Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Dirbtinio intelekto metodai radiologiniais skaitmeniniais vaizdais grįstų klausimų atsakymams prognozuoti

Baigiamasis magistro studijų projektas

Dainius Gaidamavičius

Projekto autorius

Doc. Dr. Tomas Iešmantas

Vadovas

Kaunas, 2022

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Dirbtinio intelekto metodai radiologiniais skaitmeniniais vaizdais grįstų klausimų atsakymams prognozuoti

Baigiamasis magistro studijų projektas

Taikomoji matematika (6211AX006)

Dainius Gaidamavičius

Projekto autorius

Doc. Dr. Tomas Iešmantas

Vadovas

Doc. Dr. Audrius Kabašinskas

Recenzentas

Kaunas, 2022

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Dainius Gaidamavičius

Dirbtinio intelekto metodai radiologiniais skaitmeniniais vaizdais grįstų klausimų atsakymams prognozuoti

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Dainius Gaidamavičius

Patvirtinta elektroniniu būdu

Gaidamavičius Dainius. Dirbtinio intelekto metodai radiologiniais skaitmeniniais vaizdais grįstų klausimų atsakymams prognozuoti. Magistro studijų baigiamasis projektas / Vadovas Doc. Dr. Tomas Iešmantas; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (Matematikos mokslai).

Reikšminiai žodžiai: gilieji; dirbtiniai; tinklai; radiologiniai; vaizdai; klausimais; analitika; dirbtinio; intelekto; metodai;

Kaunas, 2021. 64 p.

Santrauka

Remiantis Pasaulio Sveikatos Organizacijos duomenimis, daugiau nei 45 % pasaulio šalių turi mažiau nei vieną gydytoją tūkstančiui gyventojų. [1] Dėl šios priežasties medicinos srities darbuotojai yra priversti dirbti didelį kiekį valandų, o tai lemia, didėjantį žmogiškųjų klaidų skaičių. Kadangi nemažą gydytojų darbo dalį sudaro medicininių vaizdų nagrinėjimas, gilusis mokymas gali palengvinti specialistų darbą automatizuodamas vaizdų apdorojimą. Šiais laikais tyrėjai skiria vis didesnį dėmesį klausimais grįstai vaizdo analitikai, kuri įgalina modelius suprasti vaizde esančias problemas žymiai platesniu kontekstu nei kad tik specifiniai problemai skirti modeliai. Šiame darbe buvo nagrinėjamos giliojo mokymo metodikos, kurių tikslas buvo pateikti atsakymus į užduotus klausimus apie radiologinius vaizdus.

Darbe buvo taikomi ResNet50V2, EfficientNetB0 sąsūkų neuroniniai tinklai, ilgos-trumpos atminties, rekurentiniai neuroniniai tinklai, bei pateikiamos naujos idėjos vaizdo ir teksto požymių suliejimui. Tyrimas parodė, jog geriausia modelių architektūra naudojosi požymių vektorių apjungimu į matricą, kuriai toliau buvo pritaikoma sąsūkos operacija. Aukščiausias vidutinis tikslumas buvo pasiekiamas su ResNet50V2 sąsūkų tinklu bei LSTM neuroniniu tinklu ir buvo lygus 84,80 %.

Gaidamavičius Dainius. Artificial Intelligence Methods for Predicting Questions Answers Based on Digital Radiology Images/ Supervisor Doc. Dr. Tomas Iešmantas; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied Mathematics (Mathematical Sciences).

Keywords: deep; neural; networks; radiology; images; questions; analytic; artificial; intelligence; methods

Kaunas, 2021. 64.

Summary

According to World's Health Organization data more than 45% world countries have less than one physician for thousand residents. [1] Because of this reason medical staff are forced to work large amount of hours and this situation leads to growing human error rate. Since physician's work requires a lot of time to analyze medical images deep learning could come to hand and automize image processing. Nowadays, machine learning researchers invest more and more attention into visual question answering problems and these endeavors lead to models which can understand wider context in given image and solve problems which requires human level reasoning. In this work were analyzed deep learning methods which goal was to predict answers about digital radiography images.

In this work were applied ResNet50V2, EfficientNetB0 convolutional neural networks, long-short term memory, recurrent neural neworks and proposed new image and text feature vectors fushion options. Experiments have shown that best architecture used feature fushion by adding feature vectors into custom matrix and further applying convolution layers. Highest mean average value was reached by ResNet50V2 and LSTM networks and was equal to 84,80%.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Santrumpų ir terminų sąrašas

Santrumpos:

VGG – Viena iš sąsūkos neuroninių tinklų architektūrų.

LSTM – ilgos-trumpos atminties tinklas (sluoksnis) (angl. long-short term memory)

VQA – vaizdais grįstas atsakinėjimas (angl. visual question answering)

COCO – plačiai taikoma duomenų imtis mašininio mokymosi tematikose

ImageNet - plačiai taikoma duomenų imtis mašininio mokymosi tematikose

WORD2VEC – algoritmas žodžiams atvaizduoti į vektorius išlaikant semantinę prasmę

GloVe - algoritmas žodžiams atvaizduoti į vektorius išlaikant semantinę prasmę

Faster R-CNN – sąsūkų neuroninis tinklas skirtas objektų aptikimui

BERT – dvikryptis transformeris

MFB - multi-modalinis faktorizuotas telkimas

VQA-MED 2019 – radiologinių vaizdų ir klausimų-atsakymų rinkinys

VQA-MED 2018 – radiologinių vaizdų ir klausimų-atsakymų rinkinys

VQA- Med 2020 - radiologinių vaizdų ir klausimų-atsakymų rinkinys

Skip n Gram – metodas skirtas atvaizduoti žodžius į latentinė erdvę, išlaikant semantinę prasmę

Įvadas

Problema: Medicininius vaizdus yra sudėtinga teisingai išnagrinėti, kuomet specialistas neturi sukaupęs didelės patirties. Kadangi visame pasaulyje trūksta gydytojų, dažnai yra susiduriama su specialistų pervargimu, o tai sąlygoja žmogiškąsias klaidas, nustatant pacientų diagnozes. Tokiu atveju būtų labai palanku turėti puikiai veikiančią klausimais grįstą vaizdo analitiką, kuri galėtų pateikti papildomą nuomonę apie medicinoje nagrinėjamą konkretų atvejį[1]. Pažymime, jog, turint tokią sistemą, gydytojų darbas nagrinėjant radiologinius vaizdus ne tik sutrumpėtų, bet ir būtų žymiai tikslesnis. Kaip bebūtų, šiais laikais kuriami dirbtinio intelekto metodai dar nepasižymi labai dideliais tikslumais dėl atvirų duomenų rinkinių trūkumo. Tam, kad būtų pasiekiami vis didesni tikslumai su turimais viešais medicininiais vaizdais yra kuriami įvairūs mašininio mokymo modeliai. Šiame darbe ne tik apžvelgsime jau taikytus tyrimus, bet ir pateiksime papildomais klausimais grįstos vaizdo analitikos idėjas.

Tyrimų aktualumas: Remiantis Pasaulio Sveikatos Organizacijos duomenimis, daugiau nei 45 % pasaulio šalių turi mažiau nei vieną gydytoją tūkstančiui gyventojų[1]. Dėl to kiekvienas medikas yra priverstas išnagrinėti didelį kiekį medicininių atvejų, o visa tai lemia padidėjusią žmogiškosios klaidos riziką. Kompiuteriniais skaičiavimais paremtos medicininės diagnozės ne tik sugeba sumažinti klaidos riziką, bet, kaip ir minėjome darbo problematikoje, sugeba įžvelgti net mažiausias detales skaitmeniniuose vaizduose, kurios yra nematomos plika žmogaus akimi. Šiais laikais jau yra plačiai taikomi medicininiai portalai, kuriuose pacientai gali gauti ekspertų konsultacijas, o pritaikius klausimais grįstą vaizdo analitiką didelis gydytojų krūvis gali būti ženkliai sumažintas.

Naujumas: Nors gilusis mokymas pradėtas plačiai taikyti jau šio amžiaus pirmajame dešimtmetyje įvairioms problemoms pasaulyje spręsti, daugelis modelių nepasižymėjo gebėjimu įvykdyti kelis uždavinius tuo pačiu metu ne tik dėl skaičiavimų resursų, bet ir dėl architektūrinių modelių idėjų trūkumo. Kaip bebūtų, šiuo metu vis plačiau atliekami tyrimai bendrinio dirbtinio intelekto plėtojimo kryptimi, t. y. kuriamos įvairios metodikos, kurios įgalina modelius suprasti platesnį pasaulio kontekstą, o vis didėjantys skaičiavimo resursai lemia tai, jog dabartiniai modeliai sugeba greitai apdirbti ne tik skaitmeninius vaizdus, bet ir pateiktus apie juos klausimus. Ši priežastis bei vis didesnis tyrėjų dėmesys klausimais grįstai vaizdo analitikai lemia, jog dabartiniai modeliai, pateikus klausimą apie specifinį vaizdą, sugeba rasti tinkamą atsakymą, nepaisant to, kad vaizdų ir klausimų įvairovė yra labai plati.

Darbo tikslas – Sudaryti ir ištirti kalbos ir vaizdų modeliais grįstą algoritmą, skirtą klausimų apie radiologinius vaizdus atsakymams prognozuoti.

Darbo uždaviniai:

1. Apžvelgti ir išanalizuoti literatūrą, susijusią su giliojo mokymosi metodų taikymu klausimais grįstoje vaizdo analitikoje;
2. Pasirinkus radiologinių vaizdų, bei klausimų/atsakymų duomenų imtį, ištirti dažniausiai naudojamas kalbos ir vaizdų požymių apjungimo strategijas: sumavimą, sandaugą, prijungimą (angl. concatenation);
3. Pasiūlyti naują vaizdo ir kalbos požymių apjungimo būdą, kuris leistų padidinti modelio tikslumą.

1. Literatūros apžvalga

Šiame skyriuje pateiksime jau atliktų darbų, susijusių su klausimais grįsta vaizdo analitika, apžvalgą. Pirmiausia panagrinėsime tematiką, susijusią su duomenimis, kurie apima kasdienius vaizdus, o paskui apžvelgsime atliktus tyrimus medicininių duomenų apdorojimo srityje. Taip pat išskirstysime šiuos skyrius į mažesnius skyrelius pagal taikomas modelių architektūrines idėjas.

1.1. Atlikti tyrimai su ne medicinos srities duomenimis

1.1.1. Vaizdo bei teksto požymių sujungimas konkatenacijos bei sandaugos operacijomis

Dažniausiai darbe nagrinėjamai tematikai yra pasirenkami sąsūkų neuroninių tinklų paskutiniai sluoksniai, kurie atsakingi už vaizdinės informacijos apskaičiavimą, o tuo tarpu rekurentiniai tinklai dažnai yra atsakingi už tekstinės informacijos išgavimą. Kuomet turime šiuos vektorius, juos norime sujungti, kaip įmanoma daugiau išlaikydami kontekstinės informacijos tiek apie vaizdą, tiek apie klausimą. Šiam darbui atlikti daugelis autorių pasirenka sandaugos arba konkatenacijos operacijas. Dėl šios priežasties šiame skyrelyje apžvelgsime naudojantis šiomis operacijomis atliktus tyrimus bei pateiksime platesnį problemos apibrėžimą, pasitelkdami vaizdinius pavyzdžius.

Klausimais grįstą vaizdo analitiką galima apibrėžti kaip uždavinį, kuomet programos tikslas yra, gavus skaitmeninį vaizdą bei klausimą, sugeneruoti teisingą atsakymą. Norime pabrėžti, jog teisingas atsakymas gali būti pasirenkamas iš baigtinio skaičiaus galimų atsakymų, bet lygiai taip pat gali būti nagrinėjama problema, kai modelis turi rinktis ne iš baigtinio skaičiaus klasių, bet sugeneruoti skirtingo ilgio atsakymus. Kaip galima suprasti iš uždavinio apibrėžimo, modelis turi gebėti ne tik teisingai interpretuoti, kokie objektai, požymiai yra vaizde, bet ir sugebėti suvokti vaizdo bei klausimo kontekstą. Žemiau esančiuose paveikslėliuose pateikiame keletą vaizdų kaip pavyzdžius, kuriuos nagrinėjo [2] straipsnio autoriai.

1 pav. [2] tyrimo pirmasis pavyzdys

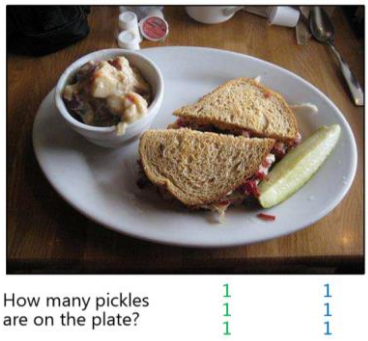
2 pav. [2] tyrimo antrasis pavyzdys

Iš 1 ir 2 paveikslėlių galima matyti, jog, uždavus sistemai klausimą „kiek bananų yra pirmoje nuotraukoje?“, modelis privalo tinkamai atskirti objektus, t. y. uždavinys tampa objektų aptikimo problema, tuo tarpu, jei užduotume klausimą „ar vyras antrajame paveikslėlyje pasižymi gera rega?“, modelis privalėtų suprasti kontekstą ir pagal tai, kaip vyras žiūri į tam tikrą objektą, pateikti atsakymą. [2] tyrimo autoriai teigia, jog anksčiau apibrėžta sistema turi pasižymėti dirbtinio intelekto savybėmis, kurios sugeba sėkmingai spręsti šias problemas: [2]

- objektų aptikimo;
- veiklos atpažinimo;

- konteksto supratimo.

[2] tyrime autoriai naudojosi MS COCO duomenų rinkiniu, kuris turėjo 204 721 vaizdus bei taip pat buvo papildytas 50 000 naujų scenų. Šis rinkinys buvo puikus pasirinkimas, kadangi vaizduose yra didelė scenų įvairovė, kurioms galima sugalvoti didelį kiekį skirtingų klausimų. Naudotoje imtyje kiekvienas vaizdas turėjo 3 klausimus. Iš viso imtyje buvo daugiau nei 760 000 klausimų ir apie 10 000 000 atsakymų. 3 ir 4 paveikslėliuose pateikiame tyrime naudotų vaizdų ir klausimų porų pavyzdžius.

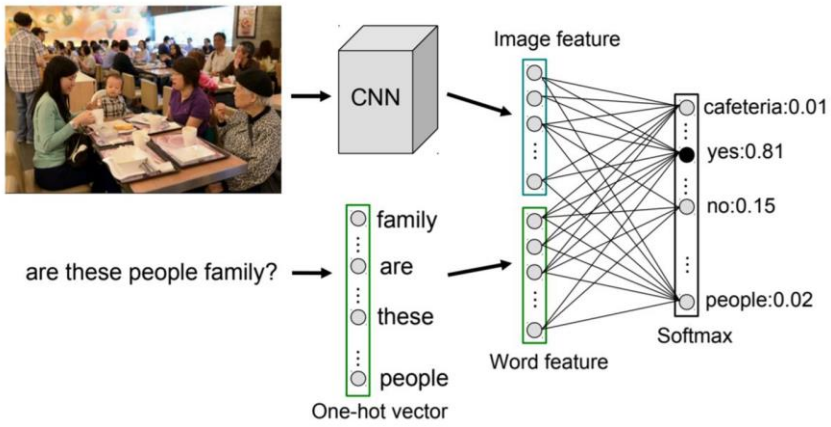


3 pav. [2] tyrimo vaizdo ir klausimų pirmasis pavyzdys

4 pav. [2] tyrimo vaizdo ir klausimų antrasis pavyzdys

Autoriai klausimų požymių vektoriams gauti naudojo neuroninį tinklą, kuris susidėjo iš 2 paslėptų sluoksnių po 1000 neuronų ir tanh aktyvacijos funkciją bei ilgos-trumpos atminties (angl. *long-short term memory*) modelį su softmax išvesties funkcija. Tuo tarpu, vaizdų požymiams apskaičiuoti buvo naudojamas VGG tinklo paskutinis sluoksnis, kuris buvo sumažinamas iki 1024 dydžio vektoriaus tam, kad atitiktų LSTM modelio išvestį. Pažymime, jog vaizdo ir teksto požymiai buvo sujungiami naudojantis sandaugos operacija. Geriausi rezultatai buvo gaunami su LSTM tinklu: buvo pasiektas 54.06 % tikslumas. Taip pat, autoriai darbe atliko eksperimentą ir apskaičiavo, jog geriausias modelis pasižymėjo tokiu pat klausimų bei vaizdo supratimu, koks būdingas 4,45 metų vaikui. Iš šio palyginimo galima matyti, jog šios tematikos modelius ateityje galima tobulinti.

Kaip bebūtų, [3] autoriai teigia, jog robastiškų modelių , gebančių spręsti darbe nagrinėjamą tematiką tikslumais, artimais žmonių lygiui, kūrimas būtų didelis žingsnis bendrinio dirbtinio intelekto link ir pateikia architektūrą, kuri remiasi žodžių rinkiniais (angl. *Bag-of-words*) ir skaitmeninių vaizdų požymiais. Autoriai pažymi, jog ši sąlyginai paprasta architektūra sugeba pasiekti didesnius tikslumus nei LSTM sluoksniais grįsti tinklai sintezuotai COCO duomenų imčiai. Kaip bebūtų, autorių pateikta architektūra nėra tokia tiksli, kuomet yra naudojama papildyta COCO VQA imtis. Šio straipsnio tyrėjai vaizdinių požymių žemėlapiams išgauti pasirinko GoogLeNet sąsūkų neuroninio tinklo architektūrą, o kaip ir minėjome jau anksčiau, teksto požymiams sugeneruoti tyrėjai naudoja naivų žodžių rinkinių požiūrį. 5 paveikslėlyje pateikiame šio tinklo architektūrą. Taip pat galima pridurti, kad žodžių ir vaizdo požymiai sujungiami naudojantis ne daugybos ar sudėties operacijomis, o pasitelkiantkonkatenacijos operaciją. [3]



5 pav. „iBOWIMG“ modelio architektūra [3]

Tuo tarpu 1 lentelėje pateikiame [3] tyrimo rezultatus, kurie, kaip ir minėjome, buvo atlikti su COCO duomenų imtimi, kurioje kiekvienas vaizdas turėjo po 3 klausimus bei 10 atsakymų.

1 lentelė. [3] tyrimo rezultatai procentais

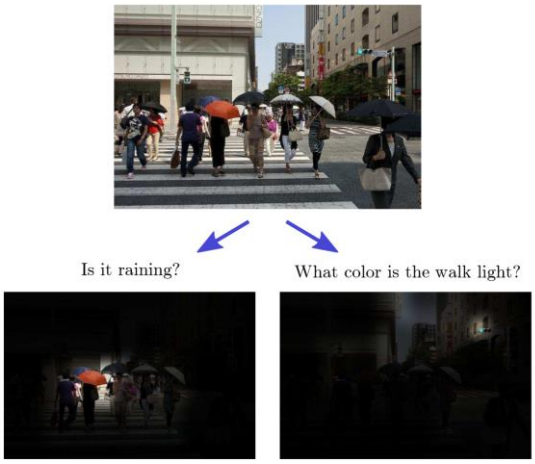
Architektūra	**Atviras atsakymas**				**Baigtinis atsakymas**			
	Bendras	**Taip/Ne**	**Skaičius**	**Kiti**	**Bendras**	**Taip/Ne**	**Skaičius**	**Kiti**
LSTMIMG	54.06	-	-	-	-	-	-	-
NMN+LSTM	55.10	-	-	-	-	-	-	-
ACK	55.98	79.05	36.10	40.61	-	-	-	-
DPPnet	57.36	80.28	36.92	42.24	62.69	80.35	38.79	52.79
iBOWIMG	**55.89**	**76.76**	**34.98**	**46.62**	**61.97**	**76.86**	**37.30**	**54.60**

Kaip galima matyti iš 1 lentelės, „iBOWIMG“ modelis ACK tinklui nusileidžia tik 0.1 %, o DPPnet tinklui nusileidžia 1.47 %, tuo tarpu likusius tinklus atviruose atsakymuose lenkia. Taip pat autorių tinklas nusileidžia DPPnet tinklui per 0.72 % baigtiniuose atsakymuose. Įvertinus, jog šis tinklas naudoja gana paprastą klausimų požymių apskaičiavimą, galima teigti, kad rezultatai yra pakankamai geri. [3]

1.1.2. Požymių sujungimas panaudojant specifinius vaizdo, požymių regionus

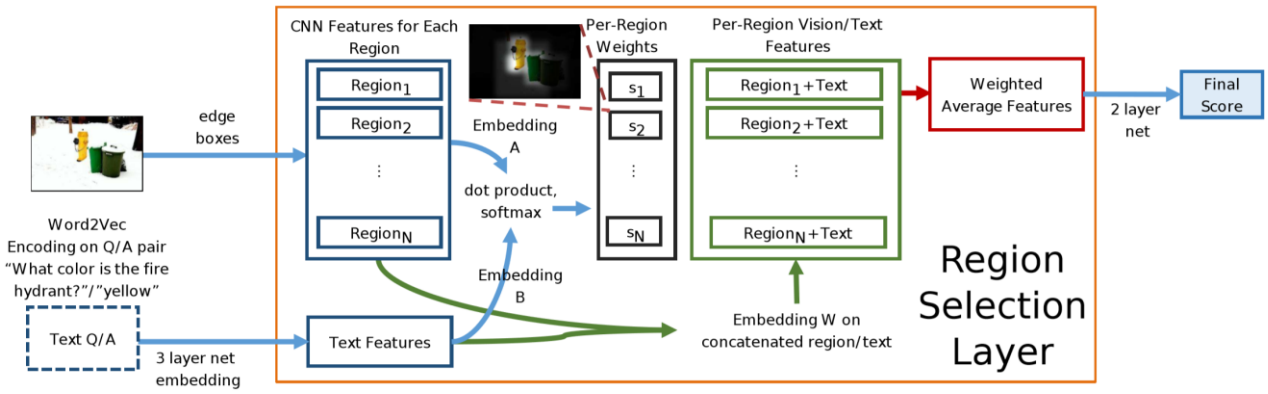
Sprendžiant darbe nagrinėjamą problemą, galima intuityviai suprasti, jog ne visi skaitmeninio vaizdo ar teksto regionai yra reikalingi, norint pateikti teisingą atsakymą į užduotą klausimą. Pavyzdžiui, pateikus 1 paveikslėlį ir uždavus klausimą „kiek bananų yra nuotraukoje“ žmogus koncentruos savo dėmesį į tą vaizdo sritį, kurioje mato bananus. Šiame skyrelyje apžvelgsime atliktus tyrimus, kurie taip pat bando tam tikriems vaizdo ar teksto bei požymių regionams suteikti didesnį svorį nei kitiems.

[4] tyrime mokslininkai pateikia modelį, kuris geba išsirinkti vaizdo regionus, aktualius užduotam klausimui atsakyti. Autoriai teigia, kad šiai sistemai būdingi didesni tikslumai, kuomet yra atsakinėjama į tokius klausimus kaip „kokia spalva?“, „koks kambarys?“ ir t. t. Vaizdo regionų išrinkimo pavyzdį pateikiame 6 paveikslėlyje tam, kad būtų lengviau suprasti modelio idėją.



6 pav. Aktualių vaizdo regionų išrinkimo pavyzdys [4]

[4] tyrime taikytas modelis mokosi atvaizduoti tekstinę informaciją ir vizualinius regionus latentinėje erdvėje, kurioje skaliarinė vektorių sandauga atitinka kiekvieno regiono svarbą. (žr. 7 pav.)



7 pav. [4] tyrimo modelio architektūra

Įvestis yra klausimas, potencialus atsakymas ir skaitmeninio vaizdo požymiai iš automatiškai atrinktų regionų kandidatų rinkinio. Klausimas ir atsakymas yra atvaizduojami vektoriais naudojantis word2vec algoritmu ir trijų sluoksniu tinklu. Vizualiniai požymiai kiekvienam regionui yra išgaunami naudojantis paskutiniais trimis sąsūkų neuroninio tinklo, kuris buvo apmokytas naudojantis ImageNet imtimi, sluoksniais. Tuomet kalbos ir vaizdo požymiai yra atvaizduojami ir palyginami naudojantis skaliarine vektorių sandauga, softmax sluoksniu tam, kad būtų gauti regionų svarbos svoriai. Naudojantis šiais sluoksniais, sukonketinuotų vaizdo ir teksto požymių svertinis vidurkis tampa įvestimi dviejų sluoksnių pilnai sujungtam tinklui, kurio išvestis yra atsakymo kandidatas. [4]

Autoriai tinklui įvertinti naudoja MS COCO VQA duomenų rinkinį, kuris susideda iš 82783 vaizdų treniravimui, 40504 validavimui ir 81434 vaizdų testavimui. Kiekvienas vaizdas turi 3 klausimus ir 10 atsakymų. Rezultatai yra įvertinami naudojantis baigtiniais klausimų atsakymais. 8 paveikslėlyje pateikiame pavyzdį, kaip atrodo klausimų, vaizdų poros naudojantis šiuo algoritmu, o 2 lentelėje gautus tikslumus. [4]

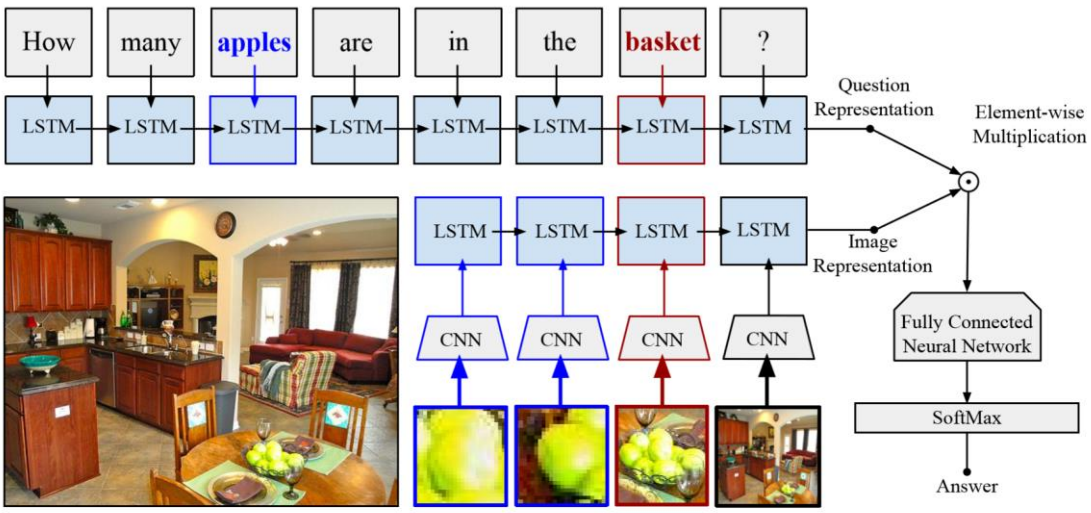
8 pav. [4] modelio dėmesio regionai

2 lentelė. [4] tyrimo rezultatai procentais

Architektūra	**Baigtinis atsakymas**			
	Bendras	**Taip/Ne**	**Skaičius**	**Kiti**
iBOWIMG	61.97	76.86	37.30	54.60
Word+Region Sel.	**62.43**	**77.18**	**33.52**	**56.09**

Iš 8 paveikslo galima matyti, kuriuos modelis regionus vaizde užfiksuoja, kaip svarbiausius, teisingo atsakymo pasirinkimui, o iš 2 lentelės galima įsitikinti, jog autorių pateikta idėja turi pranašumų prieš anksčiau nagrinėtą iBOWIMG modelį, kadangi bendras tikslumas yra didesnis per 0,46 %. [4]

Taip pat [5] straipsnio autoriai bandė spręsti darbe nagrinėjamą tematiką taikant idėją, jog ne visi vaizdo ar teksto požymiai yra reikalingi norint pateikti teisingą atsakymą. [5] straipsny autoriai pristatė sutelkto dinaminio dėmesio (angl. Focused dynamic attention) modelį, kuris naudodamas objektų aptikimą sugeba atskirti svarbius regionus ir jų požymius apjungti su globaliais požymiais pasitelkiant LSTM sluoksnį. (žr. 9 pav.)



9 pav. Sutelkto dinaminio dėmesio modelis [5]

Iš 9 paveikslėlio galima matyti, jog autoriai teksto požymių apskaičiavimui naudoja LSTM tinklą, kadangi, šis tinklas pasižymi savybe atsiminti svarbius klausimo žodžius ir sugeba elgtis kaip dėmesio sutelkimo mechanizmas pateiktiems klausimams. Tuo tarpu vaizdo analizei autoriai pasirinko gilius sąsūkų neuroninius tinklus, kurie laimėjo COCO 2015 varžybas. Tyrėjai pasirenka paskutinius sluoksnius prieš softmax sluoksnį ir jų reikšmes laiko vaizdo požymiais. Pažymime, jog modelis

apskaičiuoja ne tik viso vaizdo požymius (globalūs vaizdiniai požymiai), bet ir požymius tik iš specifinių vaizdo regionų (lokalūs vaizdiniai požymiai). Tuomet, kitaip nei kiti mokslininkai, autoriai pasirenka LSTM tinklą, kuris yra atsakingas už lokalių ir globalių požymių apjungimą. Toliau straipsnio autoriai pristato tinklo mechanizmą: kiekvienam vaizdo objektui yra naudojami word2vec žodžių vektoriai tam, kad apskaičiuoti panašumą tarp klausimo žodžių ir objektų klasių. Po to yra pasirenkami tik tie objektai, kurie turi aukštesnį nei 0.5 panašumą ir yra apskaičiuojami tų objektų požymiai naudojantis ResNet modeliu. Tuomet, laikantis žodžių tvarkos, yra perduodami lokalūs požymiai LSTM tinklui. Pabaigoje LSTM tinklui dar yra perduodami globalūs požymiai ir gaunamas rezultatas yra laikomas vaizdo reprezentacija. Taigi, dėmesio mechanizmas įgalina modelį kombinuoti lokalius bei globalius požymius į vieną atvaizdavimą, kuris yra reikalingas spręsti kompleksines darbo problematikos užduotis. Pažymime, jog, treniravimo metų objektų klasės buvo ne prognozuojamos, bet iš anksto sužymėtos, o testavimo metu, tyrėjai naudojosi [6], [7] straipsnių idėjomis, kurių šiame darbe plačiau nenagrinėsime. Visas modelio darbas pasibaigia, kuomet yra apjungiami teksto bei vaizdo požymiai pasinaudojant Tanh ir $ReLU^2$ aktyvacijos funkcijomis bei pilnai sujungtu neuroniniu tinklu, prognozuojamam atsakymui pasirinkti. [5]

Kaip ir prieš tai nagrinėtų straipsnių autoriai, taip ir [5] autoriai pasirinko COCO duomenų rinkinį, tinklo apmokymui, validavimui bei testavimui. 3 lentėje pateikiame gautus rezultatus (test-dev imčiai). [5]

3 lentelė. [5] tyrimo rezultatai procentais

Modelis	**Bendras**	**Taip/Ne**	**Skaičius**	**Kiti**
DPPnet	57.22	80.71	37.24	41.71
D-NMN	57.90	80.50	37.40	43.10
SAN	58.70	79.30	36.6	46.10
FDA	**59.24**	**81.14**	**36.16**	**45.77**

Iš aukščiau pateiktų tikslumų, galima matyti, jog modelis pasižymėjo gan gerais rezultatais ir sugebėjo aplenkti daugelį tinklų, kurie nagrinėja tą pačią tematiką.

[8] straipsnio autoriai naudojasi *iš apačios viršun* (angl. bottom up) ir *iš viršaus apačion (angl. top down)* idėjomis. Modelio įvestimi tyrėjai pasirinko klausimą ir vaizdą: klausimų 300-dimensijų pradiniams vektoriams pasinaudojo GloVe apmokytomis reikšmėmis (treniravimo metu, buvo leidžiama tinklui mokytis šiuos vektorius). Šios teksto įvestys buvo toliau siunčiamos į GRU tinklą, kurio išvestis buvo 512-dimensijų vektorius. Tuo tarpu, vaizdas buvo siunčiamas į sąsūkų neuroninį tinklą, kurio išvestis buvo *K x 2048* duomenų masyvas, kur K – vaizdo specifinių regionų skaičius. Jog gauti šiuos vektorius, autoriai naudojo *iš apačios viršun* požiūrį: šis metodas remiasi ResNet sąsūkų tinklu bei Faster R-CNN architektūra, kuri yra apmokyta aptikti specifinius vaizdo regionus (elementus, objektus) naudojantis Visual Genome rinkinio anotacijomis. Tyrėjai sąsūkų tinklui mokymo metu (kuomet buvo mokoma VQA modelis) neleido mokytis ir užfiksavo pastovius svorius. Kitaip tariant, iš *apačios viršun* požiūrį galima laikyti, kaip vaizdo paruošimą VQA modeliui. Po šių

žingsnių prasideda dėmesio sutelkimas vaizdo regionams. Autoriai pateikia *klasikinį iš viršaus apačion* požiūrį:

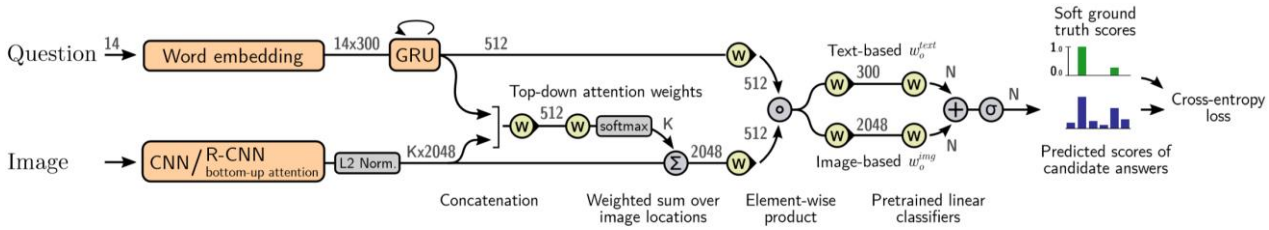
1. Kiekvieno vaizdo regiono $i = 1 .. K$ požymių vektorius yra sukonkatinuojamas su klausimo požymių vektoriumi.

2. Tuomet abiem vektoriams yra pritaikoma softmax funkcija ir tiesinis sluoksnis tam, kad gauti skaliarinius dėmesio svorius.

3. Skaliariniai dėmesio svoriai (per visus regionus) yra normalizuojami pritaikant softmax funkciją.

4. Vaizdo požymiai yra dauginami iš 3 žingsnyje gautų normalizuotų svorių ir sudedami tam, kad gauti 2048 ilgio vektorių, kuris representuoja *dėmesio vaizdą*.

VQA modelio architektūros pabaigoje yra pritaikoma klausimo požymių vektoriui ir 4 žingsnyje gautam vektoriui Hadamard sandauga (kiekvienas pirmojo vektoriaus elementas yra dauginamas su antrojo vektoriaus pirmu elementu ir t.t.). Tuomet gautas vektorius tampa bendriniu (angl. joint) vektoriumi ir yra perduodamas klasifikatoriui. Pažymime, jog šiame darbe autoriai paskutinį klasifikatorių pasirinko logistinę regresiją, kuri parodo kiekvieno atsakymo kandidato tikimybė būti teisingu atsakymu. Tam, jog būtų lengviau suprasti architektūrą, pateikiame 10 paveikslėlį. [8]



10 pav. [8] tyrime taikyta architektūra

Modelio įvertinimui autoriai pasirinko VQA v2 validavimo imtį. Gauti rezultatai, naudojant *iš apačios viršun* idėją parodė, jog bendrai klausimai buvo atsakinėjami 62.82 % tikslumu, Taip/Ne klausimai buvo atsakinėjami 79.92 % tikslumu, skaičių klausimai atsakinėjami 42.44 % tikslumu, o kiti klausimai atsakinėjami 55.35 % tikslumu. Kadangi buvo kita naudota imtis, nei prieš tai apžvelgtuose tyrimuose, tai negalim palyginti šios architektūros su kitomis, bet, kaip bebūtų, iš gautų rezultatų, galim įsitikinti, jog šis požiūris yra gana sėkmingas ir gan plačiai taikomas. [8]

Taigi, šiame skyrelyje apžvelgėme taikytus tyrimus bei gautus rezultatus kasdieniams vaizdams. Ne visus modelius pavyko tiesiogiai palygint dėl skirtingų duomenų rinkinių, bet tikime, jog pateikta informacija padėjo geriau įsisavinti taikomas idėjas darbo nagrinėjamoje tematikoje. Kitame skyrelyje apžvelgsime su medicina susijusius eksperimentus, jų specifikas bei gautus rezultatus.

1.2. Atlikti tyrimai su medicinos srities duomenimis

Kaip ir praeitame, taip ir šiame skyrelyje nagrinėsime atliktus tyrimus darbo nagrinėjamoje tematikoje, tik, šiuo atveju, apžvelgsime straipsnius, kurie koncentruojasi į medicininius duomenų rinkinius.

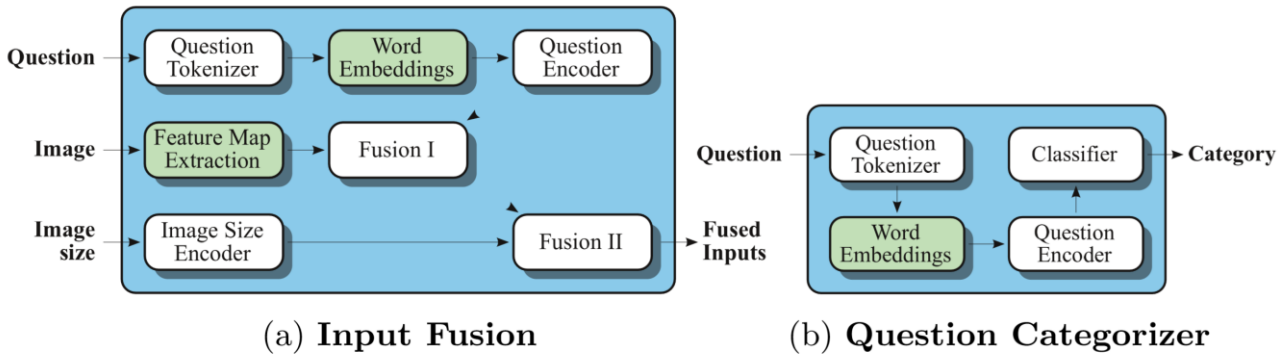
1.2.1. Keleto klasifikatorių išnaudojimas

[9] straipsnio autoriai naudojo VQA-Med 2019 duomenų rinkinį klausimais grįstoje vaizdo analitikoje. Šis rinkinys susideda iš 3200 treniravimo vaizdų bei 12792 klausimų-atsakymų porų. Validavimo imtis turi 500 vaizdų ir 2000 klausimų-atsakymų porų, o testavimo imtyje yra 500 vaizdų ir 500 klausimų. Pagrindinius klausimų tipus galima suskirstyti į šiuos: [9]

- nustatyti vaizdo modalumą (C1 kategorija);
- nustatyti vaizdo plokštumą (C2 kategorija);
- identifikuoti organus/anatomiją tam tikroje vaizdo dalyje (C3 kategorija);
- identifikuoti vaizdo anomaliją (C4 kategorija).

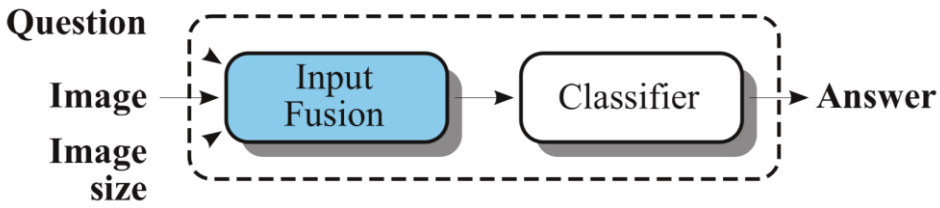
Žemiau pateikiame bandytas tyrėjų idėjas, o 11 paveikslėlyje taikytą modelį:

- įvairūs klausimų atvaizdavimo į vektorius metodai: nuo žodžių rinkinių (angl. Bag-of-words) iki įvairių rekurentinių tinklų variacijų;
- įvairūs vaizdų enkoderiai: nuo paprastų neuroninių tinklų iki jau apmokytų sąsūkų neuroninių tinklų, remiantis ImageNet imtimi;
- įvairūs duomenų apjungimo metodai.



11 pav. [9] tyrime naudotos modelių architektūros

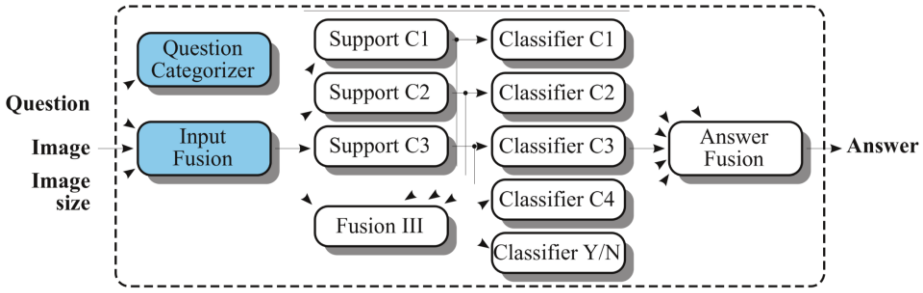
Autoriai pateiktoje architektūroje naudojo GloVe klausimų atvaizdavimus į vektorius, kurie buvo naudojami kartu su ilgos-trumpos atminties tinklu. Šio tinklo išvestis kartu su vaizdo požymių žemėlapiais, kurie buvo apskaičiuoti remiantis VGG tinklo architektūra buvo perduodami į pirmąjį suliejimo (angl. Fusion) modulį. Toliau gauti rezultatai yra perduodami į antrąjį suliejimo modulį. Pažymime, jog žali elementai modelio architektūroje reiškia, kad tose architektūrose srityse buvo remiamasi iš anksto apmokytais svoriais. Taip pat autoriai įvesties išmaišymo moduliui pasirinko pilnai sujungtąjį neuroninį tinklą (žr. 12 pav.) [9]



12 pav. [9] tyrime naudotas įvesties modulis

Pažymime, jog šis modulis iš pradžių buvo apmokytas remiantis C1, C2, C3 klausimų kategorijomis, siekiant išgauti modelį, kuris sugebėtų tinkamai atsakinėtį į sudėtingesnius klausimus. Toliau autoriai rėmėsi 5 atskirais klasifikatoriais, kuriems įvestis buvo įvesties suliejimo modulio rezultatas. Kaip autoriai teigia, visi šie klasifikatoriai buvo specializuoti skirtingoms klausimų kategorijoms ir visi turėjo savo atsakymus bei klaidos funkcijos. Visų šių klasifikatorių prognozės buvo perduodamos atsakymų suliejimo moduliui, kuris pasirinkdavo atsakymus iš tinkamų klasifikatorių, remiantis klausimų kategorizavimo moduliu (žr. 11 pav. (b)) [9]

Žemiau esančiame 13 paveikslėlyje pateikiame galutinę modulio architektūrą.



13 pav. Galutinė [9] tyrimo modelio architektūra

Autoriai padalino klasifikavimo modulius į du tinklus: palaikančiuosius tinklus (susideda iš 2 pilnai sujungtų sluoksnių) ir galutinius klasifikatorius (susideda iš 1 pilnai sujungto sluoksnio). Taip pat, kaip palaikantysis, klasifikatorius buvo C2 tipo klausimų modulis. Autoriai pažymi, jog palaikančiųjų tinklų rezultatai yra sukonkatenuojami su įvesties suliejimo rezultatais ir perduodami kaip įvesties vektoriai tolimesniems klasifikatoriams. [9]

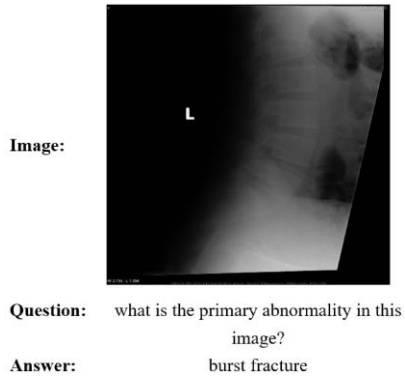
4 lentelė. [9] tyrimo rezultatai

		Apmokymo imtis			Validavimo imtis		
		Tikslumas (angl. Precision)	Atsakas (angl. Recall)	F1	Tikslumas (angl. Precision)	Atsakas (angl. Recall)	F1
IF-1C		0.683	0.497	0.545	0.690	0.499	0.548
SFN		0.753	0.692	0.707	0.762	0.704	0.717

Iš 4 lentelės galima matyti, jog SFN modelis lenkia visose rezultatų metrikose IF-1C modelį. Pažymime, jog SFN modelis yra aprašytas aukščiau, o IF-1C yra taip pat [9] autorių sukurtas modelis, bet jis naudojasi tik vienu klasifikatoriumi. [9]

1.2.2. Požymių apjungimas išnaudojant specifinius vaizdo, požymių regionus

[10] straipsno autoriai taip pat naudojosi VQA-Med 2019 imtimi tyrimams atlikti. 14 paveikslėlyje pateikiame šios imties pavyzdį.



14 pav. VQA-Med 2019 imties pavyzdys [10]

[10] straipsnio autorių naudojamas modelis susideda iš šių modulių: vaizdo požymių apskaičiavimo, klausimų semantikos enkoderio, požymių suliejimo mechanizmo, atsakymo prognozavimo (žr. 15 pav.)

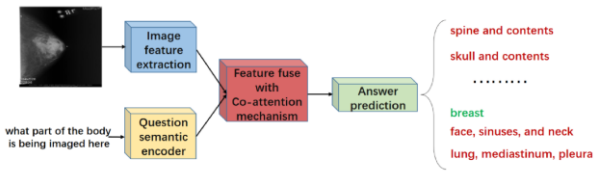
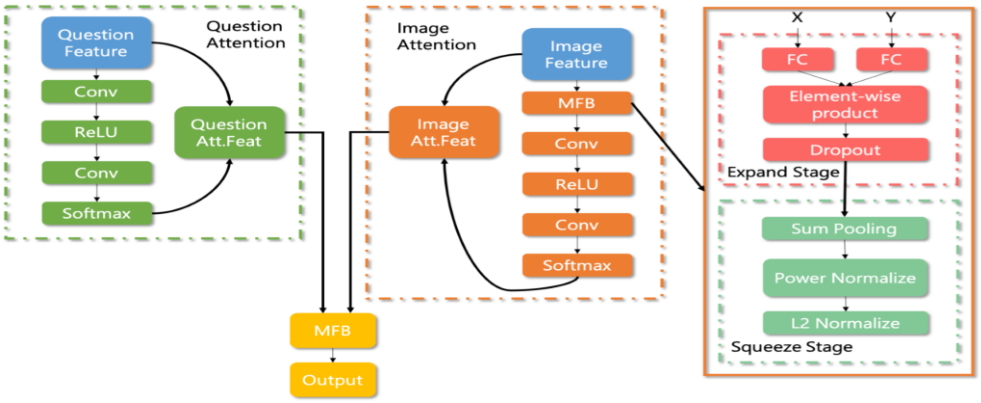


Fig. 2. Our model architecture

15 pav. [10] modelio veikimo schema

Vaizdo požymių išskyrimui autoriai pasirinko modelį, kuris remiasi VGG16 tinklu (apmokytu naudojantis ImageNet imtimi) ir globalaus vidurkinimo strategija. Straipsnyje pažymima, jog šis tinklas apskaičiuoja 1984 dimensijos vektorių, kuris reprezentuoja vaizdo požymius. Tuo tarpu klausimų požymiams išgauti, tyrėjai pasirinko dvikryptį enkoderį BERT. Žodžių atvaizdavimui buvo pasirinkti GloVe vektoriai. Autoriai pažymi, jog naudojamas BERT modelis, kuris turi 12 sluoksnių, 768 pasleptų kintamųjų ir iš viso 110 milijonų parametrų. Tam, kad gautų teksto požymius, tyrėjai suvidurkina paskutinį sluoksnį į 768 dimensijos vektorių. Tyrėjai pabrežia, jog požymių suliejimas vaidina labai svarbią rolę galutinių rezultatų tikslume ir naudoja bendrinio dėmesio mechanizmą, kuris priskiria svarbumą vaizdo regionams tam, kad išvengti nesvarbios informacijos. Autoriai naudoja multi-modalinį faktorizuotą telkimą (MFB) (žr. 16 pav.)



16 pav. [10] tyrime naudojamas modelis

Modeliui apmokyti autoriai naudoja Adam optimizavimo metodą, su mokymo žingsniu lygiu 0,0001. Reguliarizavimo koeficientas yra lygus 0,001. Modeliai mokėsi su rinkinio (angl. batch) dydžiu lygiu 32 ir naudojo 300 epochų.

Tinklui įvertinti tyrėjai naudojo tikslumo bei BLEU metrikas. Pirmoji metrika parodo santykį tarp teisingai atsakytų klausimų skaičiaus ir visų klausimo skaičiaus. Antroji metrika matuoja panašumą tarp prognozuojamo atsakymo ir teisingo atsakymo. Geriausias modelis pasiekė 0,624 tikslumo metriką, o BLEU įvertis buvo lygus 0,644. Kaip galima matyti iš rezultatų, tikslumas yra gana aukštas, žinant uždavinio sudėtingumą, bet anksčiau 1.2.1. skyrelyje pateikto modelio rezultatai buvo geresni. [10]

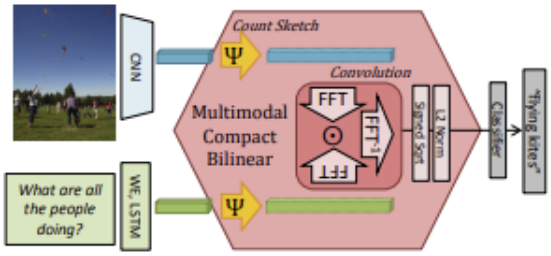
[11] tyrėjai taip pat rėmėsi idėja, jog ne visi vaizdo regionai yra vienodai svarbūs. Tyrimų duomenų rinkiniu, autoriai pasirinko VQA-MED 2018 imtį, kuri susidėjo iš: treniravimo imties su 5413 klausimų-atsakymų porų bei 2278 vaizdų, validavimo imties su 500 klausimų-atsakymų porų bei 324 skaitmeninių vaizdų, testavimo imties su 500 klausimų bei 264 vaizdų. Pagrindiniai klausimai buvo: [11]

- lokacijos. Pavyzdžiui: „kurioje vietoje sužeidimas?“;
- nustatymo. Pavyzdžiui: „ką rodo šis magnetinis rezonansas?“;
- taip/me. Pavyzdžiui: „ar smegenų magnetinis rezonansas normalus?“;
- kitų klausimų. Pavyzdžiui: „kokio tipo šis vaizdas?“.

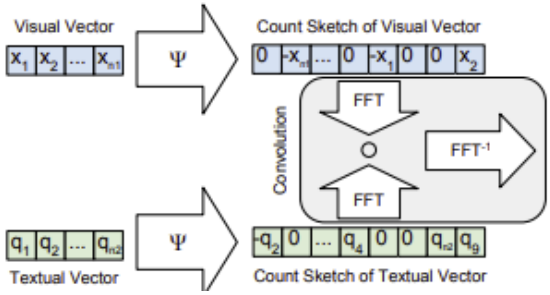
Šiame tyrime yra pristatomas sublokuotas dėmesio tinklas, kuris susidėjo iš trijų pagrindinių komponentų: sąsūkų neuroninio tinklo, kuris yra atsakingas už skaitmeninio vaizdo požymių regionų apskaičiavimą, LSTM neuroninio tinklo, kuris atvaizduoja klausimus į semantinius vektorius, išlaikant teksto kontekstą, sublokuoto dėmėsio modelio, kurio paskirtis yra atpažinti vaizdo regionus, kurie yra susiję su teisingo atsakymo prognoze. Vaizdo požymių išskirimui autoriai naudoja VGG-16 architektūrą, kuri yra iš anksto apmokyta su ImageNet duomenų imtimi. Tuo tarpu, klausimų vektoriais autoriai laiko paskutinio LSTM sluoksnio požymius. Autoriai pažymi, jog vaizdo ir klausimų požymiai buvo naudojami generuoti dėmesio pasiskirstymui vaizdo regionuose. Toliau pirmasis sublokuoto dėmesio modelis apskaičiuoja koreliaciją tarp žodžių bei vaizdo regionų, o multi-modalinis sutelkimas yra atliekamas tam, kad sugeneruoti klausimo ir teksto vektorių suliejimą, kuris yra perduodamas kitam architektūros sluoksniui. Autoriai išskiria, jog naudojo du dėmėsio sutelkimo

sluoksnius, kadangi daugelį tyrimų, šis pasirinkimas suteikdavo geriausius rezultatus. Paskutinis modelio žingsnis yra atsakymo prognozavimas. Pažymime, jog autoriai prognozavo atsakymus iš baigtinio kiekio klasių ir šiai užduočiai naudojo pilnai sujungto vieno sluoksnio neuroninį tinklą, kurio išvesčiai buvo pritaikoma Softmax funkcija. [11]

Multi-modalinis kompaktinis dvinaris sutelkimas sulaukia didelio tyrėjų susidomėjimo, kadangi ši architektūrą laimėjo CVPR-2016 klausimais grįstos analitikos mokymų iššūkį. Ši idėja pritaiko standartinę matricų sandaugą teksto ir klausimų požymiams. Pagrindiniai architektūros komponentai yra: sąsūkų neuroninis tinklas vaizdo požymiams, LSTM modelis tekstui ir multi-modalinis sutelkimas, kuris pirmiausia prognozuoja erdvinį dėmėsį (angl. spatial attention) ir tuomet apjungia dėmėsio reprezentaciją su teksto reprezentacija, tam, kad atlikti prognozes. (žr. 17 ir 18 pav.) [11]



17 pav. Multi-modalinio sutelkimo vaizdinė schema [12]



18 pav. Multi-modalinio sutelkimo schema [12]

Tyrėjai vaizdo modeliams pasirinko ResNet-152 ir ResNet-50 modelius, apmokytus su ImageNet imtimi. Klausimų modeliui pasirinko tinklą iš 2 LSTM sluoksnių su 1024 neuronų kiekviename sluoksnyje. Abiejų sluoksnių sukonkatinuotas vektorius tampa įvestimi kitam sutelkimo sluoksniui. Tuomet multi-modalinis sutelkimas yra naudojamas sulieti teksto ir vaizdo vektorius. Tam, kad išlaikyti dėmesio sutelkimą, multi-modalinis sluoksnis yra dar kartą panaudojamas, norint sujungti prieš tai gautą vektorių su kiekvienos erdvinės gardelės lokacija. [11]

Tyrėjai pažymi, jog sąsūkų neuroniniai tinklai sugeba išgauti gerus rezultatus, kai yra daugybė paruoštų duomenų. Kaip bebūtų, medicinos srityje, vaizdų kiekis yra stipriai ribotas ir yra būtina naudoti jau iš anksto apmokytus modelius su kitomis imtimis (angl. fine tuning). Autoriai išskiria, jog šie modeliai dažnai yra mokomi su ImageNet imtimi ir išmoksta bazinius požymius, kurie paskui gali būti sėkmingai taikomi ir kitoms imtims. Norime pabrėžti, jog šie tyrėjai, kitaip nei kiti, naudojo ne ImageNet svorius, bet iš pradžių apmokė tinklus su įvairiais mediciniais vaizdais. [11]

Modelių kokybei įvertinti, mokslininkai pasirinko WBSS metriką, kuri, panašiai, kaip ir BLEU įvertina prognozuojamo atsakymo panašumą į tikrąjį atsakymą. 5 lentelėje pateikiame gautus rezultatus.

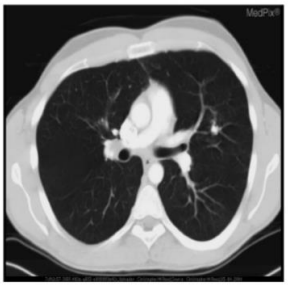
5 lentelė. [11] tyrimo rezultatai

Modelis	WBSS įvertis	BLEU įvertis
Run1-SAN	0,174	0,121
Run2-SAN	0,168	0,108
Run3-SAN	0,157	0,106
Run4-MCB	0,130	0,083
Run5-MCB	0,144	0,085

Iš 5 lentelės matome, jog geriausias modelis pasiekė 0,174 WBSS įvertį ir 0,121 BLEU įvertį. Pažymime, jog, nors iš pirmo žvilgsnio, gali pasirodyt, jog rezultatai prasti, visame ImageCLEF 2018 konkurse geriausio modelio WBSS įvertis buvo 0,186, o BLEU 0,158. Iš šių rezultatų, galima teigti, jog modelio idėjos yra pakankamai geros, o su anksčiau aprašytais tyrimais palyginimą būtų daryti neteisinga, nes modeliai buvo mokinami ir testuojami su skirtingomis imtimis. [11]

1.2.3. Tyrimai remiantis variaciniais enkoderiais

[13] tyrimo autoriai atliko eksperimentus su VQA-Med 2020 imtimi, kurioje treniravimo dalis susideda iš 4000 radiologinių vaizdų su 4000 klausimų-atsakymų poromis, testavimo dalis susideda iš 500 vaizdų ir 500 klausimų-atsakymų porų, o validavimo imtis taip pat susideda iš 500 klausimų-atsakymų porų ir 500 radiologinių vaizdų. 19 paveikslėlyje pateikiame šios imties pavyzdžius. [13]

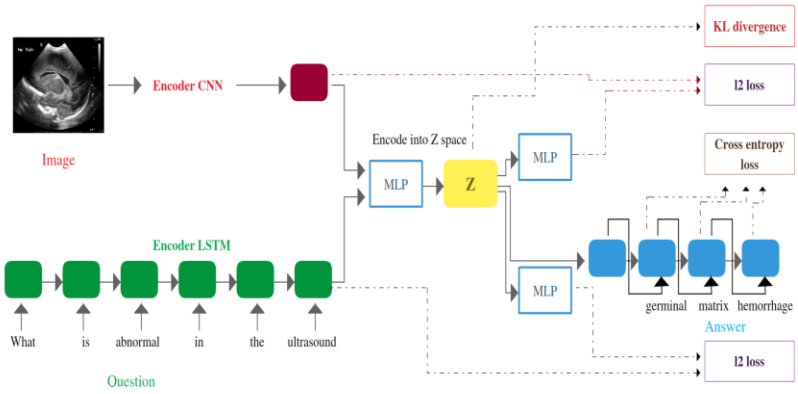


Question: what abnormality is seen in the image?
Answer: pulmonary arteriovenous malformation

Question: what is the primary abnormality in this image?
Answer: hemophilia

19 pav. VQA-Med 2020 imties pavyzdžiai [13]

[13] tyrimo autoriai pristato architektūrą, kuri remiasi variacinių autoenkoderių idėja, kuriai įvedus vaizdą ir klausimą, sistema išprognozuoja atsakymą. Modelis susideda iš dviejų neuroninių tinklų modulių: enkoderio, dekoderio. Pirmiausia, enkoderis sukuria latentinės erdvės kintamąjį z iš vaizdo ir klausimo bei apskaičiuoja tankio vektorius h_v ir h_q latentinėje z erdvėje. Sąsūkų neuroninis tinklas yra naudojamas gauti vaizdo požymių žemėlapiui v, o LSTM tinklas yra naudojamas klausimo vektoriams q apskaičiuoti. Tuomet modelis rekonstruoja įvesties požymius iš latentinės erdvės naudojantis pilnai sujungtu neuroniniu tinklu. Galiausiai LSTM dekoderis generuoja atsakymą iš z erdvės. Kitaip tariant, dekoderis gauna, kaip įvestį, z erdvės vektorių ir jį naudoja atsakymui sugeneruoti. 20 paveikslėlyje pateikiame modelio schemą. [13]



20 pav. [13] tyrime naudotas modelis

Autoriai mokindami modelius naudojo 224x224 vaizdų formatą, ADAM optimizatorių su mokymosi žingsniu lygiu 0,0001 ir rinkinio dydžiu 32, visi modeliai mokėsi 20 epochų. Žemiau pateikiame 4 eksperimentus, kuriuos atliko tyrėjai, o 6 lentelėje rezultatų metrikas:

1. Naudojami variaciniai autoenkoderiai. Modelis mokinamas su 2020 VQA-Med imtimi, be įvesties rekonstrukcijos. Išvesties ilgis lygus 3.
2. Naudojami variaciniai autoenkoderiai. Modelis mokinamas su 2020 VQA-Med ir 2019 VQA-Med imtimi, be įvesties rekonstrukcijos. Išvesties ilgis lygus 4.
3. Naudojami variaciniai autoenkoderiai. Modelis mokinamas su 2020 VQA-Med ir 2019 VQA-Med imtimi, su įvesties rekonstrukcija. Išvesties ilgis lygus 4.
4. Naudojami variaciniai autoenkoderiai. Modelis mokinamas su 2020 VQA-Med ir 2019 VQA-Med imtimi, su įvesties rekonstrukcija. Išvesties ilgis lygus 10.

6 lentelė. [13] tyrimo rezultatai

Eksperimento numeris	Tikslumas	BLEU įvertis
1	0,232	0,299
2	0,256	0,323
3	0,278	0,321
4	0,286	0,335

Iš 6 lentelės galima matyti, kad geriausi rezultatai yra pasiekiami su 4 eksperimento modeliu. Negalime lyginti šio modelio su ankstesniais tyrimais dėl tos pačios priežasties, jog naudojami skirtingi duomenų rinkiniai, bet tam jog įsivaizduoti geriau, tinklo tikslumus viso ImageCLEF 2020 konkurso kontekste, pažymime, jog pirmos vietos tikslumo metrika buvo lygi 0,496, o BLEU 0,542. [13]

Taigi, šiame skyriuje pasigilinome giliau į nagrinėjamą darbe tematiką. Išskyrėmė taikomų metodų tipus, rezultatus, architektūras. Taip pat, apžvelgėmė jau atliktus tyrimus įvairiuose konkursuose bei skirtingas duomenų imtimis. Kitame skyrelyje gilinsimės į darbe naudojamas metodikas.

2. Metodika

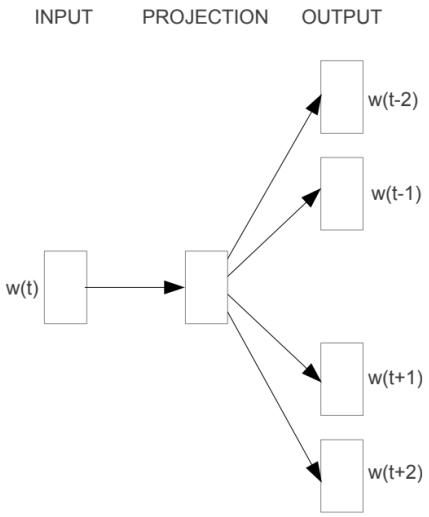
Šiame skyriuje pateiksime darbe naudojamų metodų apžvalgą. Informaciją pateiksime pagal iš anksto išskirstytus skyrelius, kurie apima vaizdo, teksto apdorojimo metodus.

2.1. Natūralios kalbos paruošimo metodai

Šiame skyrelyje apžvelgsime metodus, kurie yra taikomi natūralios kalbos apdorojime. Kadangi tekstas tiesiogiai negali būti perduodamas mašininio mokymo algoritmams, yra atlikta daugybė metodinių tyrimų, kurių tikslas yra atrasti žodžių atvaizdavimo būdus į tam tikrus vektorius. 2.1.1. skyrelyje gilinsimės į vieno iš šių metodų veikimo principą.

2.1.1. Skip n Gram metodas

Šio metodo tikslas yra atvaizduoti žodžius į latentintę erdvę taip, jog panašią semantinę prasmę turintys žodžiai būtų arčiau vieni kitų (pagal panašumo funkciją) nei kad skirtingą semantinę prasmę turintys žodžiai. [14] tyrėjų pateiktas modelis remiasi, idėja, jog panašią semantinę prasmę turintys žodžiai dažnai stovi šalia vienas kito ir siekia maksimizuoti tikimybę, jog pateikus tikslo (angl. target) žodį bus sėkmingai identifikuoti šalia esantys žodžiai, kitaip dar vadinamais konteksto žodžiais (angl. context)



21 pav. Skip n Gram modelio schema [14]

Remiantis metodo idėja ir 21 pav. galime išskirstyti modelio veikimą į šiuos žingsnius:

1. Visi žodžiai atvaizduojami į vienkartinio kodavimo vektorius (angl. one-hot encoding)
2. Perduodami vienkartinio kodavimo vektoriai projekcijos (literatūroje dar dažnai vadinama paslėptuoju) sluoksniui, kuris atvaizduoja žodį į jo atvaizdavimo *h* vektorių. Kitaip tariant, šis sluoksnis iš atvaizdavimų matricos išrenka tam tikro žodžiaus vektorių.
3. Žodžio vektorius yra sudauginamas su svorių matrica *W* ir gaunami konteksto žodžių tikimybiniai skirstiniai (pritaikius softmax funkciją).

Matematine prasme šis uždavinys siekia maksimizuoti didžiausio tikėtinumo funkciją:

$$argmax_\theta p(c_1, c_2, .., c_n | t; \theta) \quad (1)$$

Čia $c_{1..n}$ – konteksto žodžiai, t – tikslo žodis, θ – svorių matrica.

Kaip ir minėjome, šis modelis išvesties sluoksniams taiko softmax funkciją, todėl (2) išraiška galima aprašyti tikimybę, jog konteksto žodis *c* sakinyje stovi šalia tikslo žodžio *t*: [14]

$$p(c|t) = \frac{e^{W_c \cdot h}}{\sum_{i=1}^{V} e^{W_i \cdot h}} \quad (2)$$

Čia *V* – unikalių žodžių kiekis, *c* šiuo atveju simbolizuoja konteksto žodžio poziciją išvesties vektoriuje, o *i* simbolizuoja *i*-tojo žodžio poziciją išvesties vektoriuje. Pažymime, jog tam, kad supaprastinti žymėjimus šiuo atveju aprašom tik vieną konteksto, tikslo žodžių porą.

Kadangi mašininio mokymo algoritmuose yra žymiai plačiau taikomas minimizavimas, (1) išraiškai pritaikius logaritmavimą klaidos funkciją galima aprašyti (3) išraiška: [14]

$$L = -\sum_{c=1}^{C} log \frac{e^{W_c \cdot h}}{\sum_{i=1}^{V} e^{W_i \cdot h}} \quad (3)$$

Pažymime, jog (3) funkcijoje atsiranda papildomas sumavimas kontekstų žodžiams, kadangi yra prognozuojami iškart keli žodžiai. Taip pat, reikėtų įvertinti, jog daugumai modelių yra taikomas stochastinio gradiento nusileidimas ir įvesti dar papildomą sumavimą, kuris iteruoja per rinkinio (angl. batch) elementus. [14]

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} log \frac{e^{W_c \cdot h}}{\sum_{i=1}^{V} e^{W_i \cdot h}} \quad (4)$$

Čia *N* žymi rinkinio dydį.

Pažymėsime, jog šiame darbe naudosimės [15] metodu, tačiau, jis yra tik aukščiau aprašyto modelio plėtinys. [15] tyrėjai papildomai išskaido žodžius į dalis (angl. grams). Pavyzdžiui, žodis „where“ yra išskaidomas į „wh“, „whe“, „her“, „ere“, „re“, o pati žodžio reprezentacija yra šių dalių vektorių suma. Atkreipiame dėmėsį, jog tikslas yra minimizuoti klaidos funkciją *L*, o žodžių vektoriais tampa 2 žingsnyje pateiktas paslėptasis sluoksnis.

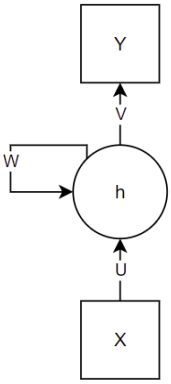
2.2. Natūralios kalbos apdorojimo metodai

Šiame skyriuje apžvelgsime metodus, kuriems perdavus sakinius (žodžių vektorius) yra apskaičiuojami teksto požymių žemėlapiai.

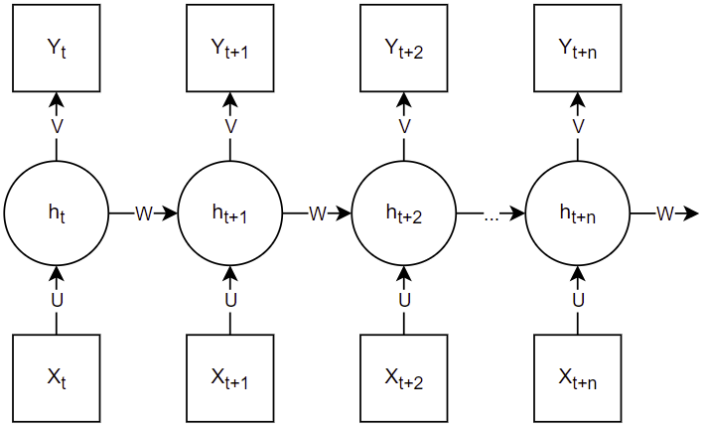
2.2.1. Rekurentiniai tinklai

Vienas didžiausių sąsūkų neuroninių tinklų trūkumų yra tai, jog šie modeliai nepasižymi gebėjimu apdoroti informacijos, kuri yra sekų pavidalo, kitaip tariant, kiekviena sekanti įvestis šiuose tinkluose nekaupia informacijos apie prieš tai buvusią įvestį. Analizuojant tekstą yra akivaizdu, jog, norint apskaičiuoti požymius, išlaikant žodžių prasmę, yra būtina tinklui suprasti ne tik laiko momentu *t*

įvedamą žodį, bet ir prieš tai buvusius. Būtent šią situaciją ir sprendžia rekurentiniai neuroniniai tinklai. (žr. 22 ir 23 pav.) [16]



22 pav. Neišskleistas rekurentinis neuroninis tinklas [19]



23 pav. Išskleistas rekurentinis neuroninis tinklas (*daug į daug*) [19]

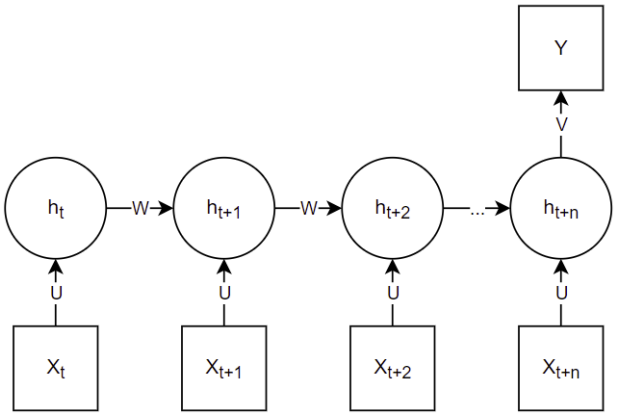
22 ir 23 paveikslėliuose galima matyti rekurentinių neuroninių tinklų schemas. Kaip ir minėjome anksčiau, pagrindinė šių tinklų idėja, jog su kiekviena įvestimi x_t yra kaupiama informacija ir apie prieš tai buvusią įvestį. Matematiškai, laiko momentu *t*, šiuos tinklus galima aprašyti 5-9 išraiškomis.

$$h_t = f(W \cdot h_{t-1} + U \cdot x_t + b_h) \quad (5)$$

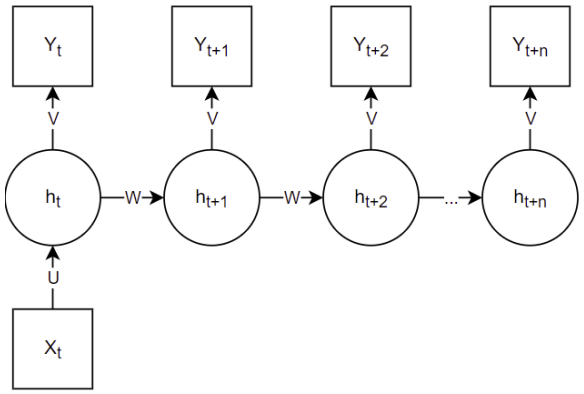
$$y_t = g(V \cdot h_t + b_y) \quad (6)$$

Čia *g*, *f* aktyvacijos funkcijos, kurios gali būti tiek ReLU, TanH, Sigmoidė bei tam tikrais atvejais Softmax, b_h ir b_y laisvieji nariai. *W* svorių matrica, kuri dauginama su paslėptuoju sluoksniu *h*, *U* svorių matrica, kuri dauginama su įvestimi, o *V* svorių matrica, kuri dauginama su paslėptuoju sluoksniu, tam, kad apskaičiuoti išvesties vektorių laiko momentu *t*. Taip pat, svarbu paminėti, jog dažniausiai paslėptas h_t sluoksnis ir išvestis y_t yra vektoriai, o 23 paveikslėlyje pateikiame tik vieną sluoksnį, t.y. praktikoje dažnai naudojami keli tokie sluoksniai, kurių įvestimis tampa prieš tai buvusių sluoksnių y_t vektoriai.

Pažymime, jog 23 paveiksle yra pateikiamas *daug į daug* tinklo tipas, o žemiau esančiuose paveikslėliuose pateikiame kitus plačiai naudojamus rekurentinių neuroninių tinklų tipus.



24 pav. *Daug į vieną* rekurentinis neuroninis tinklas [19]



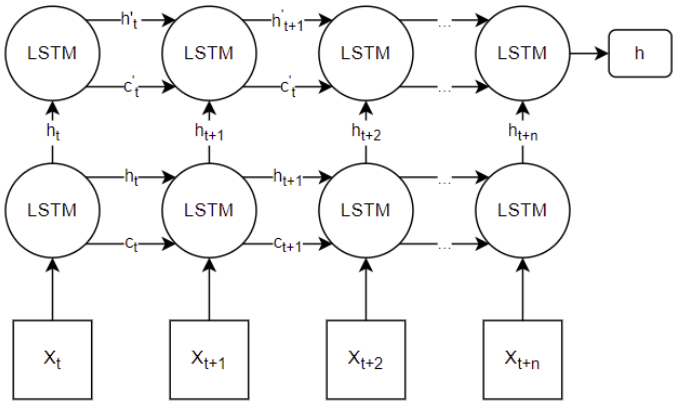
25 pav. *Vienas į daug* rekurentinis neuroninis tinklas [19]

Šiame darbe naudosime 24 paveikslėlyje pavaizduotą *daug į vieną* rekurentinį tinklą, kadangi mūsų įvestis yra sakinys ir laiko momentais *t* yra perduodami žodžių vektoriai, o išvestis yra vektorius, kuris reprezentuoja teksto požymius, šį tinklą galima aprašyti taip pat (5) ir (6) lygtimis, tik šiuo atveju išvesties vektorius *y* nebetenka laiko momento *t*:

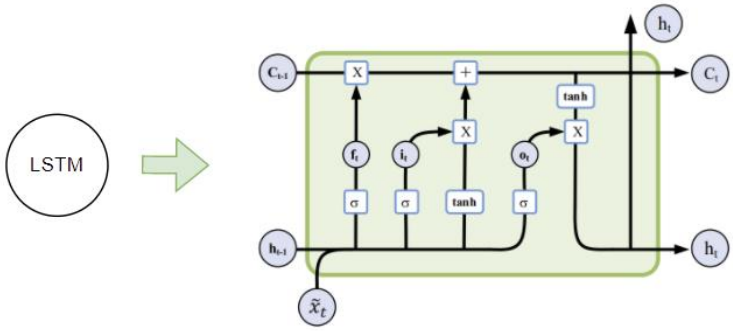
$$y = g(V \cdot h_t + b_y) \tag{7}$$

2.2.2. Ilgos-trumpos atminties tinklai

Nors rekurentiniai tinklai yra tinkamesni teksto apdorojimui nei pilnai sujungti neuroniniai tinklai ar sąsūkų neuroniniai tinklai, tačiau, jie taip pat turi trūkumų: dažnai šie tinklai susiduria su nykstančio ar sprogstančio gradiento problema, o tai lemia, jog tinklai nesugeba tinkamai mokytis. Dėl to, šie modeliai dažnai nesugeba „atsiminti“ pradinių įvesčių informacijos. Tam, kad išspręsti šią situaciją yra naudojami ilgos-trumpos atminties modeliai (angl. Long-short term memory). Pagrindinė šių tinklų idėja yra ta, jog kiekvienas šio tinklo elementas susideda iš kelių vartų (angl. Gates), kurių tikslas yra išmesti nenaudingą informaciją, įvesti bei atnaujinti naują informaciją. [17] paveikslėlyje pateikiame šio tinklo schemą, o 26 paveiksle pateikiame vieno iš šio tinklo elemento struktūrą. [17]



26 pav. Dviejų sluoksnių LSTM tinklas [19]



27 pav. LSTM sluoksnio elemento struktūra. [18]

Kaip ir minėjome anksčiau šie tinklai susideda iš 3 vartų: įvesties, atnaujinimo, išvesties. Pirmieji vartai yra atsakingi už nereikalingos informacijos iš praeito sluoksnio išmetimą, įvesties vartai yra atsakingi už reikalingos informacijos kitam sluoksniui išrinkimą, o išvesties vartai yra atsakingi už svarbiausios informacijos išvesčiai apskaičiavimą. [17]

Matematiškai šio tinklo elementą galima išreikšti (7)-(12) išraiškomis: [19]

$$f_t = \sigma(W_f \cdot h_{t-1} + U_f \cdot \widetilde{x_t} + b_f) \tag{7}$$

$$i_t = \sigma(W_i \cdot h_{t-1} + U_i \cdot \widetilde{x_t} + b_i) \tag{8}$$

$$o_t = \sigma(W_o \cdot h_{t-1} + U_o \cdot \widetilde{x_t} + b_o) \tag{9}$$

$$\widetilde{c_t} = \tau(W_c \cdot h_{t-1} + U_c \cdot \widetilde{x_t} + b_c) \tag{10}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \tag{11}$$

$$h_t = o_t * \tau(c_t) \tag{12}$$

Čia $*$ žymi Hadamard sandaugą, $\cdot$ žymi standartinę matricų sandaugą. f_t – pamiršimo vartai, i_t – įvesties vartai, o_t – išvesties vartai, $\widetilde{c_t}$ – elemento aktyvacijos vektorius, c_t – elemento vektorius, h_t

– paslėptasis vektorius. σ – sigmoidės aktyvacijos funkcija, τ – hiperbolinio tangento aktyvacijos funkcija. Atitinkamai W_f, W_i, W_c, W_o, U_f, U_o, U_i, U_c svorių matricos, o b_f, b_i, b_o, b_c laisvieji nariai. $\widetilde{x_t}$ įvesties vektorius.

Pažymime, jog būtent sigmoidės yra parinktos vartams, kadangi, šios funkcijos reikšmių sritis yra intervalas [0; 1], o tai leidžia panaikinti arba išlaikyti nenaudingą ir naudingą informaciją. Taip pat, atkreipiame dėmesį, jog teksto požymiais tampa h_t vektorius, 26 paveikslėlyje šis vektorius pažymėtas *h*.

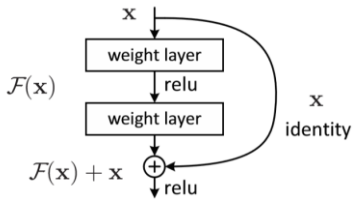
Taigi, šiame skyriuje apžvelgėmė teksto apdorojimo metodus, pateikėmė jų privalumus bei trūkumus. Taip pat, išnagrinėjome techninius veikimo aspektus.

2.3. Skaitmeninių vaizdų apdorojimo metodai

Svarbu pažymėti, jog, norint gauti aukštus rezultatus, taip pat yra svarbu ne tik apdirbti kokybiškai tekstą, bet ir vaizdą. Dėl to šiame skyrelyje pateiksime sąsūkų neuroninių tinklų architektūras, kurias naudosime šiame darbe.

2.3.1. ResNetV2 tinklo architektūra

Kaip ir rekurentinai tinklai taip ir gilūs sąsūkų neuroniniai tinklai susiduria su nykstamo gradiento problema. Ši problema buvo sprendžiama, įtraukiant normalizavimo sluoksnius, kurie leido giliems tinklams konverguoti, naudojantis stochastinio gradiento nusileidimo metodu. Tačiau, nors gilūs tinklai ir pradėjo mažiau kentėti nuo nykstančio gradiento, buvo pastebėta degradacijos problema: pridedant kuo daugiau sluoksnių į tinklą, modelio tikslumo augimas sustoja ir pradeda staigiai kristi. Pažymime, kad ši situacija atsiranda ne dėl tinklo permokymo. 2015 metais [20] šaltinio autoriai pateikė tinklo ResNet idėją, kuri kovoja su aukščiau aprašyta problema. Autoriai vietoj to, kad kiekvieno sluoksnio išvestį tiesiogiai perduotų kitam sluoksniui, leidžia tinklui išmokti liekanų atvaizdavimus (angl. Residual mapping) [20]



28 pav. ResNet tinklo blokas [20]

Atkreipiame dėmesį, jog ši architektūra vietoj tiesioginių sluoksnių naudoja blokus, kurie atitinkamai savyje turi kelis sąsūkos sluoksnius. 28 paveikslėlyje yra pavaizduotas vienas iš šių blokų. Apibrežus norimą atvaizdavimą kaip $H(x)$, autoriai leidžią sujungtiems netiesiniams sluoksniams išmokti kitą atvaizdavimą $F(x) = H(x) - x$. Tuomet originalus atvaizdavimas gali būti išreiškiamas kaip $F(x) + x$. Autorių idėja buvo, jog tinklui yra lengviau optimizuoti liekanų atvaizdavimą nei originalų atvaizdavimą. Kitais žodžiais kalbant, jei, kraštutiniu atveju, identiškas (angl. Identity) atvaizdavimas yra naudingas, tai lengviau yra liekamąjį atvaizdavimą artinti link nulio, nei kad apskaičiuoti identišką atvaizdavimą, apmokant kelis sujungtus netiesinius sluoksnius. Literatūroje dažnai $F(x) + x$

formuluotė yra minima kaip neuroninis tinklas su trumpesniųjų kelių sujungimais. Toliau pateiksime matematinį šio modelio efektyvumo pagrindimą. [20]

Pažymėjus x_l kaip *l*-tojo sluoksnio įvestį, W_l kaip *l*-tojo sluoksnio svorių matricą, *F* kaip liekamąją funkciją, x_{l+1} bloką galima išreikšti (13) formule: [21]

$$x_{l+1} = x_l + F(x_l, W_l) \tag{13}$$

Pasinaudojus rekursija bet koks gilesnis blokas x_L bus išreiškiamas (14) arba (15) formule. [21]

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{14}$$

$$x_L = x_0 + \sum_{i=0}^{L-1} F(x_i, W_i) \tag{15}$$

Iš (15) formulės matoma, jog x_L blokas yra išreiškiamas kaip visų prieš tai buvusių liekamųjų funkcijų suma, prie kurios yra pridedama įvestis x_0. Tuo tarpu, standartinių sąsūkų neuroninių tinklų sluoksniai x_L, dėl paprastumo neatsižvelgiant į ReLU aktyvacijos funkcijas ir normalizavimą gali būti išreiškiami (16) formule. [21]

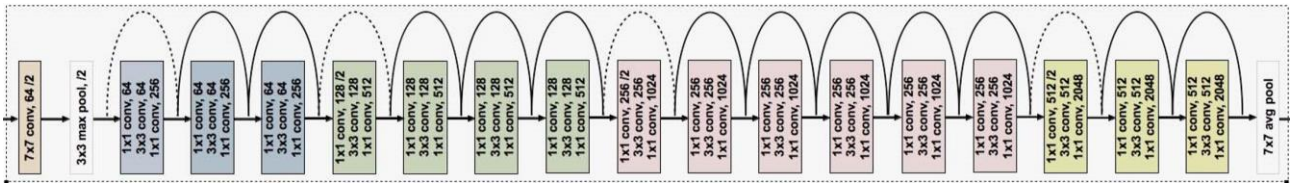
$$x_L = \prod_{i=0}^{L-1} W_i \cdot x_0 \tag{16}$$

Iš (14) formules gaunama, jog klaidos funkcijos *L* gradientas, naudojantis sudetinės funkcijos išvestinės skaičiavimo taisykle, gali būti išreiškiamas (17) formule. [21]

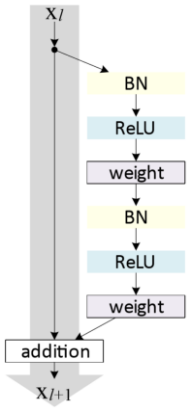
$$\frac{\partial L}{\partial x_l} = \frac{\partial L}{x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial L}{x_L} \cdot \left(1 + \frac{\partial}{x_l} \sum_{i=l}^{L-1} F(x_i, W_i)\right) \tag{17}$$

Būtina pastebėti, jog $\frac{\partial L}{\partial x_l}$ gradientas gali būti išskaidomas į dviejų narių sudetį. Pirmasis narys yra $\frac{\partial L}{x_L}$, kuris siunčia informaciją neatsižvelgdamas į jokius svorių sluoksnius, o antrasis narys yra $\frac{\partial}{x_l} \sum_{i=l}^{L-1} F(x_i, W_i)$, kuris siunčia informaciją per svorių sluoksnius. Svarbu yra tai, kad gradientas $\frac{\partial L}{\partial x_l}$ su didele tikimybe neįgys nulinės reikšmės, nes sunkiai tikėtina, kad narys $\frac{\partial}{x_l} \sum_{i=l}^{L-1} F(x_i, W_i)$ visiems rinkinio stebėjimams įgys -1 reikšmę. [21]

29 pav. pateikiame ResNet50V2 tinklo architektūrą, o 30 pav. operacijų atlikimo schemą:



29 pav. ResNet50V2 architektūra [22]



30 pav. ResNet50V2 tinklo operacijų atlikimo schema [23]

Iš 29 pav. galima matyti, jog ResNet50V2 turi:

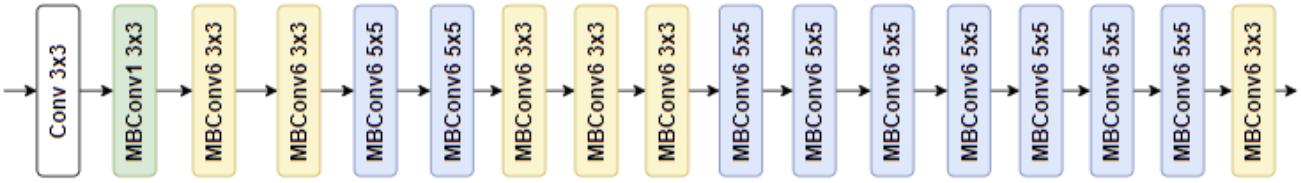
- 3 blokus, kurie susideda iš 64 1x1 formato sąsūkos filtrų, 64 3x3 formato sąsūkos filtrų ir 256 1x1 formato sąsūkos filtrų;
- 4 blokus, kurie susideda iš 128 1x1 formato sąsūkos filtrų, 128 3x3 formato sąsūkos filtrų ir 512 1x1 formato sąsūkos filtrų (išskyrus pirmąjį bloką);
- 6 blokus, kurie susideda iš 256 1x1 formato sąsūkos filtrų, 256 3x3 formato sąsūkos filtrų ir 1024 1x1 formato sąsūkos filtrų (išskyrus pirmąjį bloką);
- 3 blokus, kurie susideda iš 512 1x1 formato sąsūkos filtrų, 512 3x3 formato sąsūkos filtrų ir 2048 1x1 formato sąsūkos filtrų (išskyrus pirmąjį bloką).

Tuo tarpu, iš 30 pav. galima matyti operacijų atlikimo tvarką: rinkinio normalizavimas → ReLU aktyvacijos funkcijos pritaikymas → sąsūkos opereacija → rinkinio normalizavimas → ReLU aktyvacijos funkcijos pritaikymas → sąsūkos opereacija → rinkinio normalizavimas → ReLU aktyvacijos funkcijos pritaikymas → sąsūkos opereacija → Gauto rezultato sudėtis su įvestimi

2.3.2. EfficientNetB0 tinklo architektūra

[24] tyrime autoriai pateikė EfficientNet tinklų architektūras, kurios efektyviai sugeba apskaičiuoti tinklų gylį ir ne tik nenusileidžia kitoms architektūroms skaičiavimų tikslumais, bet daugelį iš jų lenkia. Autoriai pažymi, jog dauguma kitų tyrėjų tinklų yra konstruojami, turint omenyje fiksuotus skaičiavimų resursus, o, jei resursai leidžia, tuomet tinklai yra pagilinami. Pasak straipsnio autorių, toks požiūris nėra visiškai teisingas, nes žymiai gilesni tinklai ne visuomet suteikia aukštesnius rezultatus, o gebėjimas optimizuoti architektūrą įgalina modelius taip pat ir žymiai efektyviau atlikti

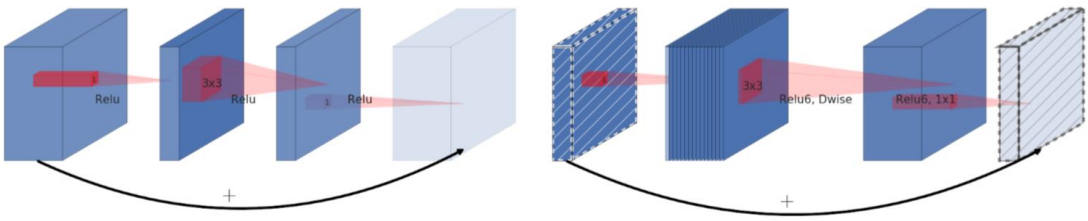
skaičiavimus. [24] paveikslėlyje pateikiame EfficientNetB0 tinklo architektūrą (šioje schemoje nežymime paskutiniųjų pilnai sujungtų sluoksnių bei liekamųjų ryšių).



31 pav. EfficientNetB0 tinklo architektūra [24]

Iš 31 pav. galima matyti jog ši architektūra susideda iš vieno MBConv1 3x3 bloko, šešių MBConv6 3x3 blokų, devynių MBConv6 5x5 blokų bei vieno sąsūkų Conv 3x3 bloko.

MBConv blokai buvo pritaikyti [25] autorių. Šių blokų idėja yra panaši į prieš tai skyrelyje aptartus liekamuosius (anngl. Residual) blokus ir dažnai yra vadinami atvirkštiniais liekamaisiais blokais (angl. Inverted residual block) (žr 32 pav.)

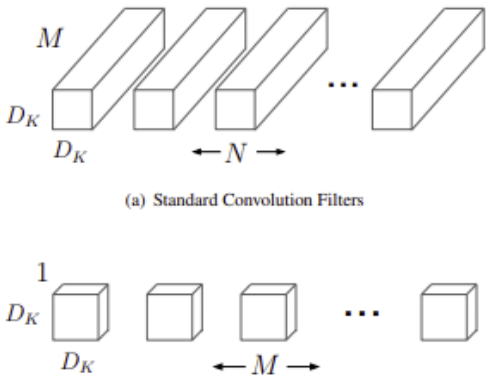


32 pav. Liekamieji blokai (kairėje) ir atvirkštiniai liekamieji blokai (dešinėje) [25]

Pagrindiniai liekamųjų blokų ir atvirkštinių liekamųjų blokų architektūriniai skirtumai yra šie:

- Liekamieji blokai naudoja žymiai daugiau požymių žemėlapių įvestyje, kuriems pirmiausia yra pritaikoma 1x1 sąsūkos operacija tuomet yra pritaikoma didesnio formato filtro sąsūkos operacija ir vėl taikoma 1x1 sąsūkos operacija tam, kad įvestis ir išvestis galėtų būti sudedamos (žr. 29 pav.)

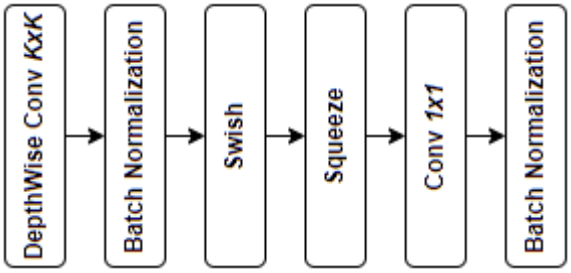
- Atvirkštiniai liekamieji blokai naudoja mažiau požymių žemėlapių įvestyje, tačiau, su 1x1 sąsūkos operacija apskaičiuoja daugiau požymių žemėlapių tuomet pritaiko 3x3 giluminę (angl. Deepwise) sąsūką bei 1x1 sąsūką tam, kad įvestis ir išvestis galėtų būti sudedamos [25]



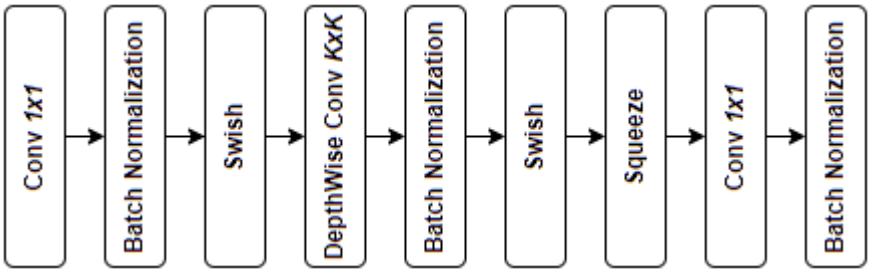
33 pav. Sąsūkos operacija (viršuje) ir giluminė sąsūkos operacija (apačioje) [26]

33 paveikslėlyje galima matyti, kaip giluminė sąsūkos operacija skiriasi nuo standartinės sąsūkos operacijos. Pažymime, jog giluminė taiko *M* filtrų vienam požymių žemėlapiui (tam, kad sumažintų skaičiavimų resursus bei išsaugotų kiekvieno filtro sukauptą informaciją), kai standartinės operacijos atveju kiekvienas filtras turi tiek pat kanalų, kiek ir įvestis. Pažymime, jog giluminės atveju nėra generuojami nauji požymių žemėlapiai, o, norint juos gauti, reiktų atlikti dar papildomai 1x1 sąsūkos operaciją. [26]

Žemiau pateikiame detalesnę informaciją, iš kokių architektūrinių idėjų susideda MBconv blokai:



34 pav. MBconv1 blokas [27]



35 pav. MBconv6 blokas [27]

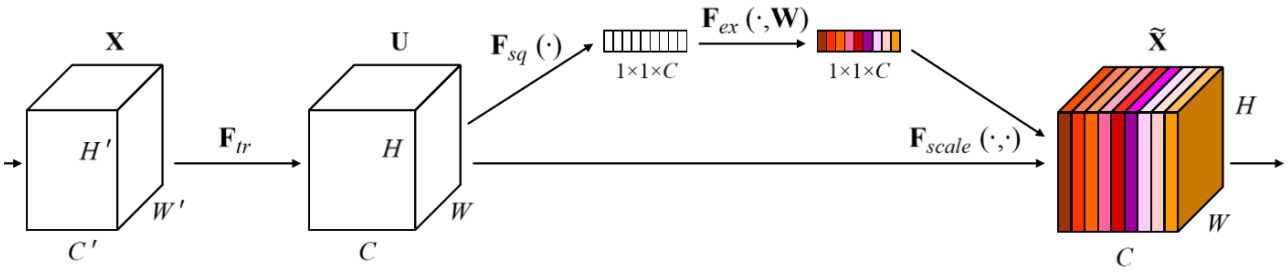
Iš 34 paveikslo galima matyti, jog MBconv1 blokas pirmiausia pritaiko giluminė sąsūkos operaciją (EfficientNetB0 atveju $K = 3$) paskui yra normalizuojamas rinkinys, pritaikoma Swish aktyvacijos funkcija (žr. 18 išraišką) tuomet pritaikomas suspaudimo sluoksnis (angl. Squeeze), o darbas pasibaigia pritaikius 1x1 sąsūkos operaciją ir rinkinio normalizavimą. [27]

Tuo tarpu, 35 paveiksle yra pateikiamas MBconv6 blokas. Pirmiausia šiame bloke yra atliekama 1x1 sąsūkos operacija, tuomet rinkinio normalizavimas ir Swish aktyvacijos funkcija, paskui vėl taikoma sąsūkos operacija, tik, šiuo atveju, KxK formato, Swish aktyvacija, suspaudimo sluoksnis ir darbas pasibaigia ties 1x1 sąsūkos operacija bei rinkinio normalizavimu. [27]

Žemiau pateikiame Swish aktyvacijos funkcijos išraišką bei informaciją apie suspaudimo sluoksnį:

$$Swish = x \cdot \frac{1}{1 + e^{-x}} \tag{18}$$

Nors matematiškai įrodyti, šios funkcijos privalumus prieš kitas aktyvacijas yra sudėtinga, tačiau, [28] autoriai mano, jog ši funkcija suteikia tam tikroms architektūroms geresnius rezultatus, dėl to, jog funkcija yra nemonotoniška, tolydi bei neapribota iš viršaus ir apribota iš apačios.



36 pav. Suspaudimo sluoksnis [29]

Tuo tarpu, pagrindinė suspaudimo sluoksnio (žr. 36 pav.) idėja yra ta, jog šis mechanizmas sugeba išryškinti labiau informatyvius požymių žemėlapius, o mažiau informatyvius suspausti. Iš esmės, sluoksnis atlieka šiuos veiksmus: įvesties požymių žemėlapiams, kurių formatas yra *HxWxC* yra pritaikomas globalus sutelkimas, pilnai sujungtas sluoksnis, ReLU aktyvacijos funkcija, pilnai sujungtas sluoksnis ir Sigmoidės aktyvacijos funkcija. Gaunamas vektorius yra *1x1xC* dimensijos, kitaip tariant, kiekvienam kanalui yra paskaičiuojamas atitinkamas svoris ir sudauginamas su pačiu kanalu. [29]

2.4. Teksto bei vaizdo požymių suliejimas

2.4.1. Požymių suliejimas sumos, konkatenacijos ir sandaugos operacijomis

Sprendžiant darbe nagrinėjamą problemaiką, apskaičiavus teksto ir vaizdo požymius, yra siekiama kaip geriau alikti jų suliejimą (angl. fushion), jog svarbiausia informacija būtų neprarandama ir kokybiškai analizuojama sekančiuose modelio etapuose. Vieni iš paprasčiausių variantų yra pritaikyti konkatenacijos, sudėties ar sandaugos operacijas. Pažymėjus vaizdo požymių vektorių $v^{\mathbb{N}}$, teksto požymių vektorių $t^{\mathbb{N}}$, o konkatenacijos operaciją $*$, šie suliejimo būdai gali būti išreiškiami (19), (20), (21) išraiškomis:

$$v + t = (v_1 + t_1, v_2 + t_2, v_3 + t_3, \dots, v_n + t_n) \tag{19}$$

$$v \cdot t = (v_1 \cdot t_1, v_2 \cdot t_2, v_3 \cdot t_3, \dots, v_n \cdot t_n) \tag{20}$$

$$v * t = (v_1, v_2, v_3, \dots\ v_n, t_1, t_2, \dots, t_n) \tag{21}$$

2.4.2. Požymių suliejimas suformuojant matricą ir pritaikant sąsūkos operaciją

Ši metodologijos idėja remiasi daugiau informacijos teikiančiu požymių suliejimu, kadangi originaliai apskaičiuoti požymiai yra suliejami ne tik sumos bei sandaugos veiksmais, o taip pat, iš jų suformavus matricą yra pritaikoma sąsūkos, bei maksimalaus sutelkimo operacijos. Taikant tokius pačius žymėjimus kaip ir 2.4.1 skyrelyje, matematiškai šią idėją galima išreikšti 22-26 išraiškomis:

$$v` = (v_n, v_{n-1}, \dots\ v_3, v_2, v_1) \tag{22}$$

$$t` = (t_n, t_{n-1}, \dots\ t_3, t_2, t_1) \tag{23}$$

$$v + t = (v_1 + t_1, v_2 + t_2, v_3 + t_3, \dots, v_n + t_n) \tag{24}$$

$$v \cdot t = (v_1 \cdot t_1, v_2 \cdot t_2, v_3 \cdot t_3, \dots, v_n \cdot t_n) \tag{25}$$

$$M = \begin{pmatrix} v_1 & t_1 & (v \cdot t)_1 & (v + t)_1 & t`_1 & v`_1 \\ v_2 & t_2 & (v \cdot t)_2 & (v + t)_2 & t`_2 & v`_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ v_n & t_n & (v \cdot t)_n & (v + t)_n & t`_n & v`_n \end{pmatrix} \tag{26}$$

Atlikus (22)-(26) išraiškų veiksmus, *M* matricai yra pritaikoma 2 kartus sąsūkos operacija (3x3 filtro dydžio) su 16 ir 8 filtrais bei maksimalaus sutelkimo sluoksniai. Gauti požymiai yra ištiesinami ir perduodami pilnai sutelktam sluoksniui, kuris apskaičiuoja klasių tikimybės. Taip pat, yra pritaikoma ir dropout reguliarizacija. (žr. 61 pav.)

2.4.3. Požymių suliejimas suformuojant matricą, suteikiant jai svorius ir pritaikant sąsūkos operaciją

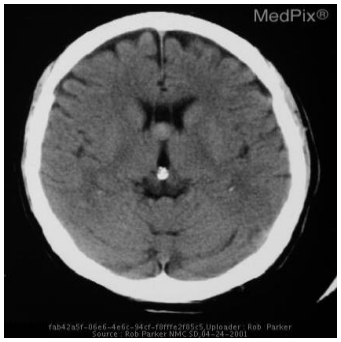
Pažymime, jog šios metodikos idėja yra ta, jog modelis gėbėtų „pasirinkti“, kurie požymiai jam yra svarbiausi teisingo atsakymo pateikimui. Tam, kad pasiektume šią implementaciją *M* matricai pritaikome sąsūkos operaciją (3x3 filtro dydžio) su 16 filrų ir sąsūkos operaciją (1x1 filro dydžio) su 1 filtru. Gautas rezultatas yra aktyvuojamas sigmoidės funkcija ir sudauginamas su *M* matrica. Pažymime, jog būtent sąsūkos operaciją padeda išrinkti naudingiausius požymius, kadangi jos reikšmių sritis yra intervalas [0; 1].

3. Tyrimas

Šiame skyriuje pateiksime informaciją apie tyrime naudotus duomenis, gautus rezultatus.

3.1. Duomenys

Darbo tyrimui naudosime VQA-Med-2019 [30] imtį. Ši imtis susideda iš radiologinių vaizdų bei klausimų, kurie gali būti suskirstomi į 4 pagrindines kategorijas: klausimai apie vaizdo kokybės normalumą, klausimai apie vaizdų anomaliją, vaizdo tipo klausimai bei klausimai apie kontrastinio tirpalo panaudojimą. 37-41 pav. pateikiame keletą šios imties vaizdų. Pažymime, jog šiame darbe naudosime tik klausimus, kurių atsakymai yra „taip“ arba „ne“.



37 pav. Darbe naudojamas imties pirmasis pavyzdys [30]

38 pav. Darbe naudojamas imties antrasis pavyzdys [30]

39 pav. Darbe naudojamas imties trečiasis pavyzdys [30]

40 pav. Darbe naudojamas imties ketvirtas pavyzdys [30]

Mokymams bus taikoma kryžminio išmaišymo metodika: visas rinkinys yra išskaidomas į 5 lygias dalis ir apmokomi 5 modeliai, kurių tikslumo metrikos galutiniui įverčiui yra suvidurkinamos. Kiekviena iš šių dalių turės 305 vaizdus testavimo dalyje ir 1218 vaizdų apmokymo dalyje (vaizdų formatas bus 512x512). Pažymime, jog validacijos imtis nėra taikoma dėl vaizdų trūkumo. Taip pat mokymų metu kiekvienam vaizdui atsitiktinai bus taikomos posūkio, postūmio, kontrasto, suliejimo, iškraipymo augmentacijos.

Imtyje iš viso turime 1560 klausimų-atsakymų porų, 24 unikalius žodžius (atlikus duomenų išvalymą), 33 unikalius klausimus. Pažymime, jog kiekvienam sakiniui buvo atliekamas paruošimas,

t.y. išvalomi prasmės nepridedantys žodžiai, pašalinami nereikalingi simboliai, suvienodinami tarpai bei didžiosios ir mažosios raidės ir taip gaunami sakiniai, kurių maksimalus ilgis yra 5 žodžiai. Taip pat, svarbu tai, jog žodžių vektoriams atvaizduoti naudosimės [31] šaltinyje pateikiamais 300 dimensijų vektoriais, kurie buvo apmokyti remiantis skip-gram metodologija. Žemiau pateikiame keletą klausimų pavyzdžių:

- Ar kompiuterinė tomografija yra normali?
- Ar tai magnetinio rezonanso vaizdas?
- Ar kontrastinis tirpalas buvo duotas pacientui?
- Ar tai t1 klasės vaizdas?

3.2. Darbe taikytų architektūrų parametrų skaičius

Žemiau esančioje lentelėje (žr. 7 lent.) pateikiame kiekvieno darbe taikyto modelio parametrų skaičių.

7 lentelė. Darbe taikomų modelių mokomų parametrų skaičius

Apjungimas / **Modeliai**	Sumavimas	Konkatenacija	Matricos formavimas	Matricos formavimas bei svorių įvedimas
ResNet50V2, LSTM	56 005 814	**33 468 342**	**55 214 762**	55 214 939
ResNet50V2, Rekurentiniai tinklai	49 934 774	27 397 302	49 143 722	49 143 899
EfficientNetB0, LSTM	**12 899 762**	13 560 242	12 108 710	**12 108 887**
EfficientNetB0, Rekurentiniai tinklai	6 828 722	7 489 202	6 037 670	6 037 847

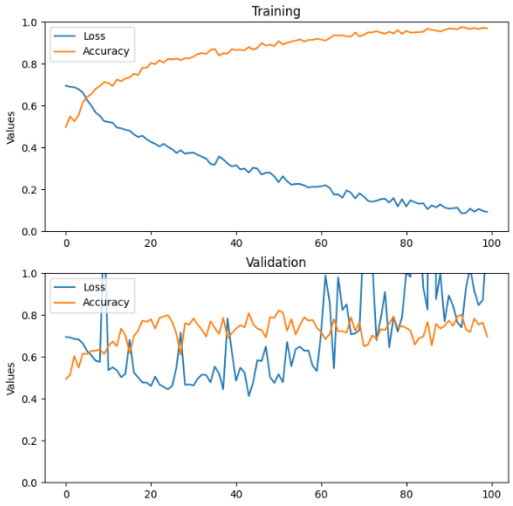
Iš 7 lentelės galima matyti, jog daugiausia parametrų turėjo ResNet50V2 ir LSTM architektūros, kurių požymių vektoriai buvo apjungiami formuojant matricą ir jai pritaikant sąsūkos operacijas, kai, tuo tarpu, mažiausiai parametrų turėjo EfficientNetB0 ir rekurentinių tinklų idėja požymių apjungimui pasitelkianti matricos formavimą su požymių svorių apskaičiavimu.

3.3. Darbe taikytos architektūros su konkatenacijos operacijos požymių apjungimu

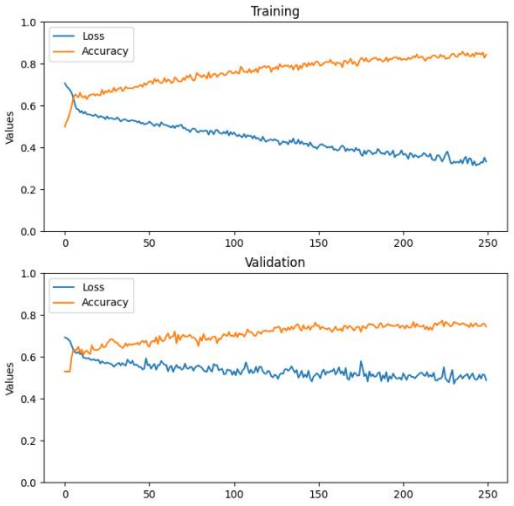
Žemiau esančioje dalyje pateiksime tiriamajame darbe taikytas architektūras, jų mokymosi grafikus, gautus tikslumo įverčius bei palyginsime juos skirtingų architektūrų lygmeny. Pažymime, jog visiems modeliams buvo taikomas Adam optimizatorius su mokymosi žingsniu lygiu 0,00002 ir rinkinio dydžiu lygiu 8. Tuo tarpu, vaizdo požymiais tapdavo paskutiniai vaizdo modelių sluoksniai, o tekstui buvo apskaičiuojami požymių vektoriai, kurių dimensija buvo lygi 1280. 3.2.1-3.2.4 skyrelių architektūrose apjungdavome požymių vektorius pasinaudodami konkatenacijos operacija ir tuomet sekdavo 4 pilnai sujungtieji sluoksniai, kurie turėjo atitinkamai po 516, 256, 16, 2 neuronus. Tam,

kad, kaip įmanoma labiau, sumažinti persimokymo problemą naudojome atsitiktinius išmetimo sluoksnius. Taip pat į tyrimą įtraukėme ir bandymus su ImageNet pradiniais svoriais, t.y. vaizdo modeliai mokėsi ne nuo atsitiktinių svorių, o nuo iš anksto apskaičiuotų ImageNet svorių. Atkreipiame dėmesį, jog pateikiame tik vieno iš penkių modelio mokymosi paveikslą, kadangi visų modelių mokymosi procesai žymiai nesiskyrė.

41 pav. LSTM ir EfficientNetB0 architektūra (konkatenuojant požymius)



42 pav. LSTM ir EfficientNetB0 su ImageNet pradiniais svoriais mokymosi grafikas

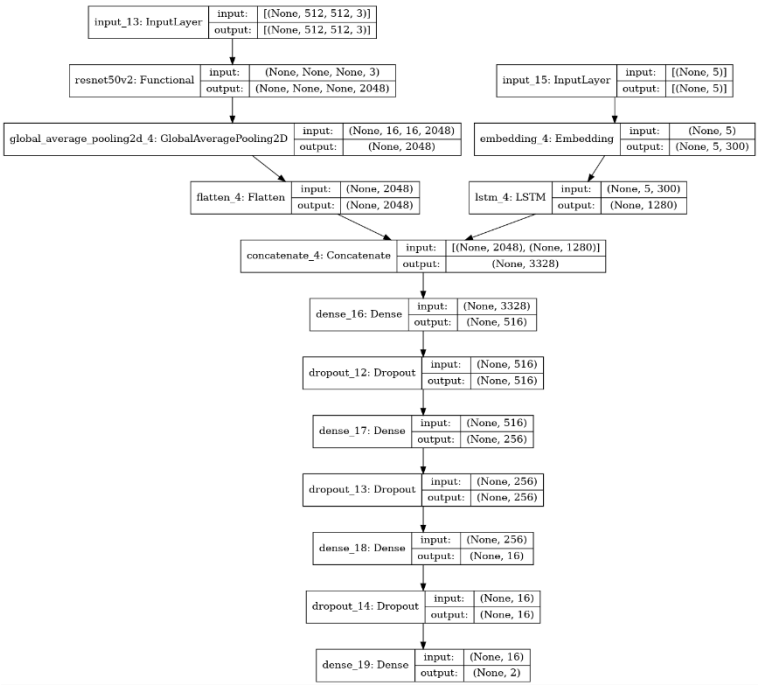


43 pav. LSTM ir EfficientNetB0 su atsitiktiniais pradiniais svoriais mokymosi grafikas

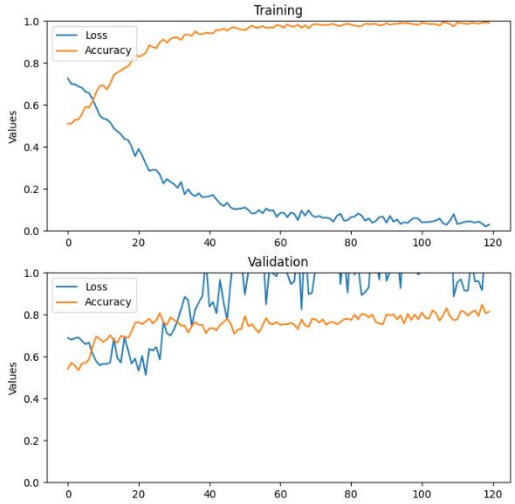
8 lentelė. LSTM ir EfficientNetB0 modelio tikslumai konkatenuojant požymius

Klausimo kategorija	Tikslumas su ImageNet svoriais	Tikslumas su atsitiktiniais svoriais
Vaizdo normalumas	76,34%	76,34%
Anomalijos vaizde	64,7%	64,7%
Vaizdo tipas	92,25%	91,47%
Kontrasto tirpalo panaudojimas	71,4%	72,01%
Bendras vidurkis	81,61%	81,53%

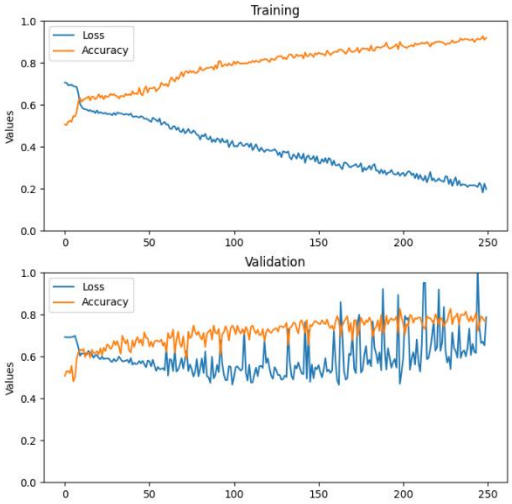
Iš 42 pav. galima matyti, jog LSTM ir EfficientNetB0 modelis ties 60 epocha pradeda persimokyti, o, tuo tarpu, iš 43 pav. galima matyti, jog naudojantis atsitiktiniais svoriais tinklo tikslumas nusistovi ties 200 epocha. Kaip bebūtų, yra gaunamas per 0,7 % geresnis rezultatas remiantis ImageNet pradiniais svoriais (žr. 8 lent.). Pažymime, jog pranašumas yra labai neryškus ir tai pagrindžia faktas, jog vaizdo normalumo ir anomalijos kategorijų klausimų tikslumo vidurkiai buvo vienodi.



44 pav. LSTM ir ResNet50V2 architektūra (konkatenuojant požymius)



45 pav. LSTM ir ResNet50V2 su ImageNet pradiniais svoriais mokymosi grafikas

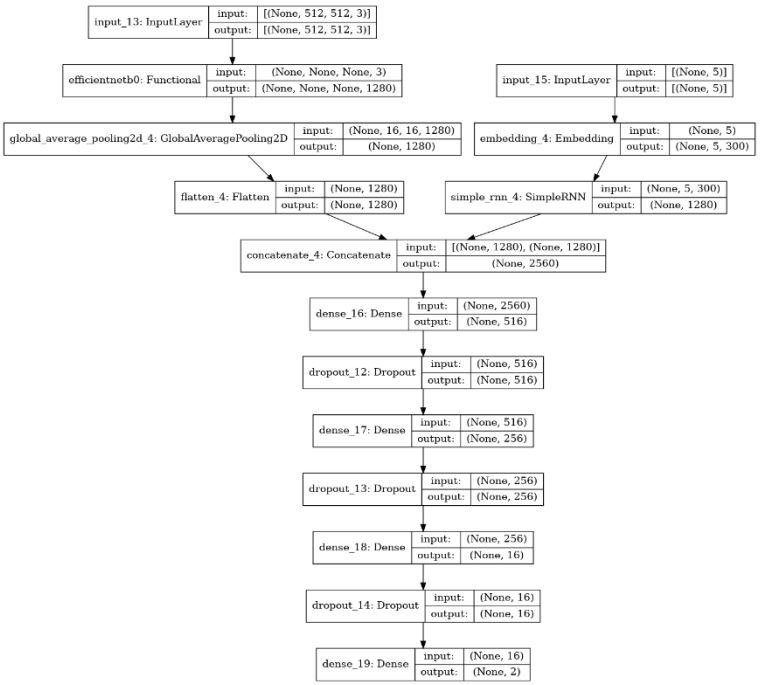


46 pav. LSTM ir ResNet50V2 su atsitiktiniais pradiniais svoriais mokymosi grafikas

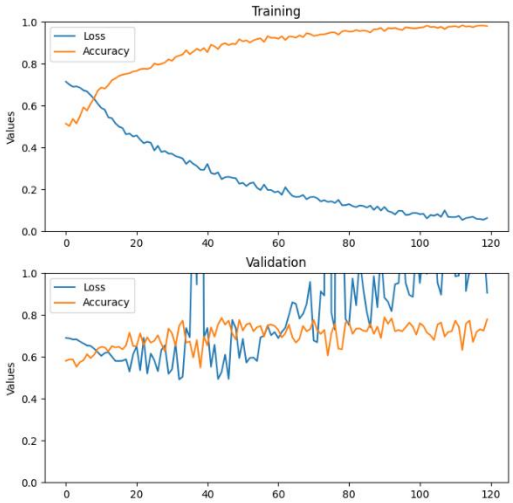
9 lentelė. LSTM ir ResNet50V2 modelio tikslumai konkatenuojant požymius

Klausimo kategorija	**Tikslumas su ImageNet svoriais**	**Tikslumas su atsitiktiniais svoriais**
Vaizdo normalumas	76,34%	80,64%
Anomalijos vaizde	72,54%	74,51%
Vaizdo tipas	84,38%	91,47%
Kontrasto tirpalo panaudojimas	62,38%	72,32%
Bendras vidurkis	74,29%	82,24%

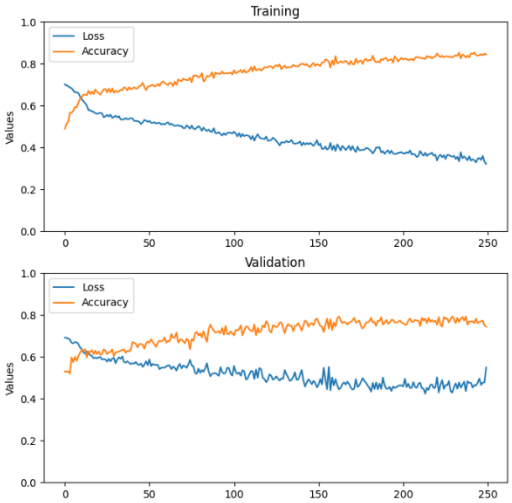
Iš abiejų LSTM ir ResNet50V2 tinklo mokymosi grafikų, galima pastebėti, jog įvyksta persimokymai: naudojantis ImageNet svoriais su šia situacija susiduriama jau ties 20 epocha, o su atsitiktiniais svoriais ties 150 epocha. Iš tikslumo metrikų (žr. 9 lent.), galima pastebėti, jog visų kategorijų tikslumai buvo aukštesni, kuomet buvo remiamasi atsitiktiniais svoriais, ir to rezultate gauname, kad bendras vidurkis yra aukštesnis per 7,95 %.



47 pav. RNN ir EfficientNetB0 architektūra (konkatenuojant požymius)



48 pav. RNN ir EfficientNetB0 su ImageNet pradiniais svoriais mokymosi grafikas

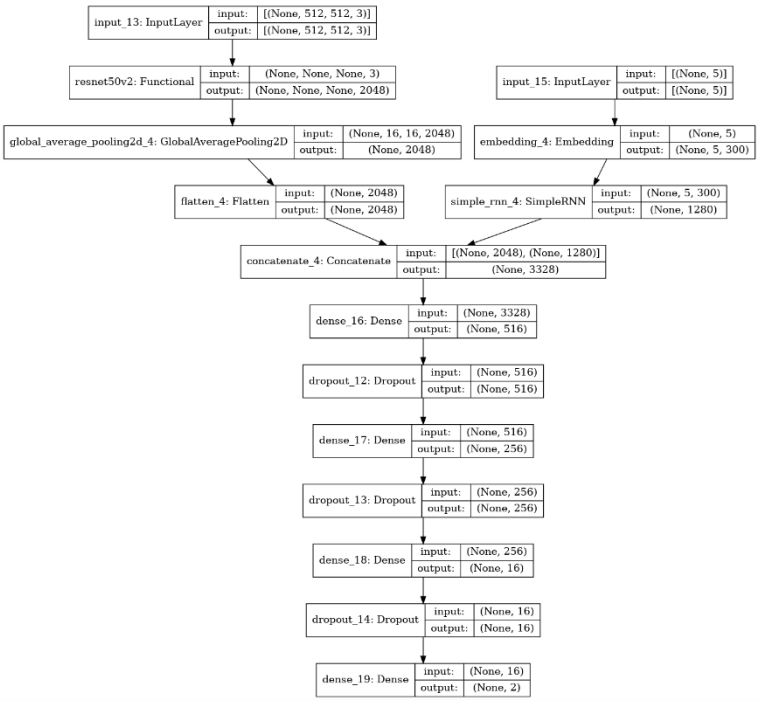


49 pav. RNN ir EfficientNetB0 su atsitiktiniais pradiniais svoriais mokymosi grafikas

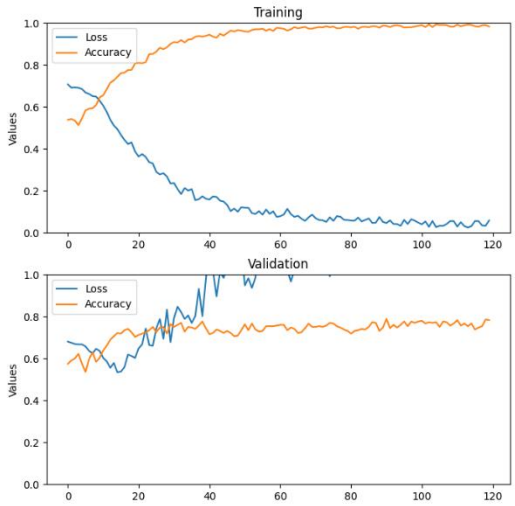
10 lentelė. RNN ir EfficientNetB0 modelio tikslumai konkatenuojant požymius

Klausimo kategorija	Tikslumas su ImageNet svoriais	Tikslumas su atsitiktiniais svoriais
Vaizdo normalumas	78,49%	74,19%
Anomalijos vaizde	68,62%	76,47%
Vaizdo tipas	89,5%	90,42%
Kontrasto tirpalo panaudojimas	70,33%	73,39%
Bendras vidurkis	80,11%	81,85%

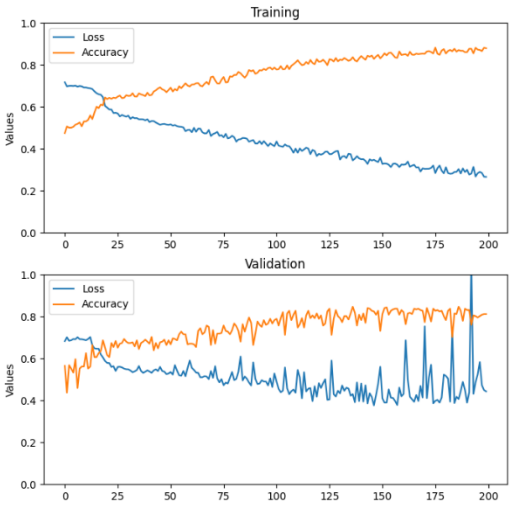
Iš RNN ir EfficientNetB0 mokymosi grafikų, galima matyti, jog naudojantis ImageNet svoriais tinklas susiduria su persimokymo problema ties 50 epocha, o, tuo tarpu, su atsitiktiniais svoriais tinklo tikslumai nusistovi ties 200 epocha. Iš gautų rezultatų (žr. 10 lent.) galima teigti, kad nežymiai geresni rezultatai yra gaunami naudojantis atsitiktiniais svoriais, kadangi bendras tikslumas yra aukštesnis per 1,74 % nei naudojantis ImageNet pradiniais svoriais.



50 pav. RNN ir ResNet50V2 architektūra (konkatenuojant požymius)



51 pav. RNN ir ResNet50V2 su ImageNet pradiniais svoriais mokymosi grafikas



52 pav. RNN ir ResNet50V2 su atsitiktiniais pradiniais svoriais mokymosi grafikas

11 lentelė. RNN ir ResNet50V2 modelio tikslumai konkatenuojant požymius

Klausimo kategorija	**Tikslumas su ImageNet svoriais**	**Tikslumas su atsitiktiniais svoriais**
Vaizdo normalumas	66,66%	72,04%
Anomalijos vaizde	62,74%	70,58%
Vaizdo tipas	79,13%	88%
Kontrasto tirpalo panaudojimas	64,37%	70,94%
Bendras vidurkis	71,66%	79,35%

Iš 51 paveikslo matome, jog su ImageNet pradiniais svoriais RNN ir ResNet50V2 tinklas persimoko ties 15 epocha, o su atsitiktiniais svoriais tikslumai nusistovi ties 150 epocha. Kalbant apie tikslumo įverčius, iš 11 lentelės matoma, jog visose kategorijose yra pranašesnis tinklas, kuris naudoja atsitiktinius svorius. Šį faktą patvirtina ir bendras klasifikavimo tikslumas, kuris yra didesnis per 7,69 %.

Taigi, šiame skyrelyje pateikėmė gautus modelių tikslumus, kuomet yra apjungiami požymiai, pasitelkiant konkatenacijos operaciją. Iš šių rezultatų galima formuluoti tokias išvadas:

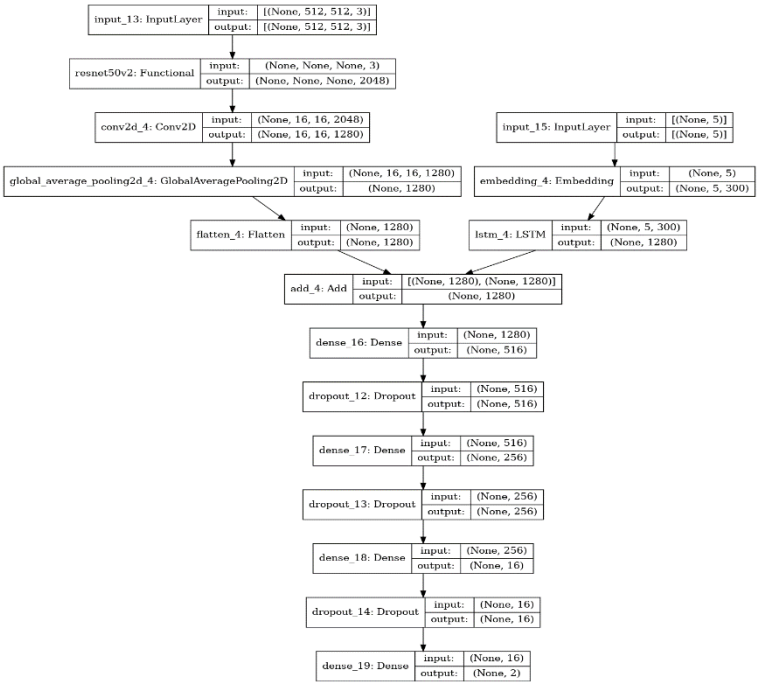
- ResNet50V2 tinklo atveju buvo gaunami geresni rezultatai, kai buvo remiamasi atsitiktiniais svoriais, o EfficientNetB0 atveju, tokio teiginio negalima teigti, nes rezultatai tarp atsitiktinių ir ImageNet svorių žymiai nesiskyrė. Tai galima sieti su faktu, jog ResNet50V2 tinklas yra gilesnis ir sugeba geriau mokytis su atsitiktiniais svoriais.

- Lyginant teksto požymių išskyrimo tinklus, EfficientNetB0 atveju gauname, jog LSTM tinklas buvo geresnis su ImageNet svoriais, o su atsitiktiniais svoriais (vaizdo) buvo geresnis RNN modelis. Kaip bebūtų skirtumai tikslumų yra per maži, jog galėtume teigti, kad kažkuris iš teksto modelių yra pranašesnis. ResNet50V2 atveju gauname, jog abejais atvejais LSTM teksto modelis generavo geresnius rezultatus nei RNN.

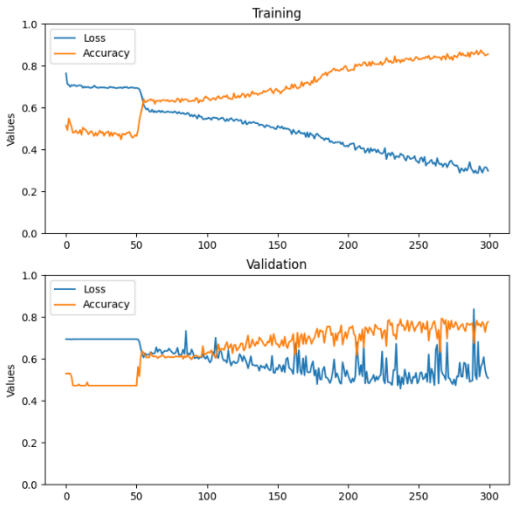
- Aukščiausias 82,24 % tikslumas buvo pasiektas naudojantis LSTM ir ResNet50V2 su atsitiktiniais vaizdo svoriais, o žemiausias tikslumas buvo gautas, taikant RNN ir ResNet50V2 modelį su ImageNet svoriais, ir buvo lygus 71,66 %.

3.4. Darbe taikytos architektūros su sudėties operacijos požymių apjungimu

Šiame skyrelyje pateiksime rezultatus, gautus sumuojant teksto ir vaizdo požymius. Pažymime, jog visi tinklų mokymo parametrai išlieka tokie patys kaip ir 3.2 skyrelyje. Norime pabrėžti, kad šioje tyrimo dalyje nemokysime modelių su ImageNet pradiniais svoriais, kadangi iš praeito skyrelio gautų rezultatų galima formuluoti išvadą, jog šis mokymo būdas nesuteikia reikšmingai aukštesnių tikslumų. Taip pat, ResNet50V2 atveju, šiek tiek keičiasi vaizdo požymių apskaičiavimo architektūra, kadangi, norint sudėti du vektorius, reikia išlaikyti tokias pačias dimensijas, o, kad tai pasiektume, įtraukėmė sąsūkų sluoksnį prieš globalų sutelkimą. (žr. 54 ir 56 pav.) Žemiau pateikiame modelių architektūras, mokymosi grafikus ir gautus tikslumo įverčius.

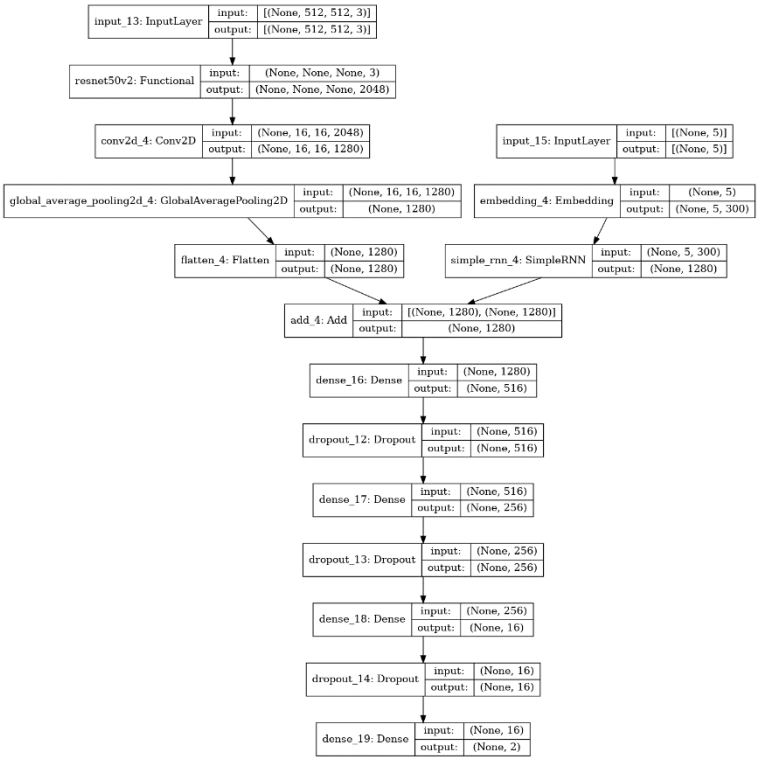


53 pav. LSTM ir ResNet50V2 architektūra (sumuojant požymius)

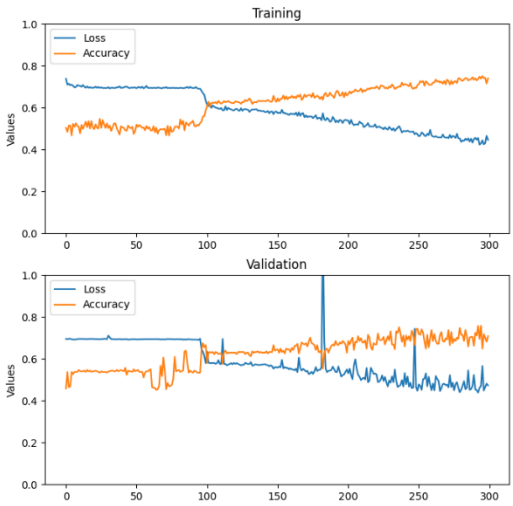


54 pav. LSTM ir ResNet50V2 mokymosi grafikas (sumuojant požymius)

Kaip ir minėjome iš 53 pav. galima matyti, jog LSTM ir ResNet50V2 architektūrai taikome papildomą sąsūkų sluoksnį bei vietoje daugybos operacijos naudojame sudėtį. Iš 54 pav. pastebime, jog modelio mokymasis nusistovi ties 230 epocha. Taip pat, yra įdomu tai, jog iki 50 epochos modelis praktiškai nesimokė ir tai siejame, jog tinklui buvo sudėtinga apieiti lokalų ekstremumą.

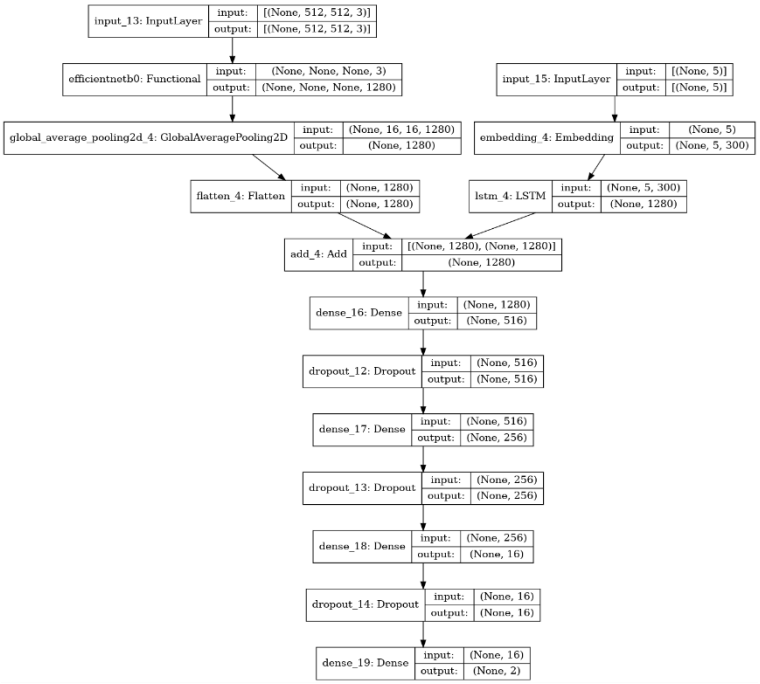


55 pav. RNN ir ResNet50V2 architektūra (sumuojant požymius)

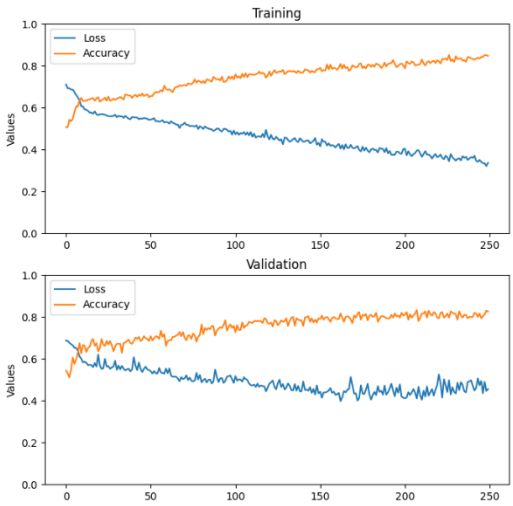


56 pav. RNN ir ResNet50V2 mokymosi grafikas (sumuojant požymius)

55 paveikslėlyje pateikiame RNN ir ResNet50V2 tinklo architektūrą, kuomet yra sumuojami požymių vektoriai, o 56 pav. pateikiame mokymosi grafiką iš kurio galim matyti, jog klaidos funkcijos nusistovėjimas įvyksta ties 250 epocha. Taip pat, kaip ir praeitu atveju, matome, kad iki 100 epochos modeliui sudėtingai sekasi tinkamai minimizuoti klaidos funkciją.

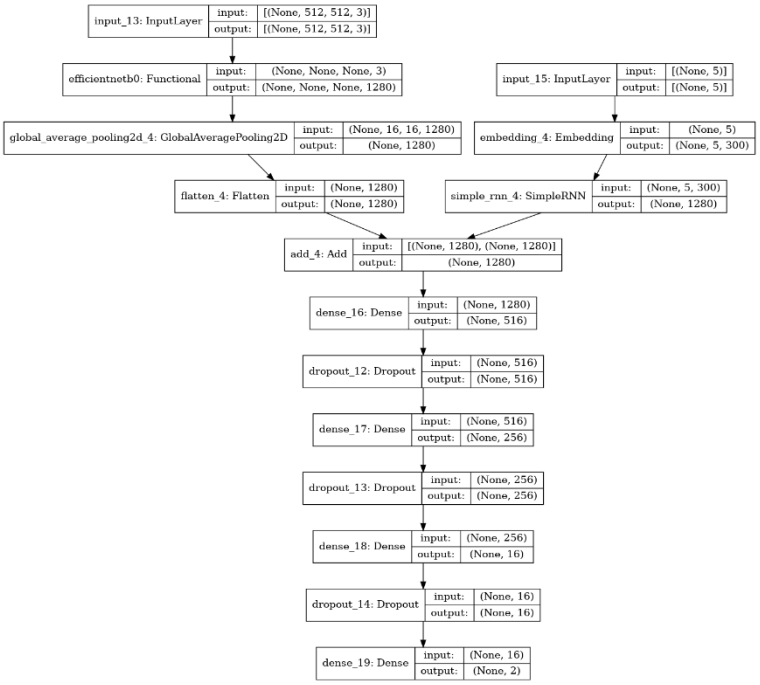


57 pav. LSTM ir EfficientNetB0 architektūra (sumuojant požymius)

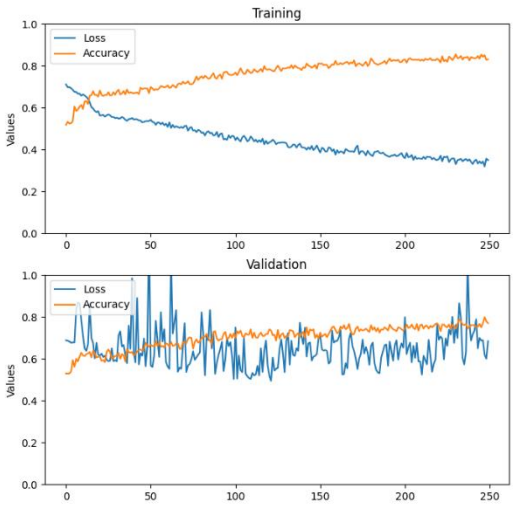


58 pav. LSTM ir EfficientNetB0 mokymosi grafikas (sumuojant požymius)

Iš 58 paveikslo galima matyti, jog LSTM ir EfficientNetB0 atveju tinklo tikslumai nusistovi ties 200 epocha, o 57 paveiksle yra pateikiama šio modelio architektūra.



59 pav. RNN ir EfficientNetB0 architektūra (sumuojant požymius)



60 pav. RNN ir EfficientNetB0 mokymosi grafikas (sumuojant požymius)

Iš 60 pav. galime matyti, jog RNN ir EfficientNetB0 mokymasis nusistovi ties 150 epocha. Žemiau esančioje lentelėje pateikiame gautų modelių tikslumo metrikas.

12 lentelė. Modelių tikslumai naudojantis sudėties apjungimu

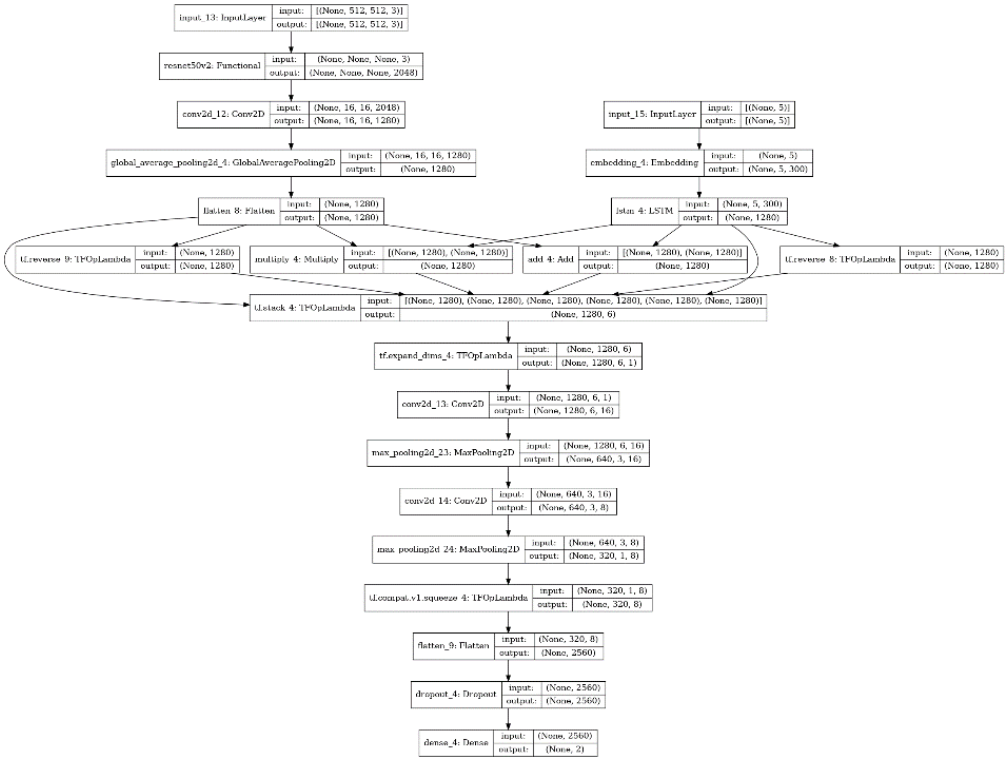
Klausimo kategorija	**LSTM ir EfficientNetB0 tikslumas**	**RNN ir EfficientNetB0 tikslumas**	**LSTM ir ResNet50V2 tikslumas**	**RNN ir ResNet50V2 tikslumas**
Vaizdo normalumas	77,41%	77,41%	80,64%	77,41%
Anomalijos vaizde	74,50%	70,58%	68,62%	78,43%
Vaizdo tipas	90,02%	88,84%	91,46%	87%
Kontrasto tirpalo panaudojimas	73,70%	72,93%	71,71%	66,05%
Bendras vidurkis	81,91%	80,89%	81,78%	77,36%

Iš 12 lentelės rezultatų, galima formuluoti tokias išvadas:

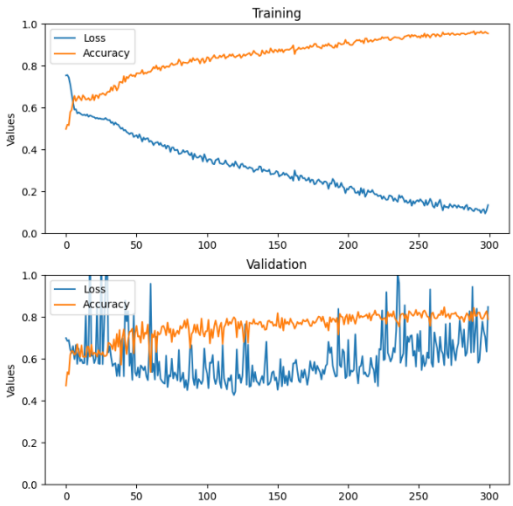
- LSTM ir EfficientNetB0 architektų atveju aukštesnis vidutinis tikslumas yra pasiekiamas sudedant požymių vektorius
- RNN ir EfficientNetB0 architektūrų atveju aukštesnis vidutinis tikslumas yra pasiekiamas konkatenuojant požymių vektorius
- LSTM ir ResNet50V2 architektūrų atveju aukštesnis vidutinis tikslumas yra pasiekiamas konkatenuojant požymių vektorius
- RNN ir ResNet50V2 architektūrų atveju aukštesnis vidutinis tikslumas yra pasiekiamas konkatenuojant požymių vektorius
- Galima teigti, jog konkatenacijos operacija leido pasiekti geresnius rezultatus, kadangi 3 iš 4 architektūrų buvo tikslesnės, kuomet buvo taikomas šis požymių suliejimo būdas, o ne suliejimas taikant sudėties operaciją.

3.5. Darbe taikytos architektūros požymius sudedant į matricą ir taikant sąsūkos operaciją

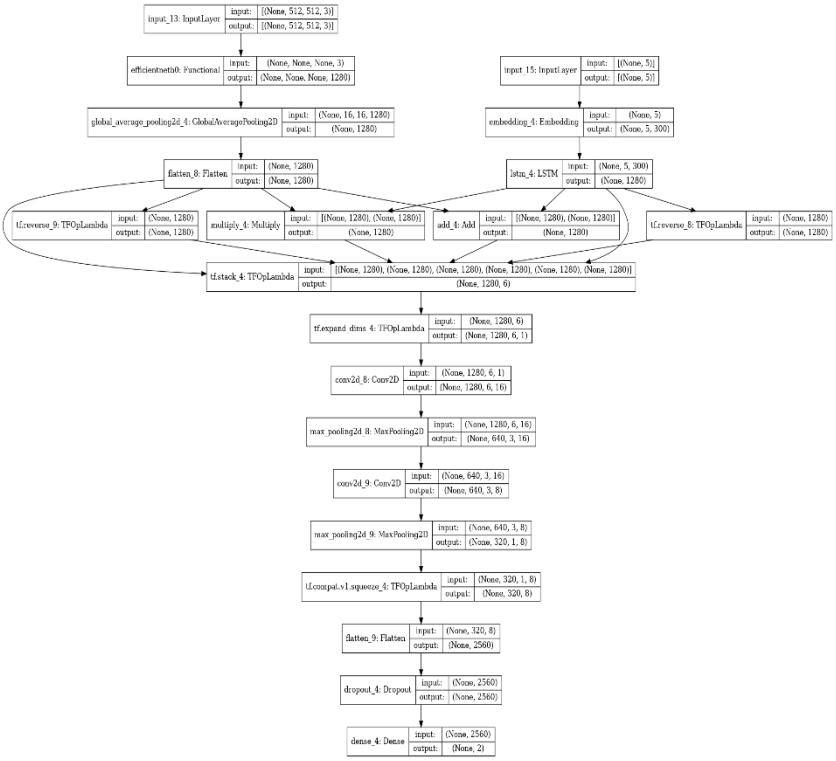
Pažymime, jog šiame skyriuje taip pat, kaip ir praeituose pateiksime kiekvienos architektūros grafinius pavyzdžius bei jų mokymosi grafikus. Architektūrų plačiau nekomentuosime, kadangi šio modelio idėja yra aprašyta 2.4.2 skyrelyje. Taip pat, pateikiame žemiau mokymosi grafikus, iš kurių galima matyti, jog ResNet50V2 tinklo mokymasis turėjo daugiau triukšmo nei EfficientNetB0. Skyrelio pabaigoje, 13 lentelėje, pateikiame gautų tikslumų įverčius, o mokymo parametrai išlieka tokie patys, kaip ir ankstesniuose skyreliuose.



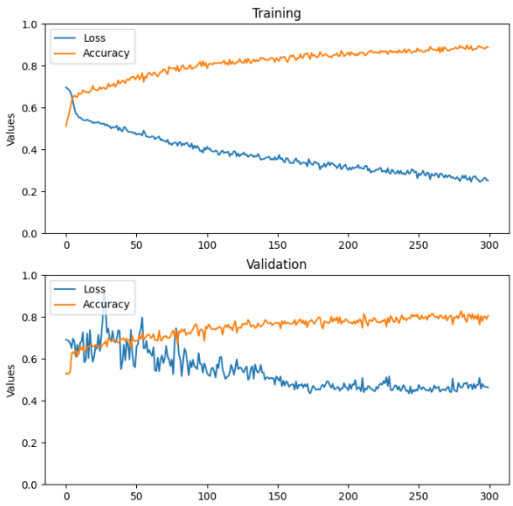
61 pav. LSTM ir ResNet50v2 architektūra (taikant sąsūką požymiams)



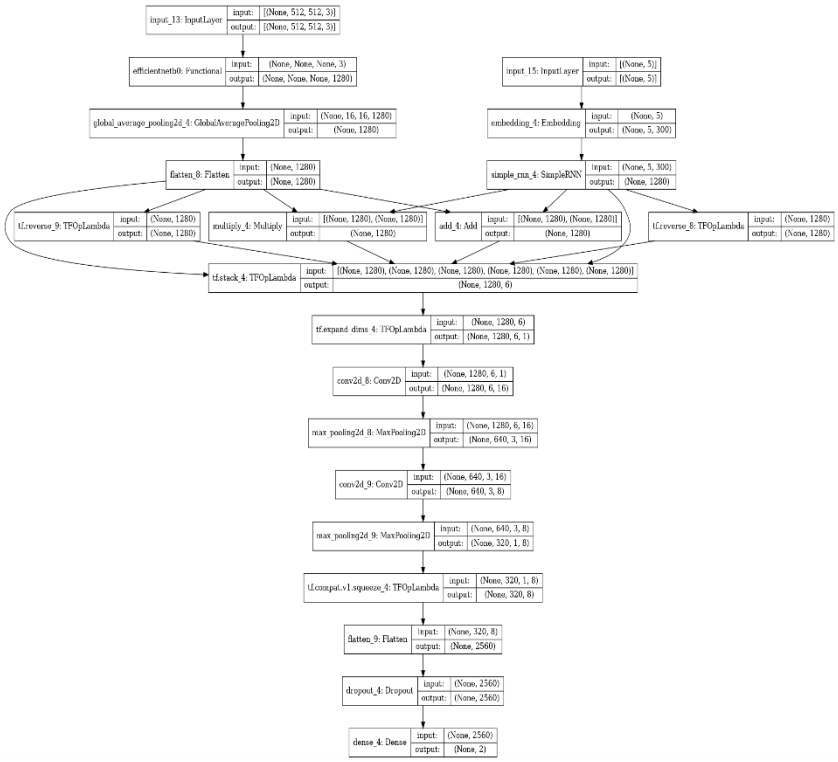
62 pav. LSTM ir ResNet50v2 mokymosi grafikas (taikant sąsūką požymiams)



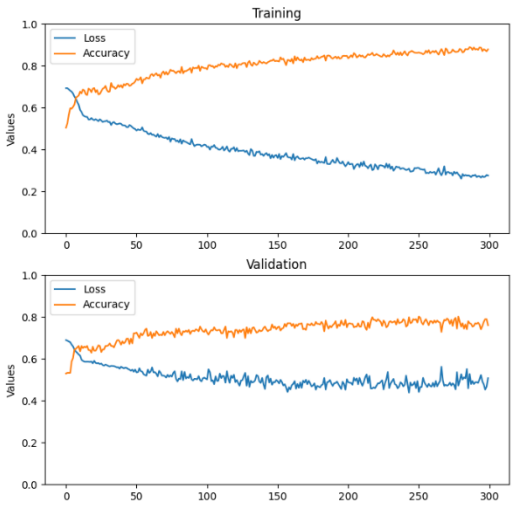
63 pav. LSTM ir EfficientNetB0 architektūra (taikant sąsūką požymiams)



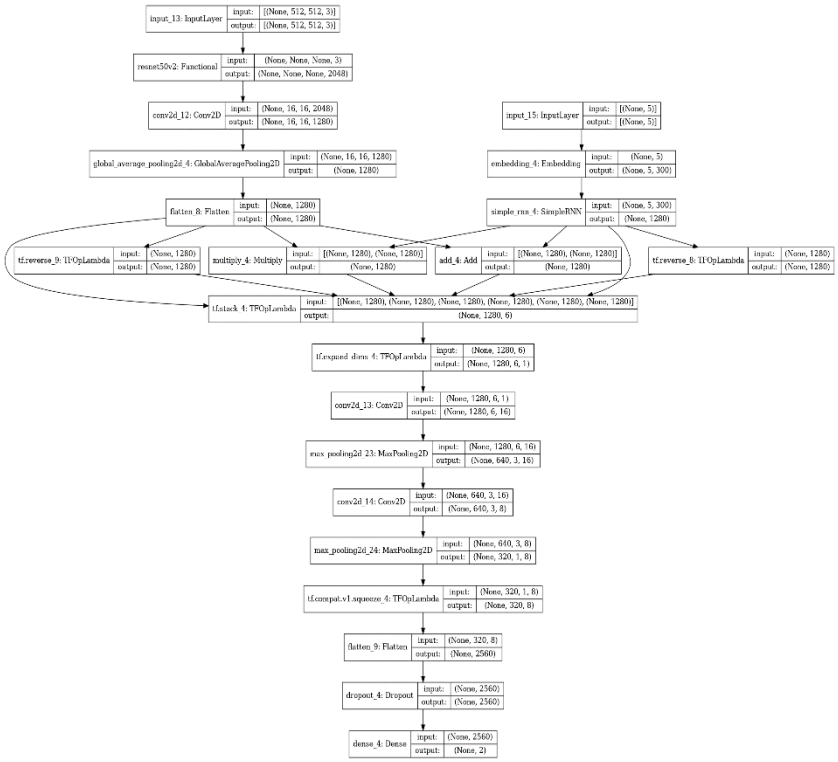
64 pav. LSTM ir EfficientNetB0 mokymosi grafikas (taikant sąsūką požymiams)



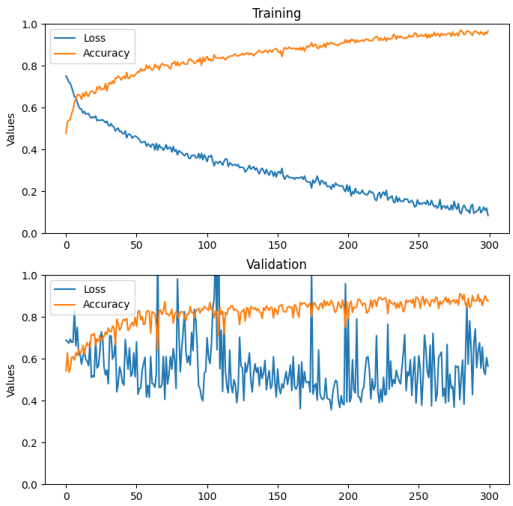
65 pav. RNN ir EfficientNetB0 architektūra (taikant sąsūką požymiams)



66 pav. RNN ir EfficientNetB0 mokymosi grafikas (taikant sąsūką požymiams)



67 pav. RNN ir ResNet50v2 architektūra (taikant sąsūką požymiams)



68 pav. RNN ir ResNet50v2 mokymosi grafikas (taikant sąsūką požymiams)

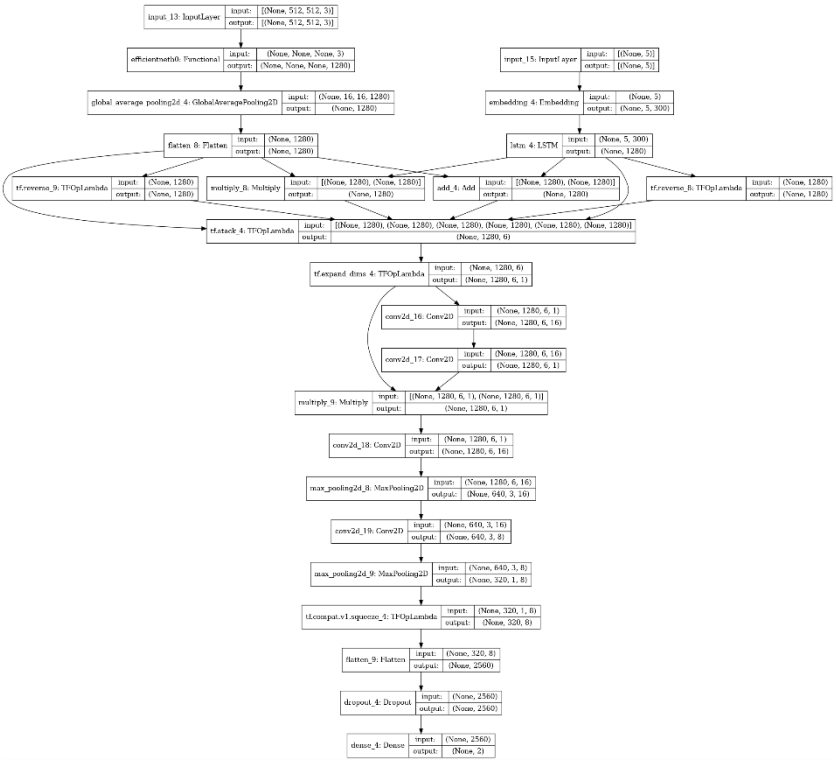
13 lentelė. Modelių tikslumai naudojantis 2.4.2 skyrelyje aprašyta metodologija

Klausimo kategorija	**LSTM ir EfficientNetB0 tikslumas**	**RNN ir EfficientNetB0 tikslumas**	**LSTM ir ResNet50V2 tikslumas**	**RNN ir ResNet50V2 tikslumas**
Vaizdo normalumas	80,64%	76,34%	80,64%	79,56%
Anomalijos vaizde	68,62%	72,54%	76,47%	72,54%
Vaizdo tipas	91,33%	91,33%	92,91%	90,15%
Kontrasto tirpalo panaudojimas	74,31%	76,14%	76,60%	73,39%
Bendras vidurkis	82,81%	83,45%	84,80%	81,91%

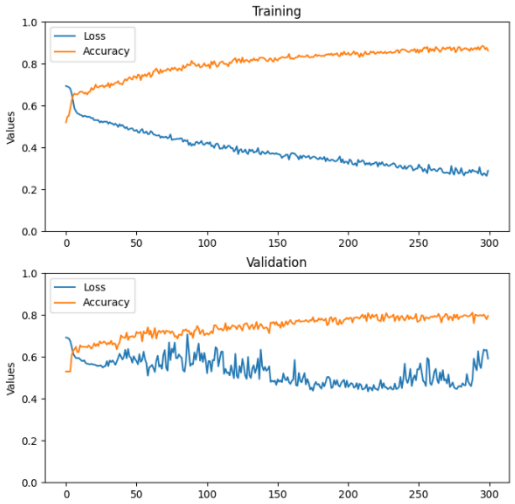
Iš 13 lentelės, galima matyti, jog aukščiausias bendro vidurkio tikslumas buvo pasiektas naudojantis LSTM ir ResNet50v2 modeliu. Lyginant šio požymių suliejimo būdą su ankstesniais darbe taikytais metodais, galima formuluoti išvada, jog ši metodologija lenkia tiek požymių konkatenavimą, tiek sumavimą, kadangi visų architektūrų tikslumai yra aukštesni.

3.6. Darbe taikytos architektūros požymius sudedant į matricą, įvedant jai svorius ir taikant sąsūkos operaciją

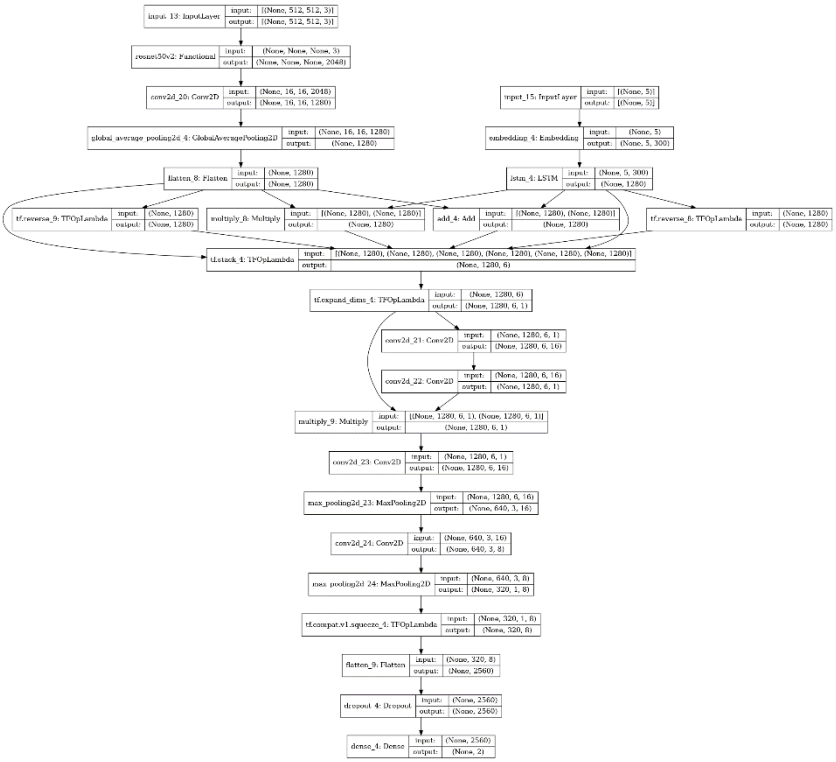
Šiame skyriuje taip pat, kaip ir praeituose pateiksime kiekvienos architektūros grafinius pavyzdžius bei jų mokymosi grafikus. Architektūrų plačiau nekomentuosime, kadangi šio modelio idėja yra aprašyta 2.4.3 skyrelyje. Taip pat, pateikiame žemiau mokymosi grafikus, iš kurių galima matyti, jog ResNet50V2 tinklo mokymasis turėjo daugiau triukšmo nei EfficientNetB0. Skyrelio pabaigoje, 14 lentelėje, pateikiame gautų tikslumų įverčius, o mokymo parametrai išlieka tokie patys, kaip ir ankstesniuose skyreliuose.



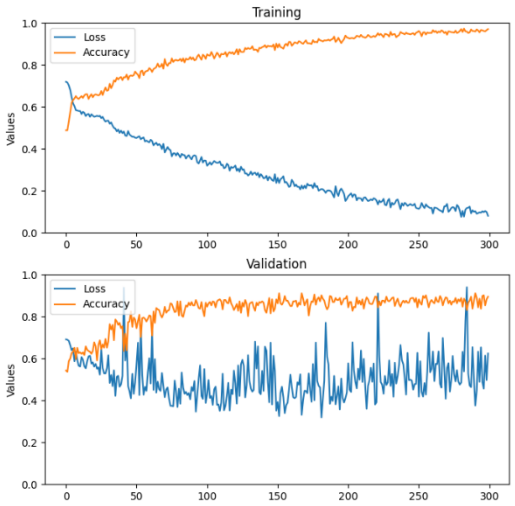
69 pav. LSTM ir EfficientNetB0 architektūra (taikant sąsūką bei svorius požymiams)



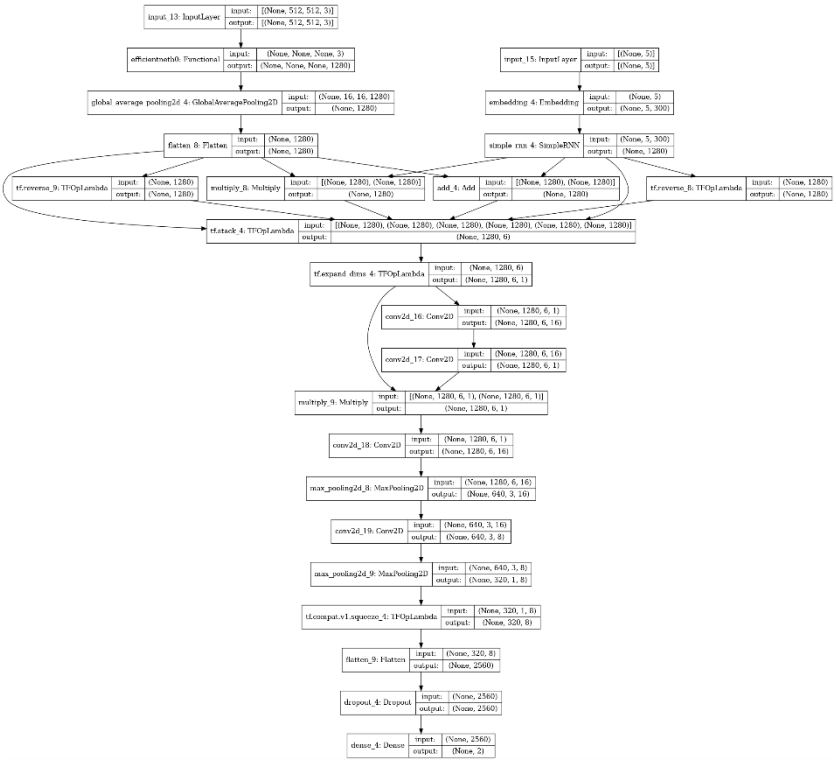
70 pav. LSTM ir EfficientNetB0 mokymosi grafikas (taikant sąsūką bei svorius požymiams)



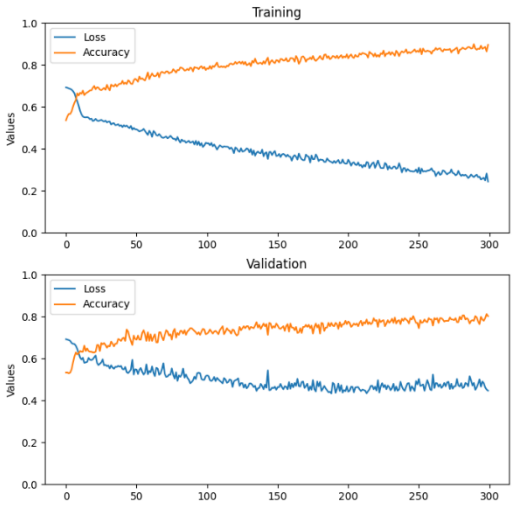
71 pav. LSTM ir ResNet50v2 architektūra (taikant sąsūką bei svorius požymiams)



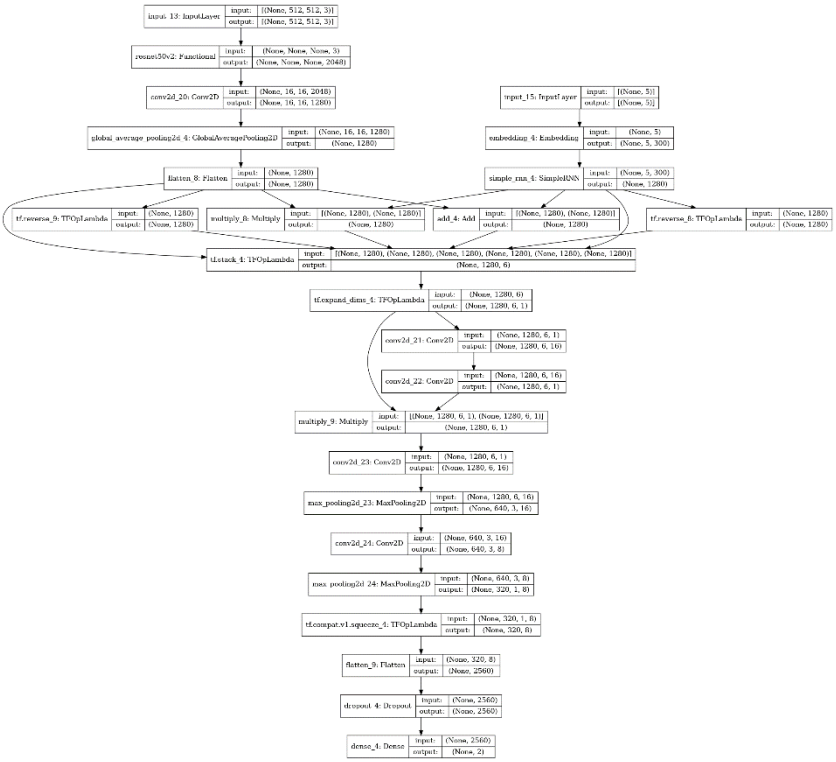
72 pav. LSTM ir ResNet50v2 mokymosi grafikas (taikant sąsūką bei svorius požymiams)



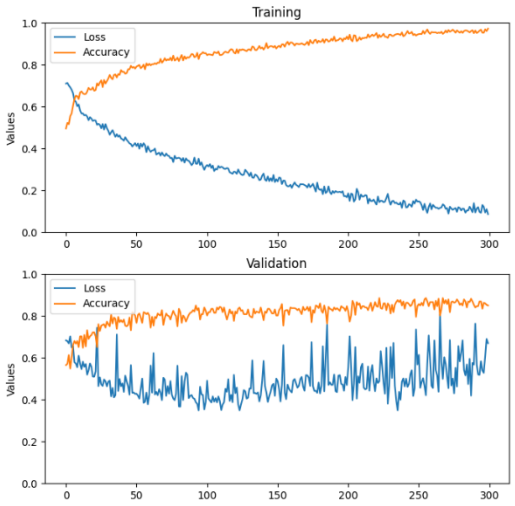
73 pav. RNN ir EfficientNetB0 architektūra (taikant sąsūką bei svorius požymiams)



74 pav. RNN ir EfficientNetB0 mokymosi grafikas (taikant sąsūką bei svorius požymiams)



75 pav. RNN ir ResNet50v2 architektūra (taikant sąsūką bei svorius požymiams)



76 pav. RNN ir ResNet50v2 mokymosi grafikas (taikant sąsūką bei svorius požymiams)

14 lentelė. Modelių tikslumai naudojantis 2.4.3 skyrelyje aprašyta metodologija

Klausimo kategorija	**LSTM ir EfficientNetB0 tikslumas**	**RNN ir EfficientNetB0 tikslumas**	**LSTM ir ResNet50V2 tikslumas**	**RNN ir ResNet50V2 tikslumas**
Vaizdo normalumas	78,49%	75,26%	79,56%	76,34%
Anomalijos vaizde	72,54%	76,47%	70,58%	76,47%
Vaizdo tipas	91,86%	91,07%	91,07%	92,25%
Kontrasto tirpalo panaudojimas	75,68%	73,70%	76,14%	73,08%
Bendras vidurkis	83,65%	82,36%	83,45%	82,74%

Iš 14 lentelės galima matyti, jog LSTM ir EfficientNetB0 bei RNN ir ResNet50v2 modeliai turėjo geresnius rezultatus, kuomet buvo pridedami svoriai *M* matricai, o, tuo tarpu, RNN ir EfficientNetB0 bei LSTM ir ResNet50v2 modeliams geriau sekėsi klasifikuoti rezultatus be papildomų svorių. Kaip bebūtų, negalima teigti, jog ši architektūrinė idėja pasiteisino, nes geriausi darbe rezultatai yra pasiekti netaikant papildomų svorių *M* matricai.

Išvados

- Išanalizavus literatūros medžiagą, pastebėjome, jog vaizdais grįsta vaizdo analitika sulaukia vis didesnio tyrėjų susidomėjimo. Iš nagrinėtų straipsnių, galima matyti, jog autoriai pasitelkia panašius modelius (paprastai, giliaisiais tinklais grįstus) vaizdo bei teksto požymiams apskaičiuoti, bet šių požymių suliejimo metodikos pasižymi didele įvairove: yra taikomos sudėties, sandaugos, prijungimo operacijos, dėmesio sutelkimo mechanizmai bei kiti specifiniai metodai.

- Darbe pasiūlytas kalbos ir vaizdų požymių agregavimo būdas leido gauti geresnius rezultatus, lyginant su tradiciškai mokslinėje literatūroje naudojamais sumavimo, ar konkatenacijos būdais. Naudojantis ResNet50V2, LSTM tinklais ir šių modelių požymių vektorius apdorojant suforamvus matricą bei pritaikius sąsūkos operacijas buvo pasiekiamas 84,40 % vidutinis tikslumas, kai, tuo tarpu, sumavimo atveju geriausias modelis remėsi EfficientNetB0 ir LSTM tinklais bei pasiekė 81.91 % vidutinį tikslumą, o požymius apjungiant konkatenacijos operacija buvo gaunamas 82,24 % vidutinis tikslumas pasitelkiant ResNet50V2 ir LSTM tinklus.

- Apjungtų požymių matricai dar papildomai įvedant svorius, rezultatai nepagerėjo. Šios architektūros geriausias modelis rėmėsi LSTM tinklu bei EfficientNetB0 sąsūkų neuroniniu tinklu ir sugebėjo pasiekti 83,65 % vidutinį tikslumą.

- Prasčiausiai darbe nagrinėjamą problematiką sprendė architektūra, kuri rėmėsi požymių sutelkimu sudedant teksto bei vaizdo vektorius. Šios idėjos geriausias rezultatas buvo pasiektas remiantis LSTM ir EfficientNetB0 architektūromis ir buvo lygūs 81,91 %. Šis agregavimo būdas greičiausiai pernelyg sumažina informacijos kiekį esantį atskiruose kalbos ir vaizdo požymių vektoriuose.

Literatūros šaltiniai

1. SHARMA, D. - PURUSHOTHAM, S. - REDDY, C. MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. In . 2022. [žiūrėta 2022-02-15]

2. ANTOL, S. - AGRAWAL, A. - LU, J. - MITCHELL, M. - BATRA, D. - ZITNICK, C. - PARIKH, D. VQA: Visual Question Answering. In Openaccess.thecvf.com [interaktyvus]. 2022. [žiūrėta 2022-02-15]. Prieiga per internetą: <https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html>.

3. ZHOU, B. - TIAN, Y. - SUKHBAATAR, S. - SZLAM, A. - FERGUS, R. Simple Baseline for Visual Question Answering. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-02-27]. Prieiga per internetą: <https://arxiv.org/abs/1512.02167>.

4. SHIH, K. - SINGH, S. - HOIEM, D. Where to Look: Focus Regions for Visual Question Answering. In Openaccess.thecvf.com [interaktyvus]. 2022. [žiūrėta 2022-02-15]. Prieiga per internetą: <https://openaccess.thecvf.com/content_cvpr_2016/html/Shih_Where_to_Look_CVPR_2016_paper.html>.

5. ILIEVSKI, I. - YAN, S. - FENG, J. A Focused Dynamic Attention Model for Visual Question Answering. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-02-16]. Prieiga per internetą: <https://arxiv.org/abs/1604.01485>.

6. PONT-TUSET, J. - ARBELAEZ, P. - T.BARRON, J. - MARQUES, F. - MALIK, J. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation. In IEEE Transactions on Pattern Analysis and Machine Intelligence . 2017. Vol. 39, no. 1, p. 128-140. [žiūrėta 2022-02-16]

7. HE, K. - ZHANG, X. - REN, S. - SUN, J. Deep Residual Learning for Image Recognition. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-02-16]. Prieiga per internetą: <https://arxiv.org/abs/1512.03385>.

8. TENEY, D. - ANDERSON, P. - HE, X. - VAN DEN HENGEL, A. Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge. In Openaccess.thecvf.com [interaktyvus]. 2022. [žiūrėta 2022-02-16]. Prieiga per internetą: <https://openaccess.thecvf.com/content_cvpr_2018/html/Teney_Tips_and_Tricks_CVPR_2018_paper.html>.

9. KORNUTA, T. - RAJAN, D. - SHIVADE, C. - ASSEMAN, A. - OZCAN, A. Leveraging Medical Visual Question Answering with Supporting Facts. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-02-16]. Prieiga per internetą: <https://arxiv.org/abs/1905.12008>.

10. YAN, X. - LI, L. - XIE, C. - XIAO, J. - GU, L. Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domai. In [interaktyvus]. 2019. [žiūrėta 2022-02- 27]. Prieiga per internetą: <https://www.semanticscholar.org/paper/Zhejiang-University-at-ImageCLEF-2019-Visual-in-the-Yan-i/c7dbeea4258174e0eece7239e24f7bd909f2d606>.

11. ABACHA, A. - RAJARAMAN, S. NLM at ImageCLEF 2018 Visual question Answering in the Medical Domain. In [interaktyvus]. 2018. [žiūrėta 2022-02-17]. Prieiga per internetą: <https://www.researchgate.net/publication/328491475_NLM_at_ImageCLEF_2018_Visual_Question_Answering_in_the_Medical_Domain/citation/download>.

12. FUKUI, A. - PARK, D. - YANG, D. - ROHRBACH, A. - DARRELL, T. - ROHRBACH, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-02-17]. Prieiga per internetą: <https://arxiv.org/abs/1606.01847>.

13. SARROUTI, M. NLM at VQA-Med 2020: Visual Question Answering and Generation in the Medical Domain. In [interaktyvus]. 2020. [žiūrėta 2022-02-17]. Prieiga per internetą: <https://www.semanticscholar.org/paper/NLM-at-VQA-Med-2020%3A-Visual-Question-Answering-and-Sarrouti/791c950feebeae3130e04fa008419ac99afb41f0>.

14. MIKOLOV, T. - CHEN, K. - CORRADO, G. - DEAN, J. Efficient Estimation of Word Representations in Vector Space. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-02]. Prieiga per internetą: <https://arxiv.org/abs/1301.3781>.

15. BOJANOWSKI, P. - GRAVE, E. - JOULIN, A. - MIKOLOV, T. Enriching Word Vectors with Subword Information. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-02]. Prieiga per internetą: <https://arxiv.org/abs/1607.04606>.

16. LIU, P. - QIU, X. - HUANG, X. Recurrent Neural Network for Text Classification with Multi-Task Learning. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-02]. Prieiga per internetą: <https://arxiv.org/abs/1605.05101>.

17. WANG, H. - SONG, Y. - TANG, S. LSTM-based Flow Prediction. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-02]. Prieiga per internetą: <https://arxiv.org/abs/1908.03571>.

18. Peng J., Kimmig A., Wang J., Liu X., Niu Z., Ovtcharova J., Dual-stage attention-based long-short-term memory neural networks for energy demand prediction, Energy and Buildings, Volume 249, 2021. [žiūrėta 2022-03-02]. [interaktyvus]. Prieiga per internetą: https://www.sciencedirect.com/science/article/pii/S0378778821004953

19. SONG, S. - HUANG, H. - RUAN, T. Abstractive text summarization using LSTM-CNN based deep learning. In Multimedia Tools and Applications . 2018. Vol. 78, no. 1, p. 857-875. [žiūrėta 2022-03-03]

20. HE, K. - ZHANG, X. - REN, S. - SUN, J. Deep Residual Learning for Image Recognition. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-03]. Prieiga per internetą: <https://arxiv.org/abs/1512.03385>.

21. HE, K. - ZHANG, X. - REN, S. - SUN, J. Identity Mappings in Deep Residual Networks. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-03]. Prieiga per internetą: <https://arxiv.org/abs/1603.05027>.

22. REZENDE, E. - RUPPERT, G. - CARVALHO, T. - RAMOS, F. - GEUS, P. Malicious Software Classification Using Transfer Learning of ResNet-50 Deep Neural Network. In

Ieeexplore.ieee.org [interaktyvus]. 2022. [žiūrėta 2022-03-03]. Prieiga per internetą: <https://ieeexplore.ieee.org/document/8260773>.

23. HE, K. - ZHANG, X. - REN, S. - SUN, J. Identity Mappings in Deep Residual Networks. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-04]. Prieiga per internetą: <https://arxiv.org/abs/1603.05027>.

24. TAN, M. - LE, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-04]. Prieiga per internetą: <https://arxiv.org/abs/1905.11946>.

25. SANDLER, M. - HOWARD, A. - ZHU, M. - ZHMOGINOV, A. - CHEN, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-04]. Prieiga per internetą: <https://arxiv.org/abs/1801.04381v4>.

26. HOWARD, A. - ZHU, M. - CHEN, B. - KALENICHENKO, D. - WANG, W. - WEYAND, T. - ANDREETTO, M. - ADAM, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-05]. Prieiga per internetą: <https://arxiv.org/abs/1704.04861>.

27. GANG, S. CHUNG, D. Character Recognition of Components Mounted on Printed Circuit Board Using Deep Learning. 2021. [žiūrėta 2022-03-04]. Prieiga per internetą: https://www.researchgate.net/publication/351057828_Character_Recognition_of_Components_Mounted_on_Printed_Circuit_Board_Using_Deep_Learning

28. RAMACHANDRAN, P. - ZOPH, B. - LE, Q. Swish: a Self-Gated Activation Function. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-05]. Prieiga per internetą: <https://arxiv.org/abs/1710.05941v1?source=post_page---------->.

29. HU, J. - SHEN, L. - ALBANIE, S. - SUN, G. - WU, E. Squeeze-and-Excitation Networks. In arXiv.org [interaktyvus]. 2022. [žiūrėta 2022-03-05]. Prieiga per internetą: <https//arxiv.org/abs/1709.01507>.

30. Visual Question Answering in the Medical Domain VQA-Med 2019 [interaktyvus]. 2019. [žiūrėta 2022-02-18]. Prieiga per internetą: https://github.com/abachaa/VQA-Med-2019

31. FastText. [interaktyvus]. 2019. [žiūrėta 2022-02-18]. Prieiga per internetą: https://fasttext.cc/