Universitatea Tehnică a Moldovei

Identificarea mesajelor instigatoare a imaginilor multimodale

Student: **Ungureanu Mihail**

Coordonator: **Peca Ludmila**
lect. univ., mag.

Chișinău, 2022

MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei
Facultatea Calculatoare Informatică şi Microelectronică

Departamentul Ingineria Software şi Automatică

Admis la susţinere
Şef departament:
Fiodorov I. dr., conf.univ.

_____

„____”_____ 2022

Identificarea mesajelor instigatoare a imaginilor multimodale

Teză de master

Student: **Ungureanu Mihail, TIA-201M**
Conducător: **Peca Ludmila, lect. univ., mag.**
Consultant: **Bodoga Cristina, asis. univ.**

Chişinău, 2022

REZUMAT

Lucrarea este alcătuită din introducere, 3 capitole, concluzii și bibliografie.

Capitolul 1 descrie importanța temei, analiza domeniului și importanța cercetărilor din acest domeniu. În acest capitol sunt analizate materialele teoretice MML(Multimodal Machine Learning) și sublinierea provocărilor care trebuie rezolvate pentru a crea un model funcțional.

Capitolul 2 descrierea și analiza modelelor existente. Este descrisă arhitectura celor mai de success modele în domeniu, specificarea atât avantajelor cât și dezavantajele fiecărui model. Sunt prezentate diagrame, figuri și formule cu exemple care demonstrează acest fapt.

Capitolul 3 descrie integrarea a mai multor modele existente precum BERT, UNITER într-un întreg pentru a acoperi probleme și cazuri mai diverse. Sunt aduse exemple de optimizare a modelului prin diversificarea setului de date cât și îmbunatățirea algoritmului de predicție.

Aplicația teză de master reprezintă un model scris în limbajul Python. Acest fapt a fost atins în mai multe etape.Prima fiind asamblarea unui set de date, adnotarea datelor și reeveluarea acestuia în timpul antrenării în scopul ridicării preciziei a rezultatelor. A doua etapa e crearea modelului propriuzis dupa analiza solutiilor existente.

ABSTRACT

The paper consists of an introduction, 3 chapters, conclusions and bibliography.

Chapter 1 describes the importance of the topic, the analysis of the field and the importance of research in this area. In this chapter, we take a look at the theoretical side of MML(Multimodal Machine Learning) and at the challenges that need to be solved to create a functional model are outlined.

Chapter 2 describes and analyses the existing models. The architecture of the most successful models in the field is described, specifying both advantages and disadvantages of each model. Diagrams, figures and formulae are presented with examples to demonstrate this.

Chapter 3 describes the integration of several existing models such as BERT, UNITER into a whole to cover more diverse problems and cases. Examples are given of model optimization by diversifying the dataset as well as improving the prediction algorithm.

The master thesis application represents a model written in the Python. This was achieved in several steps.The first being assembling a dataset, annotating the data and re-developing it during training in order to raise the accuracy of the results. The second stage is the creation of the model itself after analysis of existing solutions.

CUPRINS

INTRODUCERE

Pentru a evita transmiterea informațiilor ironate despre obiectele din lumea înconjuratorare, diverse semnale cognitive care descriu diferite aspecte ale aceluiași obiect sunt înregistrate în diferite moduri, cum ar fi text, imagine, video, sunet. Cuvântul „modalitate” se referă la un anumit mecanism de codificare a informațiilor. Prin urmare, diferitele tipuri de media enumerate mai sus se referă la modalități, iar sarcinile de învățare a reprezentării care implică mai multe modalități vor fi caracterizate ca multimodale.

Deoarece datele multimodale descriu un obiect din puncte de vedere diferite, de obicei complementare sau suplimentare în conținut, ele sunt mai informative decât datele unimodale. De exemplu, cercetările privind recunoașterea vorbirii au arătat că modalitatea vizuală oferă informații despre mișcările buzelor și articulațiile gurii, inclusiv deschiderea și închiderea, astfel încât poate ajuta la îmbunătățirea performanței de recunoaștere a vorbirii. Prin urmare, este valoros să se exploateze semantica cuprinzătoare oferită de mai multe modalități. Deși este ușor pentru ființele umane să perceapă lumea prin informații cuprinzătoare de la mai multe organe senzoriale cum ar fi auz,miros, senzații tactile,vedere, cum să transmitem aceste informații mașinilor cu capacități cognitive analoge este încă o întrebare deschisă, aceasta fiind una dintre provocările cu care ne confruntăm [1]. O metodă populară pentru abordarea acestei probleme este proiectarea caracteristicilor eterogene într-un subspațiu comun, unde datele multimodale cu semantică similară vor fi reprezentate de vectori similari [2].

Astfel, obiectivul principal al învățării reprezentării multimodale este reducerea decalajului de distribuție într-un subspațiu semantic comun, păstrând în același timp intactă semantica specifică modalității. Pentru a reduce decalajul de eterogenitate, în ultimele decenii au fost efectuate numeroase cercetări cu abordări diferite. Ca rezultat, progresul învățării reprezentării multimodale a beneficiat de o mulțime de aplicații. De exemplu, se poate obține performanță îmbunătățită în sarcinile de analiză cross-media, cum ar fi clasificarea video, detectarea evenimentelor și analiza emoțiilor. În plus, prin exploatarea similitudinii intermodale sau a corelației intermodale, devine posibil pentru noi să căutăm imagini folosind o propoziție ca input, sau invers, acest process fiind numit cross-modal retrieval. [3].

Scopul acestei lucrări e de a oferi un studiu complet asupra procesului de deep learning, a reprezentării multimodale și de a sugera direcția viitoare în acest domeniu. În general, sarcinile de automatizare bazate pe date multimodale includ trei pași necesari: extragerea caracteristicilor specifice modalității, învățarea reprezentării multimodale care urmărește să integreze diverse caracteristici din diferite modalități într-un subspațiu comun și o etapă de raționament cum ar fi clasificarea sau gruparea. O caracteristică crucială a provocării este faptul că includem așa-numiții "factori de confuzie benigni" pentru a contracara posibilitatea ca modelele să exploateze prejudecăți unimodale: pentru fiecare exemplu negative, găsim imagini sau subtitrări alternative care fac ca eticheta să se întoarcă la starea inițială.

BIBLIOGRAFIE

1. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A. Y. Ng, "*Multimodal deep learning*", p. 689-696, 2011.
2. N. Rasiwasia et al., "*A new approach to cross-modal multimedia retrieval*", p. 251-260, 2010.
3. F. Feng, X. Wang and R. Li, "*Cross-modal retrieval with correspondence autoencoder*", p. 7-16, 2014.
4. L. Deng and Y. Liu, "*Deep Learning in Natural Language Processing*". Springer, 2018.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "*BERT: Pretraining of deep bidirectional transformers for language understanding*", 2019.
6. S. Kumar and R. Udupa, "*Learning hash functions for cross-view similarity search*," in IJCAI, 2011.
7. M. Soleymani, M. Pantic, and T. Pun, "*Multimodal emotion recognition in response to videos*," TAC, 2012
8. T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda, "*Text-to-Visual Speech Synthesis Based on Parameter Generation from HMM*," in ICASSP, 1998
9. C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler, "*What are you talking about? Text-to-Image Coreference*," in CVPR, 2014.
10. R. Bernardi, R. Cakici, D. Elliott, Muscat, and B. Plank, "*Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures*" ,2016
11. K. Sjolander, "*An HMM-based system for automatic segmentation and alignment of speech*," in Proceedings of Fonetik, 2003
12. K. Xu, "*Show, attend and tell: Neural image caption generation with visual attention*"
13. Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai. "*VL-BERT: Pre-training of Generic Visual-Linguistic Representations"*, 2020
14. Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "*Improving language understanding by generative pre-training"*. 2018.
15. Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. "*Stacked cross attention for image-text matching."* In ECCV, 2018