IMPACT OF NORMALIZATION IN MULTIMODAL AI

Neeraj Kumar

BHARTI SCHOOL OF TELECOMMUNICATION TECHNOLOGY & MANAGEMENT

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2024

IMPACT OF NORMALIZATION IN MULTIMODAL AI

by

Neeraj Kumar

Bharti School of Telecommunication Technology & Management

Submitted

in fulfillment of the requirements of the degree of **Doctor of Philosophy**

to the

Indian Institute of Technology Delhi

JULY 2024

Certificate

This is to certify that the thesis titled **"IMPACT OF NORMALIZATION IN MULTIMODAL AI"** being submitted by **NEERAJ KUMAR** for the award of **Doctor of Philosophy** in **Bharti School of Telecommunication Technology & Management,** is a record of bona fide work carried out by him under my guidance and supervision at the **Bharti School of Telecommunication Technology & Management, Indian Institute of Technology Delhi**. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Ankur Narang

Ankur Narang
VP of AI
Hike Private Limited
New Delhi- 110016

Brejesh Lall
Professor
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi- 110016

Acknowledgements

I owe my ability to pursue my PhD to the unwavering support of my family, especially my parents, who have shown incredible understanding and made significant sacrifices throughout their lives. I am also immensely grateful to my wife, Shinjini, for her constant presence and support. It was the nurturing and supportive environment at home that allowed me to dedicate the majority of my waking hours to my academic pursuits.

I extend my deepest appreciation to my advisors, Prof. Brejesh Lall and Dr. Ankur Narang. Your faith in me and the freedom you granted me to explore topics of personal interest have been invaluable. I am thankful for the continuous encouragement, feedback, guidance, and support you provided, serving as a constant source of inspiration.

Throughout my doctoral journey, I held a full-time position at Hike Private Limited, a workplace I consider exceptional. I have gained a wealth of knowledge from my colleagues, and my PhD would not have been achievable without the unwavering support of my friends and coworkers at Hike. I express my gratitude to my colleagues and teammates, including Srishti, Pranshu, Dipankar, and Kush, for their invaluable support and technical insights.

I would like to acknowledge all my friends at IITD for their technical discussions, which greatly contributed to my research and enhanced my comprehension. Additionally, they made life enjoyable and thought-provoking with their shared experiences.

A special mention goes to my remarkable circle of friends, including Sohil, Pallav, Mohit, Ayan, Kaushal, Jayant and Vikas for the joy, humor, and laughter they brought into my life.

Lastly, I am thankful to God for everything and all that I have achieved.

Neeraj Kumar

Abstract

Multimodal deep learning systems, which leverage multiple modalities such as text, image, audio, and video, have demonstrated superior performance compared to single-modality systems. The process of multimodal machine learning encompasses several stages, including representation, translation, alignment, fusion, and co-learning.

Three distinct fusion techniques are examined: early fusion, intermediate fusion, and late fusion. Early fusion involves amalgamating raw data from diverse modalities into a unified representation before the learning phase. Intermediate fusion entails transforming raw inputs into a higher-level representation through a series of layers. Late fusion, on the other hand, entails merging predictions from each modality to arrive at a final decision. Various fusion methods, including element-wise summation, weighted averaging, bilinear product, rank minimization, and attention mechanisms, have been proposed previously.

However, existing methods face several challenges. These include discrepancies or misaligned spatial dimensions of intermediate features from different modalities, overfitting issues in late fusion networks due to the increased parameter count, and complications in utilizing pre-trained weights from uni-modal networks when incorporating intermediate fusion. There is also the potential for performance degradation in multimodal models compared to uni-modal ones when employing a joint training strategy, which contradicts the goal of enhancing performance through multimodal integration. Different modalities may converge at varying rates, resulting in uncoordinated convergence issues.

Our research has delved into the role of normalization in various multimodal applications, such as audio-to-video synthesis, text-to-speech synthesis, and uni-modal feature extraction. We have introduced innovative normalization techniques that have proven effective across different multimodal applications. Our approach, known as multimodal normalization, is designed to capture the interdependence between different domains by utilizing a mixture of multivariate Gaussian distributions. This method helps prevent the model from experiencing mode collapse, a common issue when trying to model data distribution solely with multivariate Gaussian

distributions. The integration of multimodal normalization has yielded outstanding results in our proposed video synthesis, involving the transformation of audio and a single image, outperforming previous methods on various metrics such as SSIM (structural similarity index), PSNR (peak signal-to-noise ratio), CPBD (image sharpness), WER (word error rate), blinks per second, and LMD (landmark distance). In the context of video emotion detection, we harnessed the power of the mel-spectrogram and optical flow within the multimodal normalization approach to learn affine parameters, thereby contributing to enhanced accuracy.

We have introduced a novel normalization framework for text-to-speech applications, which is designed to capture stylistic features at both the speaker and frame levels, in addition to accommodating style-agnostic features. This is achieved through the utilization of learnable parameters denoted as γ (scale) and β (bias) within the normalization framework. To compute the various learnable parameters associated with speaker embeddings from the speaker encoder, pitch, and energy, we have proposed two distinct approaches: one based on convolutional networks and the other employing a multi-head attention network. We establish the effectiveness of our proposed architecture through comprehensive evaluations on the VCTK and LibriTTS datasets. Our assessments include visualizing the Hessian matrix of the proposed model, employing various quantitative metrics to measure speech distortion, evaluating Mean Opinion Scores (MOS), and conducting an in-depth analysis of speaker embeddings generated by our novel speaker encoder model.

We introduce a novel approach known as Kullback-Leibler (KL) Regularized Normalization (KL-Norm). This technique enhances the stability of normalized data, thereby contributing to improved generalization by reducing overfitting. It generalises well in adapting to out-of-domain distributions while eliminating irrelevant biases and features, all accomplished with minimal impact on model parameters and memory resources. Our extensive empirical assessments across various low-resource natural language processing (NLP) and speech-related tasks substantiate the superior performance of KL-Norm when compared to well-known normalization and regularization methods.

Lastly, Our contributions encompass the introduction of innovative multimodal normalization techniques, which have proven advantageous in diverse applications such as speech-driven video generation and video emotion detection, the development of a normalization framework within text-to-speech systems to capture both stylistic and style-agnostic features, and the introduction of KL Regularized normalization for addressing low-resource natural language processing challenges. This research topic has resulted in the creation of novel normalization architectures and an extensive exploration of the impact of normalization techniques across a wide range of multimodal applications.

सार

मल्टीमॉडल डीप लर्निंग सिस्टम, जो पाठ, छवि, ध्वनि और वीडियो का उपयोग करते हैं, एकल-मॉडलिटी सिस्टम की तुलना में बेहतर प्रदर्शन करते हैं। मल्टीमॉडल मशीन लर्निंग में प्रतिनिधित्व, अनुवाद, संरेखण, संयोजन और सह-अधिगम जैसे चरण शामिल हैं। संयोजन तकनीकों में प्रारंभिक, मध्यवर्ती और अंतिम संयोजन शामिल हैं। प्रारंभिक संयोजन में विभिन्न माध्यमों के कच्चे डेटा को एकीकृत प्रतिनिधित्व में मिलाया जाता है। मध्यवर्ती संयोजन में कच्चे इनपुट को परतों के माध्यम से उच्च-स्तरीय प्रतिनिधित्व में बदल दिया जाता है। अंतिम संयोजन में प्रत्येक माध्यम से प्राप्त पूर्वानुमानों को अंतिम निर्णय में मिलाया जाता है। संयोजन विधियों में तत्व-वार जोड़, भारित औसत, द्विखंडित उत्पाद, रैंक न्यूनतमकरण और ध्यान तंत्र शामिल हैं। चुनौतियों में विभिन्न माध्यमों के मध्यवर्ती विशेषताओं के विसंगत या असंगत स्थानिक आयाम, अंतिम संयोजन नेटवर्क्स में अधिक अभ्यस्तता के मुद्दे और मध्यवर्ती संयोजन में पूर्व-प्रशिक्षित भार का उपयोग करने में कठिनाइयाँ शामिल हैं। संयुक्त प्रशिक्षण के दौरान विभिन्न माध्यमों के असंयोजित समापन दर के कारण मल्टीमॉडल मॉडल का प्रदर्शन एकल-मॉडल की तुलना में खराब हो सकता है।

हमारा शोध ध्वनि-से-वीडियो निर्माण, पाठ-से-भाषण निर्माण और एकल-माध्यम विशेषता निष्कर्षण जैसे मल्टीमॉडल अनुप्रयोगों में सामान्यीकरण का पता लगाता है। हम विभिन्न क्षेत्रों के बीच परस्पर निर्भरता को पकड़ने के लिए बहुसांख्यिक गॉसियन वितरणों के मिश्रण का उपयोग करके मल्टीमॉडल सामान्यीकरण प्रस्तुत करते हैं, जो मोड कोलैप्स को रोकता है। यह वीडियो निर्माण प्रदर्शन में सुधार करता है, जैसा कि संरचनात्मक समानता सूचकांक, पीक सिग्नल-टू-शोर अनुपात, छवि तीक्ष्णता, शब्द त्रुटि दर, प्रति सेकंड झपकी और लैंडमार्क दूरी जैसे मापदंडों द्वारा प्रदर्शित किया गया है। पाठ-से-भाषण के लिए, हम एक सामान्यीकरण ढांचा प्रस्तुत करते हैं जो वक्ता और फ्रेम-स्तरीय शैलीगत विशेषताओं को पकड़ता है, जो गामा और बीटा जैसे सीखने योग्य मापदंडों का उपयोग करता है। हम इन मापदंडों की गणना के लिए संपूर्ण नेटवर्क और बहु-प्रमुख ध्यान नेटवर्क का प्रस्ताव रखते हैं। वीसीटीके और लिब्रिट्टीएस डेटासेट पर हमारे आर्किटेक्चर की प्रभावशीलता हेसियन मैट्रिक्स, भाषण विकृति मापदंड, औसत राय स्कोर और वक्ता एम्बेडिंग द्वारा मापी गई है। हम कुलबैक-लेबलर (केएल) नियमितीकृत सामान्यीकरण भी प्रस्तुत करते हैं, जो डेटा स्थिरता को बढ़ाता है और अधिक अभ्यस्तता को कम करते हुए सामान्यीकरण में सुधार करता है। केएल सामान्यीकरण बाहरी-डोमेन वितरणों के अनुकूलन में अच्छा प्रदर्शन करता है और अप्रासंगिक पूर्वाग्रह और विशेषताओं को समाप्त करता है, सभी न्यूनतम मापदंड और स्मृति प्रभाव के साथ। विभिन्न निम्न-संसाधन प्राकृतिक भाषा संसाधन और भाषण कार्यों में केएल सामान्यीकरण का उत्कृष्ट प्रदर्शन हमारे व्यापक अनुभवजन्य मूल्यांकनों द्वारा सिद्ध है।

हमारे योगदान में भाषण-चालित वीडियो निर्माण और वीडियो भाव पहचान के लिए नवाचारी मल्टीमॉडल सामान्यीकरण तकनीकें, पाठ-से-भाषण के भीतर शैलीगत और शैली-अज्ञेय विशेषताओं को पकड़ने वाला सामान्यीकरण ढांचा, और निम्न-संसाधन प्राकृतिक भाषा संसाधन चुनौतियों के लिए केएल नियमितीकृत सामान्यीकरण शामिल हैं। इस शोध ने सामान्यीकरण आर्किटेक्चर में नवाचार किया है और विभिन्न मल्टीमॉडल अनुप्रयोगों में सामान्यीकरण तकनीकों के प्रभाव का व्यापक रूप से अन्वेषण किया है।

Contents

List of Figures

List of Tables