Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 621.39
DOI: https://doi.org/10.24867/28BE43Djurkic

МУЛТИМОДАЛНО ПРЕПОЗНАВАЊЕ ЕМОЦИЈА ПОМОЋУ КОМПРИМОВАНИХ ГРАФОВСКИХ НЕУРОНСКИХ МРЕЖА

MULTIMODAL EMOTION RECOGNITION USING COMPRESSED GRAPH NEURAL NETWORKS

Тијана Ђуркић, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *У раду је прво описан један алгоритам за препознавање емоција на основу звука, текста и видео записа, уз коришћење графовских неуронских мрежа. Потом су представљени резултати примене компресије на модел, која је потребна како би се обимни модели могли користити и на мањим уређајима.*

Кључне речи: *графовске неуронске мреже, препознавање емоција, мултимодални подаци, компресија*

Abstract – *In this paper, first, an algorithm for emotion recognition based on sound, text and video, using graph neural networks is described. Then, the results of applying compression to the model are present, which is needed so that large models can be used on smaller devices.*

Keywords: *graph neural networks, emotion recognition, multimodal data, compression*

1. УВОД

Различити уређаји постају све доступунији и заступљенији у свакодневном животу. Због тога човек више времена него раније проводи користећи уређаје. Тиме се смањило време које је раније било посвећено разговору са другим људима. Управо је због тога пожељно да интеракција са уређајима што је могуће више личи на на комуникацију између два човека.

Пошто се комуникација одвија путем видеа, звука и текста, постоји захтев за моделом који би успешно обрађивао сва три модалитета. Алгоритам који се показао успешним за решавање овог проблема јесте графовске неуронске мреже, *GNN* (енгл. – *Graph Neural Networks*). Главни задатак овог рада јесте да покаже теоријски преглед овог алгоритма, детаљнију анализу рада који се бавио овом темом, као и резултате компресије параметара. У другом поглављу дате су опште информације о емоцијама и теоријски преглед *GNN* методе и компресије, у трећем детаљна анализа рада који је послужио као узор, у четвртом допринос и резултати, а у петом закључак и даљи рад.

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био доцент др Синиша Сузић.

2. ТЕОРИЈСКИ ПРЕГЛЕД

У овом поглављу ће детаљније бити представљена теорија која стоји иза *GNN* методе и компресије.

2.1 Емоције

Емоције су одговор на дешавања око нас. Према Нику Фрицди, емоције су исходи човековог приступања свету односно његовог доживљаја околине које касније модификују реакције на дешавања око њега и поступке који се предузимају као одговор на та дешавања [1]. Због значаја које емоције имају у природној комуникацији, пожељно је да буду укључене и у интеракцију човек – машина како би коришћење уређаја било приближније човековој природи..

2.2 Графовске неуронске мреже

Графовске неуронске мреже су подгрупа неуронских мрежа. Основа неуронске мреже јесте перцептрон (неурон) [2]. Узорак се класификује у зависности од тога да ли је вредност излаза испод или изнад одређеног (бинарна класификација). Уместо прага, користи се слободан члан *p* који се назива померајем (енгл. – *bias*). За сложеније проблеме користи се комбиновање неурона у неуронске мреже.

Корен графовских неуронских мрежа су графови. Граф представља везе између групе ентитета који се називају чворови. Ивице или гране повезују чворове и описују њихову повезаност [3]. Основна идеја *GNN* алгоритма је да се сваки чвор графа моделује на основу веза са суседним чворовима чиме се обезбеђује да се науче сложене репрезентације чворова и целокупног графа. То се постиже итеративним ажурирањем стања или карактеристика чвора на основу информација из „суседства“. Прво се прикупе информације из околине и комбинују се, најчешће помоћу неке функције. Ти прикупљени подаци се користе за ажурирање стања чвора кроз неку функцију која укључује неуронску мрежу. Врши се комбиновање тренутних карактеристика чвора и информација из суседних.

Најважнији параметри за *GNN* алгоритам су тежине веза (енгл. - *weights*) и померај (енгл. – *bias*). Тежине представљају јачине веза између чворова у суседним слојевима неуронске мреже, а утичу и на смер преноса информација. Што је утицај једног чвора на други већи, то је и вредност тежинског коефицијента већа. Слично као што је раније поменуто, померај се додаје излазу сваког чвора пре примене активационе

функције. Он доприноси померању излаза дуж осе вредности и тиме даје могућност моделу да уочи структуре у подацима. Без *bias*-а би се вршило само линеарно пресликавање улаза на излаз. Подешавањем оптималних тежина и помераја приликом обуке, модел учи и прилагођава се подацима из скупа за обуку што касније доводи до успешнијих перформанси приликом решавања проблема.

2.3 Компресија

Компликованији задаци изискују више слојева и параметара. Ипак, велики модели у погледу броја параметара нису практични када их је потребно применити у реалном времену због ограничености меморијских ресурса и хардверских компоненти, нпр. у мобилним телефонима. Такође се показало да су ти модели препараметризовани (енгл. – *over-parametrized*), односно да толики број параметара заправо није потребан [4]. Да би се постигла боља ефикасност извршавања, са задовољавајућим очувањем резултата, потребно је вршити компресију.

У овом раду су примењене две методе квантизације и оне ће бити објашњене у наредним поглављима.

3. ОСНОВНИ РАД

Рад који је послужио као основа за израду овог рада заснован је на архитектури названој COGMEN описаној у [5]. Као узорци за препознавање емоција користе се аудио, видео и текст разговора јер се информације из ова три различита модалитета међусобно надопуњују и дају потпунији утисак. Оно што издваја овај рад од осталих на сличну или исту тему јесте да се узимају у обзир и утицај контекста разговора (глобалне информације) и локалне информације односно међусобна зависност саговорника, као и зависност од самог појединачног говорника на временски блиске реченице. Циљ је препознати емоцију изражену у једној реченици говорника уз све горе поменуте новитете који су уведени.

Како би се препознао контекст (глобална информација) и његов утицај на сваку појединачну реченицу (узорак) користи се трансформерски кодер [6] који користи механизам пажње за уочавање односа између различитих делова секвенци и њихово обрађивање. Што се тиче локалних информација, емоција изражена у једној реченици је често плод утицаја околних, суседних реченица, па је било потребно утврдити утицај међу саговорницима и утицај саговорника на самог себе. За ове потребе развијен је граф где свака реченица представља чвор, а усмерене ивице различите повезаности где је важан редослед реченица. Битно је напоменути да се једна реченица састоји из аудио, видео и текстуалне репрезентације, па је самим тим и чвор мултимодалне природе. Разликује се усмерена веза између реченица које изговара један говорник и усмерена веза између реченица које потичу од више њих.

3.1 Архитектура модела

На слици 1 приказана је оригинална архитектура модела. Улазне реченице прво наилазе на издвајач контекста, у слободном преводу (енгл. – *context extractor*). Спојена обележја сва три модалитета (видео, аудио и текст) користе се као улаз за сваку реченицу дијалога. Како би се препознао контекст (глобална информација) и његов утицај на сваку појединачну реченицу (узорак) издвајач контекста користи кодер трансформера [6].

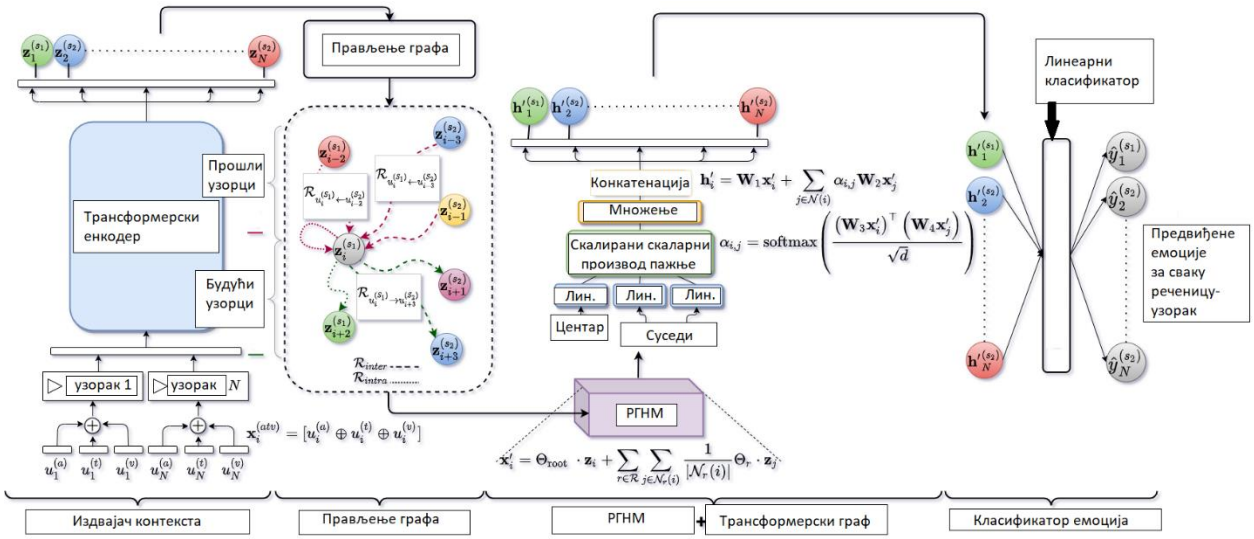
Главни део трансформерског кодера јесте механизам вишеструке пажње (енгл. – *multi-head attention mechanism*) чија је улога да омогући декодеру да искористи делове улазних секвенци који су најрелевантнији додељивањем тежинских коефицијената енкодованим улазним векторима, где се највеће вредности коефицијената приписују најбитнијим узорцима. Последњи слој кодера је мрежа са пропагацијом унапред. На крају се добија вектор обележја за сваку реченицу.

На основу добијених обележја из кодера, прави се граф који уочава везе између реченица дијалога. Као што је већ напоменуто, сваки чвор графа представљен је једном реченицом у сва три модалитета које су повезане усмереним ивицама (гранама). Разликујемо гране које повезују реченице које изговара исти говорник и гране које повезују реченице различитих говорника. Такође, прате се будуће и претходне релације сваке реченице односно за сваку реченицу се зна које су јој све претходиле, а које све следовале.

Наредна целина је релациона графовска неуронска мрежа (енгл. – *Relational Graph Convolutional Network (RGCN)*) која је предложена у ранијем раду [7]. Њена улога је прикупљање трансформација специфичних за однос међу суседнимчворовима које зависе од типа и смера грана и то све кроз нормализовану суму. За разлику од класичног *GNN* алгоритма који све ивице третира подједнако, ова мрежа уочава разлику између типова ивица узимајући у обзир и специфичности веза које те чворове повезују. На тај начин *RGCN* кодира и информације о самим узорцима, али и њихове међусобне релације.

У раду од интереса, овај модел мреже примећује зависност повезаних реченица од самог говорника и више саговорника. Графовска неуронска мрежа у овом раду има 52 слоја где као параметри фигуришу тежине (енгл. – *weights*) и помераји (енгл. – *biases*). Има 8 слојева који садрже преко милион параметара. За издвајање лабела и обележја из чворова користи се трансформерски граф (енгл. – *graph transformer*) [8] који те информације користи као улаз у пропагацију кроз мрежу. Циљ је да на различите начине приступа суседним чворовима на основу њихових међусобних веза и да тиме уочава структуру графа. Предност трансформерског графа је што уочава дугорочне зависности кроз механизам самопажње (енгл. – *self-attention mechanism*), који омогућава моделу да одреди значај улазне секвенце у односу на друге делове улаза.

Додатне предности су да се паралелним рачунањем смањује време извршавања и обуке, а узимањем у обзир дугорочних зависности нема ограничења у погледу меморијских или рачунарских ресурса. На самом крају целокупног модела налази се класификатор емоција који се састоји из једног линеарног слоја.



Слика 1. *Архитектура оригиналног модела [5]*

3.2 Резултати оригиналног рада

Перформансе модела проверене су на две базе података.

IEMOCAP (енгл. – *The Interactive Emotional Dyadic Motion Capture*) је мултимодална база података за препознавање емоција где је свака реченица обележена једном од 6 емоција [9]. Она се састоји из снимљених видео записа дијалога глумаца од којих је тражено да одглуме одређене емоције.

MOSEI (енгл. – *Mulitmodal Opinion Sentiment and Emotion Intensity)* је друга широко распострањена мултимодална база са 6 емоција: срећа, туга, гађење, страх, изненађеност и љутња [10]. Метрике које су коришћење за евалуацију оригиналног модела су макро средња вредност прецизности, осетљивости и *F*-мере која износи 82.39%, 81.24% и 81.66% респективно, потом пондерисана средња вредност прецизности, осетљивости и *F*-мере, од 82.22%, 81.97% и 81.96% респективно и просечна тачност од 81.97% и просечна *F*-мера од 81.96%.

Показан је велик утицај контекста на одлучивање о присутној емоцији јер се повећањем броја секвенци из дијалога, добија бољу увида у контекст целокупног дијалога, постижу веће вредности F-мере. Такође, уколико се изостави *GNN* из архитектуре модела примећује се пад перформанси, јер се уз помоћ *GNN* боље издвајају и групишу реченице у којима су изражене исте емоције. Тиме се даје на значају локалним информацијама које су раније објашњене.

4. ДОПРИНОС ОРИГИНАЛНОМ РАДУ

Као надоградња на постојећи рад, примењена је квантизација на већ истрениран модел чији је циљ смањити величину модела у погледу величине фајла и убрзати његово извршавање смањењем прецизности параметара, пре свега тежина. Тиме се уједно смањује и потрошња енергије. На основу хистограма слојева мреже тешко је било утврдити неку одређену расподелу, па је одлучено да се примени бинарна квантизација као веома лака за имплементирање и *FP8* (енгл. – *Floating Point 8*) као веома корисна и све коришћенија метода.

4.1 Бинарна квантизација

Бинарна квантизација мења улазну вредност користећи само један бит. Један од изазова бинарне квантизације јесте балансирање између компресије података и очувања квалитета.

Претварање континуираних или вредности широког опсега у само две могуће може довести до губитка битних детаља и информација, што у појединим ситуацијама може бити неприхватљиво.

4.2 *Floating point 8*

Користећи *Floating Point,* реални бројеви се представљају са фракционим делом. Фракциона компонента су цифре после зареза у децималном запису бројева. *Floating Point* бројеви се састоје из три компоненте: бит знака (енгл. – *the sign bit*), експонент и мантиса (фракција). Бит знака даје информацију о томе да ли је број позитиван или негативан, експонент одређује скалу, а мантиса прецизност или значајне цифре [11].

За *FP8* важи да је то формат смањене прецизности који користи 8 бита за енкодовање. Улазни подаци биће квантизовани на вредности које се могу представити овим бројем бита, а на основу оригиналних података и формула које иду уз овај начин квантизације.

Као надоградња на постојећи рад, бинарна квантизација и *FP8* су прво примењени на слојевима који имају преко милион параметара и њих има 8. Пошто они заузимају највише меморијског простора због свог броја, а један од главних циљева квантизације јесте уштеда меморијског простора, очекивано је да ти слојеви буду први квантизовани. Потом су сви слојеви квантизовани и анализиран је утицај квантизације на резултате предвиђања модела.

4.3 Резултати

Пошто су резултати бољи када се примени квантизација на све параметре, само ће они бити представљени у раду. Може се приметити знатан пад у перформансама модела приликом предвиђања емоција када се примени бинарна квантизација.

Макро средња вредност прецизности, осетљивости и *F*-мере износи 49.42%, 36.85%, 35.25% респективно, потом пондерисана средња вредност прецизности, осетљивости и *F*-мере, од 48.13%, 44.75% и 35.25% респективно и просечна тачност од 44.75% и просечна *F*-мера од 39.54%. Закључено је да неуниформна расподела података над којима је примењена квантизација утиче на пад перформанси [12], што је могуће да је узрок толико слабим резултатима јер се из хистограма види да расподела вредности не одговара Лапласовој расподели око нуле, а у том случају би бинарна квантизација била погодан метод за компресију.

У општем случају, уколико је узрок лошијих резултата неуниформна расподела, тада се врши нормализација током финог подешавања (енгл. – *fine tuning*) или нелинеарна квантизација (нпр. логаритамска). Такође, могуће је да комбинација обичних графовских неуронских мрежа и детерминистичке бинарне функције није предодређена за постизање задовољавајућих резултата.

Овој архитектури модела много више одговара *FP8* метода за добијање жељених резултата, Могуће објашњење је да је можда оригинални модел могао боље да се истренира, односно да је његова обука заустављена у локалном минимуму. Узрок је то што се *FP8* методом вредности параметара заокружују на најближе квантизационе нивое који се само мало разликују од оригиналних вредности.

Приликом тренирања модела, вредности параметара се такође мењају у малим корацима. Када се тачност мало побољша, то значи да је хипотетички, примењено даље учење модела јер су вредности параметара мењане за мале вредности које одговарају кораку учења, тако да је и у раду [11] постигнуто побољшање модела.

Макро средња вредност прецизности, осетљивости и *F*-мере износи 80.90.42%, 81.93%, 81.39% респективно, потом пондерисана средња вредност прецизности, осетљивости и *F*-мере, од 81.86%, 81.76% и 81.78% респективно и просечна тачност од 81.76% и просечна *F*-мера од 81.78%.

5. ЗАКЉУЧАК И ДАЉИ РАД

На основу изнетих резултата, може се закључити да компресија модела неком од врста квантизације, на овом раду *FP8*, не мора да има негативан утицај на саму тачност модела, чак може и да је мало побољша. *FP8* метода се показала као много успешнија од бинарне квантизације.

У даљем раду могло би да се испита каквим проблемима компресије и типовима модела више одговара примена бинаризације. Ово би било значајно јер бинаризација није ни временски ни рачунски захтевна метода, па би се резултати добијали много брже, а ако се покаже да за одређене случајеве не нарушава перформансе, била би јако добра за примену. Такође, могли би се испитати неки други начини компресије или њихова комбинација и проверити да ли то додатно поспешује или оштећује модел.

6. ЛИТЕРАТУРА

[1] N.H. Frijda, The emotions, Cambridge University Press; 1986.

[2] Т. Носек, Б. Бркљач, Д. Деспотовић, М. Сечујски, Т. Лончар-Турукало, „Практикум из машинског учења“, Факултет техничких наука, Универзитет у Новом Саду, 2020.

[3] H. Abdi, D. Valentin, & B. Edelman, *Neural networks* (No. 124). Sage, 1999.

[4] J. O. Neill, An overview of neural network compression. *arXiv preprint arXiv:2006.03669*, 2020

[5] A. Joshi, A. Bhat, A. Jain, A. V. Singh, & A. Modi, COGMEN: COntextualized GNN based multimodal emotion recognitioN. *arXiv preprint arXiv:2205.02455*, 2022

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... & I. Polosukhin, Attention is all you need. *Advances in neural information processing systems*, *30*, 2017

[7] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, & M. Welling, Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15* (pp. 593-607). Springer International Publishing, 2018

[8] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, Y. Sun, Masked label prediction: Unified message passing model for semi-supervised classification. arXiv preprint arXiv:2009.03509, Sep 8, 2020

[9] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, ... & S. S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*, 335-359, 2008

[10] A. B. Zadeh, P. P .Liang, S. Poria, E. Cambria, & L. P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2236-2246), July, 2018

[11] Н. Симић, С. Сузић, Т. Носек, М. Вујовић, З. Перић, М. Савић, и В. Делић, Speaker recognition using constrained convolutional neural networks in emotional speech. *Entropy*, *24*(3), 414, 2020

[12] T. Liang, J. Glossner, L. Wang, S. Shi, & X. Zhang, Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, *461*, 370-403, 2021

Кратка биографија:

Тијана Ђуркић рођена је у Новом Саду 1998. године, а одрасла у Бечеју. Дипломски рад одбранила је на Факултету техничких наука из области машинског учења. Контакт: tijana.djurkic@gmail.com