworks include OmniQuant[13], Quarot[14], and AffineQuant[15], which explore mixed-precision quantization, adaptive scaling, and fine-grained channel-level adjustments to preserve model accuracy under aggressive compression. These methods highlight the growing trend of selectively addressing quantization challenges in LLMs, rather than uniformly applying transformations across all weights or activations.

Ⅳ. Methodology

In this section, we introduce MaskSmooth, a selective channel-smoothing framework designed to enhance the robustness of post-training quantized Large Language Models (LLMs). Unlike prior works that uniformly smooth all channels, MaskSmooth identifies and selectively smooths only those channels exhibiting significant outlier behavior. This approach effectively reduces unnecessary computation and preserves model integrity while maintaining low quantization error.

1. Motivation

Empirical analysis of LLM activations reveals that certain channels consistently exhibit high-magnitude spikes, deviating significantly from the average channel distribution. These spikes, or outliers, introduce substantial variance during quantization, often leading to degraded model accuracy. Interestingly, our investigation shows that such outlier behavior is persistent across tokens and samples, suggesting that specific channels inherently tend to amplify activation magnitudes. Previous methods, such as SmoothQuant[6], addressed this problem by applying a scaling transformation that transfers part of the activation variance to the corresponding weight channels. This operation smooths activations globally, enabling reduced quantization error. However, these techniques typically perform uniform smoothing across all channels, even for those that do not exhibit any outlier behavior, introducing unnecessary computational overhead and potential information distortion.

2. Mask Generation

To effectively distinguish between channels that contain outliers and those that do not, MaskSmooth begins by running the model on a small calibration dataset. During this phase, we record the maximum activation value for each channel across all layers and tokens. This collection of statistics provides a comprehensive view of the activation distribution and highlights channels exhibiting extreme magnitude behavior. Following this data collection, we perform a statistical analysis to determine which channels can be categorized as outlier channels. Specifically, we employ classical statistical tools such as the interquartile range (IQR) and its variants to establish the outlier limits for each layer. Channels whose activation maxima exceed these limits are flagged as outlier channels and are thus included in the smoothing mask.

After determining the per-channel thresholds that distinguish inliers from outliers, we construct a binary mask for each activation channel. Activation values falling outside the threshold range (outliers) are preserved, while values within the range (inliers) are replaced with the corresponding upper fence value. An analogous procedure is applied to the weights: channel-wise bounds are computed, and the channels aligned with the masked activation chan-

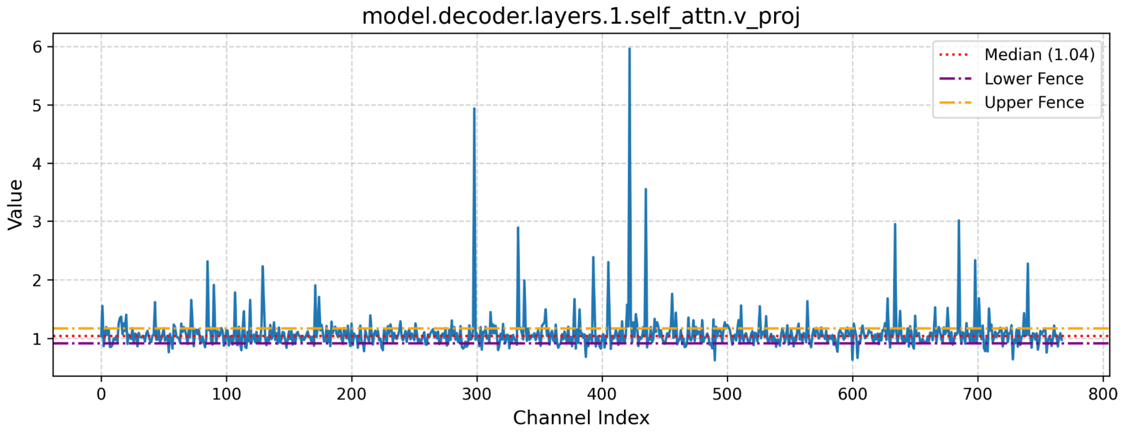

그림 1. OPT-2.7B 모델에서 특정 레이어의 채널별 활성화 값의 최대값을 시각화한 그림
Fig. 1. Visualization of maximum values of a certain layer's activation values per channel for OPT-2.7B model

nels are replaced using the lower fence value of the weights. This selective replacement introduces an additional smoothing effect, stabilizing activation‑weight interactions during quantization. Replacing all inlier activations with their channel-specific upper threshold addresses an inherent weakness of calibration data: many channels do not exhibit true outliers during calibration, leading to insufficient correction in those channels. By enforcing threshold-based replacement even in the absence of observed outliers, our method compensates for under-represented outlier behavior, strengthens the smoothing effect across channels, and ultimately yields a consistent and measurable improvement in perplexity.

Figure 1 shows the per channel maximum activation values for the first decoder layer of OPT-2.7B[16] model which clearly shows that we can isolate outliers and scale only them.

However, we observe that activation fluctuations between channels can make the masking scheme highly sensitive to the selected outlier threshold. To mitigate this, we empirically determine that maintaining the outlier window around the median value yields more stable and effective results. For our work, a threshold of ~2% around the median was used, with channels within this range considered normal and those outside it identified as outliers. Through this process, approximately 10% of the channels that fall within the defined threshold range are identified as non-outliers and subsequently masked or excluded from smoothing. The remaining ~90% of channels, whose values lie outside the threshold range, are then selected for the smoothing operation. The detailed values for different models can be seen in Table 1. This configuration ensures that only a limited subset of channels, those with pro-nounced deviations, are masked for smoothing while the majority of channels are preserved for direct scaling.

표 1. 이상치에 대한 제안된 마진을 사용하여 채널 마스킹을 구현
Table 1. Channel masking achieved with proposed margin for outliers

Model	Channel Mask (%)
OPT-125M	9.47%
OPT-1.3B	9.81%
OPT-2.7B	9.50%
OPT-6.7B	9.95%

3. Masked Channel Smoothing

Once the mask is generated, we modify the weight and activation scales accordingly to calculate the scaling factor using the offline migration introduced by smoothquant.

$$W' = max(W) \circ M \tag{5}$$

$$X' = max(X) \circ M \tag{6}$$

$$s = sqrt((X') / (W')) \tag{7}$$

Here, W and X denote the weights and activations, respectively. The function max () computes the per-channel maximum weight and activation value, upon which the mask M is applied, here denoted as $\circ$, to prevent non-outlier channels from undergoing the smoothing operation. Since the weights are already available in memory and the activation maxima are obtained from the calibration set, the per-channel scale values can be precomputed efficiently as s.

Ⅴ. Experiments and Results

We conduct our study using Meta’s OPT[16] model family, a series of open-source autoregressive language models trained on large-scale text corpora. OPT serves as a suitable testbed for our experiments due to its architectural similarity to GPT-style transformer models and its widespread use in quantization research. All experiments are performed on the OPT model variants to ensure comparability with prior PTQ methods such as SmoothQuant[6]. To extract activation statistics and generate channel masks, we utilize the validation split of the WikiText-2 dataset[17]. This dataset provides a diverse yet compact corpus suitable for com

표 2. 제안된 양자화 방식의 성능과 기준 방식의 성능 비교
Table 2. Performance of proposed quantization scheme with baseline

	Accuracy (↑)				Perplexity (↓)			
	FP16	MinMax	Smoothquant	MaskSmooth	FP16	MinMax	Smoothquant	MaskSmooth
OPT-125M	63.07%	61.19%	62.60%	**62.62%**	38.12	42.81	41.14	**41.09**
OPT-1.3B	75.49%	73.08%	74.31%	**74.38%**	22.53	25.09	23.05	**23.01**
OPT-2.7B	77.84%	77.12%	**78.18%**	77.62%	19.54	21.92	19.69	**19.66**
OPT-6.7B	81.25%	45.18%	**81.43%**	81.39%	17.35	56.8	**17.49**	**17.49**

puting representative activation ranges and outlier characteristics without incurring excessive computational cost.

For evaluation, we employ Last-Word Prediction Accuracy (LWPA) and Perplexity (PPL), two widely adopted metrics that reflect both generative quality and prediction accuracy. We compute these metrics for the LAMBADA[18] test set.

Detailed results of the experiments performed can be seen in Table 2. Here 4 different OPT models are evaluated under 4 different scenarios. Firstly models are evaluated in full precision (FP16) and naive min-max quantization to establish the upper and lower bounds respectively. We then compare our proposed methodology with smoothquant to show that we achieve better performance in terms of perplexity whereas accuracy does go down a little for certain models.

Ⅵ. Performance Analysis

In this section, we investigate the performance of the proposed approach under different thresholding criteria.

To illustrate the effect, we use the OPT-1.3B model and conduct experiments across different threshold values. This

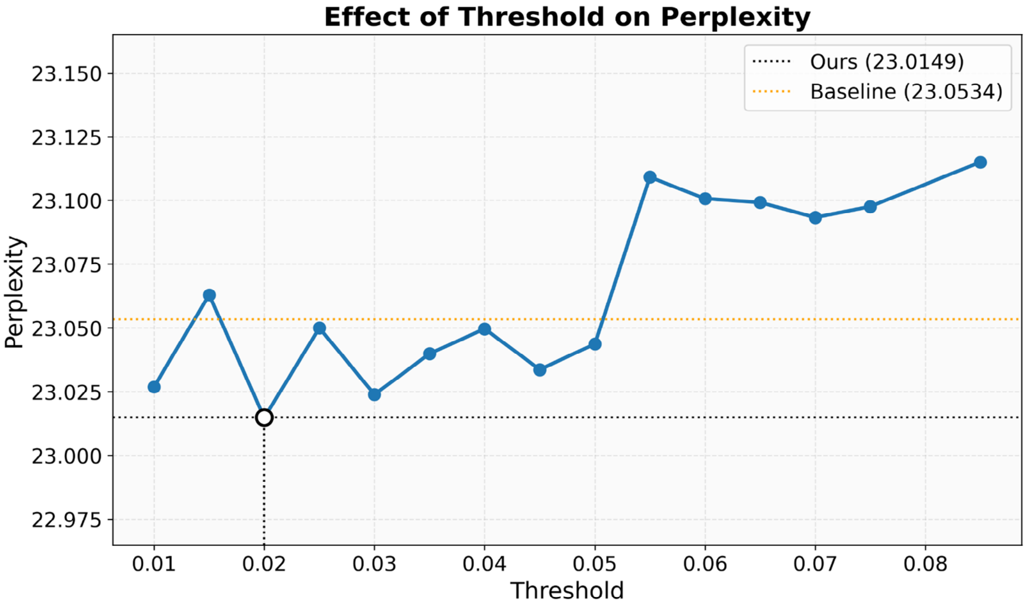

그림 2. OPT-1.3B 모델의 임계값 변화에 따른 혼란도 변화
Fig. 2. Variation in perplexity with change in threshold value for OPT-1.3B model

allows us to observe a clear trend: lower thresholds help balance the treatment of outliers and non-outliers, whereas higher thresholds lead to a degradation in model perplexity. Figure 2 demonstrates this behavior, showing how increasing the threshold negatively impacts performance.

Ⅶ. Conclusion

In this work, we presented MaskSmooth, an efficient and targeted framework for post-training quantization (PTQ) of large language models. Unlike prior approaches that uniformly apply smoothing to all activation channels, Mask Smooth introduces a mask-based selective smoothing mechanism that focuses only on channels exhibiting outlier behavior. This selective approach reduces unnecessary computation while retaining or even improving quantization robustness. Through extensive experimentation on Meta's OPT models, using WikiText-2 for calibration and evaluating on the LAMBADA benchmark, we demonstrated that MaskSmooth achieves competitive or superior performance across key metrics. These findings indicate that channel-aware selective smoothing provides an effective balance between efficiency and precision in PTQ settings.

Future work will extend this framework to explore dynamic masking strategies, ultra low bit settings, adaptive thresholds based on task-specific statistics, and integration with mixed-precision quantization. We believe Mask Smooth opens a new direction for fine-grained quantization strategies that make large-scale language models more practical for deployment in resource-constrained environments.

참 고 문 헌 (References)

[1] Brown, Tom, et al. “Language models are few-shot learners.” Advances in neural information processing systems 33 (2020): 1877-1901.
doi: https://doi.org/10.48550/arXiv.2005.14165

[2] Wei, Jason, et al. “Emergent abilities of large language models.” arXiv preprint arXiv:2206.07682 (2022).
doi: https://doi.org/10.48550/arXiv.2206.07682

[3] Girija, Sanjay Surendranath, et al. “Optimizing llms for resource-constrained environments: A survey of model compression techniques.” 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2025.
doi: https://doi.org/10.48550/arXiv.2505.02309

[4] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2704–2713, 2018.
doi: https://doi.org/10.48550/arXiv.1712.05877

[5] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv: 1806.08342, 2018.
doi: https://doi.org/10.48550/arXiv.1806.08342

[6] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In International Conference on Machine Learning, pp. 38087–38099. PMLR, 22023.
doi: https://doi.org/10.48550/arXiv.2211.10438

[7] Han, S., Mao, H., and Dally, W. J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. In ICLR, 2016.
doi: https://doi.org/10.48550/arXiv.1510.00149

[8] Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C., and Blankevoort, T. Up or down? Adaptive rounding for post-training quantization. In International Conference on Machine Learning, pp. 7197–7206. PMLR, 2020.
doi: https://doi.org/10.48550/arXiv.2004.10568

[9] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323, 2022.
doi: https://doi.org/10.48550/arXiv.2210.17323

[10] Yao, Zhewei, et al. “Zeroquant: Efficient and affordable post-training quantization for large-scale transformers.” Advances in neural information processing systems 35 (2022): 27168-27183.
doi: https://doi.org/10.48550/arXiv.2206.01861

[11] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. Proceedings of Machine Learning and Systems, 6: 87–100, 2024.
doi: https://doi.org/10.48550/arXiv.2306.00978

[12] Lee, Changhun, et al. “Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol.

38. No. 12. 2024.
doi: https://doi.org/10.48550/arXiv.2306.02272

[13] Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In The Twelfth International Conference on Learning Representations, 2024.
doi: https://doi.org/10.48550/arXiv.2308.13137

[14] Ashkboos, Saleh, et al. "Quarot: Outlier-free 4-bit inference in rotated llms." Advances in Neural Information Processing Systems 37 (2024): 100213-100240.
doi: https://doi.org/10.48550/arXiv.2404.00456

[15] Ma, Yuexiao, et al. "Affinequant: Affine transformation quantization for large language models." arXiv preprint arXiv:2403.12544 (2024).
doi: https://doi.org/10.48550/arXiv.2403.12544

[16] Zhang, Susan, et al. "Opt: Open pre-trained transformer language models." arXiv preprint arXiv:2205.01068 (2022).
doi: https://doi.org/10.48550/arXiv.2205.01068

[17] Merity, Stephen, et al. "Pointer sentinel mixture models." arXiv preprint arXiv:1609.07843 (2016).
doi: https://doi.org/10.48550/arXiv.1609.07843

[18] Denis Paperno, Germ´ an Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fern´ andez. The lambada dataset: Word prediction requiring a broad discourse context. arXiv preprint arXiv:1606.06031, 2016.
doi: https://doi.org/10.18653/v1/P16-1144

저자소개

Tahir Khalil

- Sept. 2014 ~ Aug. 2018 : Bachelor's in Department of Mechatronics and Control Engineering at University of Engineering and Technology, Lahore, Pakistan
- Sept. 2019 ~ Aug. 2022 : Master's in Department of Computer Science at Information Technology University, Lahore, Pakistan
- Sept. 2023 ~ Present : Ph.D. in Department of Computer Science at Kyung Hee University, South Korea
- ORCID : https://orcid.org/0009-0006-3366-1712
- Research Interests : deep learning, image processing, generative models, and model compressionrative models, and model compression

배성호

- Mar. 2004 ~ Feb. 2011 : Bachelor's degree in Department of Computer Engineering and Electronic Engineering (dual majors) at Kyung Hee University, South Korea
- Feb. 2011 ~ Aug. 2016 : Ph.D. in Department of Electrical Engineering at KAIST, South Korea
- Jul. 2016 ~ Aug. 2017 : Postdoc. Associate in MIT Computer Science and Artificial Intelligence Laboratory (CSAIL)
- Sept. 2017 ~ Present : Associate Professor at Kyung Hee University in school of Computing, South Korea
- ORCID : https://orcid.org/0000-0003-2677-3186
- Research Interests : Inverse problems in image processing, video compression, model compression, generative AI