LEVERAGING METADATA AND MULTIMODAL INFORMATION IN DEEP EXTREME CLASSIFICATION

ANSHUL MITTAL

AMAR NATH AND SHASHI KHOSLA
SCHOOL OF INFORMATION TECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI

MAY 2024

LEVERAGING METADATA AND MULTIMODAL INFORMATION IN DEEP EXTREME CLASSIFICATION

by

ANSHUL MITTAL

AMAR NATH AND SHASHI KHOSLA
SCHOOL OF INFORMATION TECHNOLOGY

Submitted

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

to the

Indian Institute of Technology Delhi
May 2024

Certificate

This is to certify that the thesis titled **LEVERAGING METADATA AND MULTIMODAL INFORMATION IN DEEP EXTREME CLASSIFICATION** being submitted by **Mr. ANSHUL MITTAL** for the award of **Doctor of Philosophy** in **Extreme Classification** is a record of bonafide work carried out by him under my guidance and supervision at the Amar Nath and Shashi Khosla School of Information Technology, 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.

Sumeet Agarwal
Associate Professor
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi - 110016

Manik Varma
Distinguished Scientist
& Vice President
Microsoft Research India
Bengaluru - 560001

Acknowledgements

This doctoral journey has been an incredible voyage of intellectual exploration, personal growth, and immense challenges. The completion of this dissertation would not have been possible without the support and guidance of many wonderful individuals.

First and foremost, I owe a deep debt of gratitude to Dr. Manik Varma, Dr. Sumeet Agrwal, and Dr. Purushottam Kar for their unwavering support, insightful guidance, and trust in my abilities have been instrumental in shaping my research and development as a scholar. Their mentorship extends far beyond academia, as they have also offered invaluable personal support throughout my Ph.D. Their belief in me, even when I doubted myself, propelled me forward, especially when I ventured into seemingly risky projects.

My sincere thanks also go to my exceptional colleagues, Kunal Dhahiya and Deepak Saini. Countless late-night discussions, insightful lunch debates, and collaborative brainstorming sessions fueled my research and enriched my experience. I deeply appreciate their intellectual companionship and camaraderie.

I am incredibly grateful to my collaborators, Sheshansh Agrawal, Shreya Malani, Shikhar Mohan, and Naveen Sachdeva. Collaborating with these brilliant minds has been a rewarding experience, pushing my research boundaries and fostering meaningful intellectual exchange.

Words cannot express my heartfelt gratitude to Prateek Chhabra, Vyom Vashishta, Shruti Garg, and Ritika Singh. Their unwavering friendship, emotional support, and unwavering encouragement were like a beacon of hope during my darkest hours. They kept me grounded, motivated, and helped me navigate the emotional roller coaster of this journey.

Lastly but most importantly, my deepest love and gratitude go to my family: my sister, Dr. Megha, and my mother, Rajni, who have stood as pillars of strength throughout my life. I am forever indebted to the memory of my late father, Lect. Vinay Mittal, whose shield of protection, affection, and unwavering trust continue to inspire resilience in the face of adversity. His spirit motivates me to continue contributing to the world of knowledge.

To all of you, thank you from the bottom of my heart. Your presence has made this incredible journey possible.

Anshul Mittal

Abstract

Deep Extreme Classification (XC) refers to accurately tagging a data point with a relevant subset of labels from a large universe of labels. The challenge is to predict data scarce tail labels. Additionally, tagging every data point comprehensively becomes impractical due to the sheer number of labels, leading to potential omissions or "missing labels". Metadata about labels, including descriptions, image features, correlations, and relationships, offers valuable insights into document-label connections and holds promise for addressing these challenges. This thesis introduces four algorithms – DECAF, MUFIN, ECLARE, and RAMEN – that leverage the metadata to improve both accuracy and diverse recommendations.

Traditional XC methods like MACH and Parabel struggle with label semantics. However, DECAF tackles this challenge by understanding label meaning through the associated text. By capturing semantic relationships between labels, DECAF delivers accurate recommendations, outperforming baselines by 2-3% in accuracy.

Real-world product recommendations often involve both text and visual data. While vision-based methods like ADDE-O and CLIP excel at handling visuals, they face

difficulties in scaling. MUFIN bridges the gap between existing multi-modal approaches and XC through its scalable training and architecture, surpassing XC and vision-based methods by 4-10% in accuracy.

Leveraging multi-modality enhances accuracy for tail labels, but gains are limited due to significant data imbalance. Methods like X-Transformer and AttentionXML, despite utilizing label semantics, struggle with recommending tail labels. ECLARE harnesses textual descriptions and inferred label correlation graphs to "borrow strength" from more common head labels, allowing for accurate predictions of tail labels, outperforming baseline XC approaches by 4-5% in accuracy.

These label correlations are often incomplete due to missing labels. Graph metadata, such as connections between related products or webpages, can help identify the missing labels. Traditional methods like GraphFormers and PINA rely on computationally expensive graph convolution networks (GCNs) to utilize such data. RAMEN introduces a cost-effective alternative to GCNs, leveraging graph metadata to infer missing labels and subsequently enhancing performance by 3-5% in accuracy.

Collectively, these algorithms demonstrate the power of metadata. They pave the way for accurate and personalized recommendations, improved content tagging, and richer user experiences across various applications. Code for each method is available at the XML Repository [4].

सारांश

दीप एक्सट्रीम वर्गीकरण (एक्ससी) बड़ी संख्या में लेबलों में से प्रासंगिक उपसमूह के साथ डेटा बिंदु को सटीक रूप से टैग करने को संदर्भित करता है। चुनौती यह है कि डेटा की कमी वाले टेल लेबल की भविष्यवाणी की जाए। इसके अतिरिक्त, लेबलों की भारी संख्या के कारण प्रत्येक डेटा बिंदु को व्यापक रूप से टैग करना अव्यावहारिक हो जाता है, जिससे संभावित चूक या "लापता लेबल" हो जाते हैं। लेबल विवरण, छवि सुविधाओं, सहसंबंधों और संबंधों सहित लेबल के बारे में मेटाडेटा दस्तावेज़-लेबल कनेक्शन में मूल्यवान अंतर्दृष्टि प्रदान करता है और इन चुनौतियों का समाधान करने का वादा करता है। यह थीसिस चार एल्गोरिदम - DECAF, MUFIN, ECLARE और RAMEN - का परिचय देता है जो सटीकता और विविध अनुशंसाओं दोनों को बेहतर बनाने के लिए मेटाडेटा का लाभ उठाते हैं।

MACH और Parabel जैसी पारंपरिक XC विधियां लेबल अर्थशास्त्र के साथ संघर्ष करती हैं। हालांकि, DECAF संबंधित टेक्स्ट के माध्यम से लेबल अर्थ को समझकर इस चुनौती का सामना करता है। लेबल के बीच अर्थ संबंधों को कैप्चर करके, DECAF सटीक अनुशंसाएं प्रदान करता है, जो सटीकता में आधारभूत प्रदर्शन से 2-3% बेहतर प्रदर्शन करता है।

वास्तविक दुनिया में उत्पाद अनुशंसाओं में अक्सर पाठ और दृश्य दोनों डेटा शामिल होते हैं। जबकि दृष्टि-आधारित तरीके जैसे ADDE-O और CLIP दृश्यों को संभालने में उत्कृष्ट हैं, उन्हें स्केलिंग में कठिनाइयों का सामना करना पड़ता है। MUFIN अपने स्केलेबल प्रशिक्षण और आर्किटेक्चर के माध्यम से मौजूदा मल्टी-मोडल दृष्टिकोणों और XC के बीच के अंतर को पाटता है, जो सटीकता में XC और विज़न-आधारित विधियों को 4-10% से अधिक पार कर जाता है।

मल्टी-मोडैलिटी का लाभ उठाना टेल लेबल के लिए सटीकता को बढ़ाता है, लेकिन महत्वपूर्ण डेटा असंतुलन के कारण लाभ सीमित हैं। X- ट्रांसफॉर्मर और अटेंशनएक्सएमएल जैसी विधियां, लेबल अर्थशास्त्र का उपयोग करने के बावजूद, टेल लेबल की सिफारिश करने में संघर्ष करती हैं। ECLARE पाठ विवरणों और अनुमानित लेबल सहसंबंध ग्राफ़ का उपयोग करता है ताकि अधिक सामान्य हेड लेबल से "शक्ति उधार" ले सके, जिससे टेल लेबल की सटीक भविष्यवाणी की जा सके, आधारभूत XC दृष्टिकोणों को सटीकता में 4-5% से अधिक आउटपरफॉर्म किया जा सके।

ये लेबल सहसंबंध अक्सर लापता लेबलों के कारण अधूरे होते हैं। ग्राफ़ मेटाडेटा, जैसे संबंधित उत्पादों या वेबपृष्ठों के बीच संबंध, लापता लेबलों की पहचान करने में मदद कर सकता है। ग्राफफॉर्मर्स और पीआईएनए जैसी पारंपरिक विधियां ऐसे डेटा का

उपयोग करने के लिए कम्प्यूटेशनल रूप से महंगे ग्राफ कनवल्शन नेटवर्क (जीसीएन) पर भरोसा करती हैं। RAMEN GCN के लिए एक लागत- प्रभावी विकल्प प्रस्तुत करता है, जो लापता लेबल का अनुमान लगाने और बाद में सटीकता में 3-5% तक प्रदर्शन को बढ़ाने के लिए ग्राफ मेटाडेटा का लाभ उठाता है।

सामूहिक रूप से, ये एल्गोरिदम मेटाडेटा की शक्ति का प्रदर्शन करते हैं। वे विभिन्न अनुप्रयोगों में सटीक और व्यक्तिगत अनुशंसाओं, बेहतर सामग्री टैगिंग और समृद्ध उपयोगकर्ता अनुभवों का मार्ग प्रशस्त करते हैं। प्रत्येक विधि के लिए कोड XML रिपॉजिटरी में उपलब्ध है।

Contents

List of Figures

List of Tables