

Հ.Հ. ԽԱՉԱՏՐՅԱՆ, Դ.Ս. ՔԱՐԱՍՅԱՆ, Տ.Ս. ՔԱՐԱՍՅԱՆ

**ՌԵԿՈՒՐԵՆՏ ՆԵՅՐՈՆԱՅԻՆ ՑԱՆՑԵՐԻ ՈՐՈՇՈՒՄՆԵՐԻ
ԲԱՑԱՏՐՈՒԹՅՈՒՆԸ ՊԱՏԿԵՐԱՎՈՐՄԱՆ ՄԻՋՈՑՈՎ**

Ժամանակակից նեյրոնային ցանցերն ունակ են մոտարկելու բարդ ֆունկցիաներ, ինչի հետևանքով դրանք բարձր արդյունավետությամբ աշխատում են պատկերների ճանաչման, բնական լեզուների մշակման (*NLP*) մի շարք խնդիրների լուծման դեպքում: Չնայած դրան, այնքան էլ պարզ չէ, թե ինչի հիման վրա են նեյրոնային ցանցերը կատարում իրենց կանխատեսումները: Մասնավորապես, խնդիր է դրվում՝ հասկանալու, թե մուտքի որ նկարագրերին (*features*) նայելով է ցանցը կայացնում իր որոշումը: Բժշկության մեջ ցանցերի կանխատեսումները ոչ միայն պետք է լինեն ճշգրիտ, այլև պետք է ունենան հիմնավորումներ: Օրինակ, երբ նեյրոնային ցանցը կանխատեսել է հիվանդության առկայությունը, անհրաժեշտ է հասկանալ, թե մուտքային որ տվյալներն են մեծ դեր խաղացել այդ որոշման մեջ:

Աշխատանքում փորձ է արվել լուսաբանել նեյրոնային ցանցերի աշխատանքը կամ, այլ կերպ ասած, պատկերավորել դրանց աշխատանքը: Որպես հետազոտության առարկա վերցվել է MIMIC-III բժշկական տվյալների բազայով աշխատող ռեկուրենտ նեյրոնային ցանց, որը մուտքում ստանում է վերականգնման քաժամանակում գտնվող հիվանդի կլինիկական տվյալների ժամանակային շարքը և փորձում գուշակել հիվանդի՝ վերականգնման քաժամանակում մահվան ելքի հավանականությունը: Աշխատանքում հաջողվել է պատկերների ճանաչման և բնական լեզվի մշակման խնդիրների համար մշակված երկու մեթոդ հարմարեցնել կլինիկական տվյալների վրա աշխատող նեյրոնային ցանցերին և պատկերավորել այդ ցանցերի աշխատանքը:

Առանցքային բառեր: ռեկուրենտ նեյրոնային ցանցեր, գրադիենտային հիմքով պատկերավորում, կարևորության հետընթաց մակարդակային շարժմամբ պատկերավորում, MIMIC-III, ժամանակային շարքեր:

Ներածություն: Կանխարգելվող հիվանդությունների վաղ հայտնաբերումը կարևոր դեր է խաղում հիվանդության արդյունավետ բուժման կազմակերպման, առողջական խնամքի ռեսուրսների ճիշտ օգտագործման գործընթացներում: Հիվանդների առողջական ցուցանիշների (*electronic health care records or EHRs*) թվայնացումը և հավաքագրումը հնարավորություն է ստեղծում՝ օգտագործելու մեքենայական ուսուցման ներուժը և ստեղծելու համակարգեր, որոնք թույլ կտան անընդհատ հետևել հիվանդի վիճակին, կանխատեսել թաքնված հիվանդություններ:

ները, ինչպես նաև կանխատեսել հիվանդի վիճակը ապագայում՝ հիմք ընդունելով նախորդ ժամերին չափած առողջական տվյալները:

Ներկայումս *EHR* համակարգերը նախատեսված են տվյալների ճշգրիտ պահպանման և հիվանդի վիճակը ժամանակի ընթացքում հետևելու համար: Դա թույլ է տալիս խուսափել նախկինում արված բժշկական հետազոտությունների անհարկի կրկնօրինակումներից, բժշկական քարտերի կորստից և ունենալով տվյալների միայն մեկ կենտրոնացված միասնական աղբյուր՝ ավելի ճշգրիտ հետևել հիվանդի վիճակի դինամիկային:

Աշխատանքում փորձել ենք պատկերավորել [1]-ում նկարագրված ցանցերից մեկը, որը, մուտքում ստանալով վերակենդանացման բաժանմունքում գտնվող հիվանդի 48-ժամյա առողջական ցուցանիշների ժամանակային շարքը, փորձում է գուշակել մահվան ելքի հավանականությունը:

Պատկերավորումը մուտքային տվյալների կարևորության քարտեզի ստացումն է: Պատկերավորման համար օգտագործվել են երկու տեսակ գործիքակազմ՝ հիմնված մուտքային տվյալների գրադիենտի (*gradient based visualization*) և կարևորության արժեքի հետընթաց մակարդակային շարժման վրա (*Layer wise relevance propagation*): Համեմատենք և մեկնաբանենք ստացված կարևորության քարտեզները:

Աշխատանքը բաղկացած է երեք մասից, որի առաջին մասում ներկայացված են ներդրումային ցանցերի պատկերավորման վերոնշյալ մեթոդները: Երկրորդ մասում ներկայացված են *MIMIC-III* տվյալների հենքը, խնդրի մաթեմատիկական ձևակերպումը և այն մոդելը, որը վարժեցվել է այդ տվյալների վրա: Երրորդ մասում դիտարկվում են արդյունքները և մեկնաբանվում կարևորության քարտեզները:

Պատկերավորման մեթոդները պատկերների ճանաչման մեջ. [2] աշխատանքում դիտարկվել է հետևյալ խնդիրը. տրված վարժեցված փաթեթային ցանցի ու տրված դասի համար անհրաժեշտ է նկարների տարածությունից գտնել այնպիսի նկար, որը կմաքսիմալացնի նպատակային դասի արժեքը: Այլ կերպ՝ պետք է լուծել հետևյալ օպտիմալացման խնդիրը. $\operatorname{argmax}_I (S_c(I) - \lambda \|I\|_2^2)$, որտեղ $S_c(I)$ -ին նպատակային դասի արժեքն է, որը ստացվում է ցանցի կանխատեսող շերտից՝ I -մուտքային նկարի համար: Անհրաժեշտ է գտնել այնպիսի $L2$ կարգավորված նկար, որի դեպքում $S_c(I)$ - ին կլինի առավելագույնը: Տեղային օպտիմալ I հնարավոր է գտնել հետընթաց-տարածման (*back-propagation*) ալգորիթմով: Օպտիմալացումը սկսվում է 0-ներով սկզբնարժեքավորված նկարից, որը յուրաքանչյուր հետընթաց-տարածման քայլի ժամանակ ենթարկվում է փոփոխության՝ $I_t = I_{t-1} + \alpha \times \operatorname{grad}_I (S_c(I) - \lambda \|I\|_2^2)$: Վերջնական արդյունքին գումարվում է վարժեցնող բազմության (*training set*) միջինը:

[3]-ում խնդիր է դրված պատկերավորել ոչ թե դասը, այլ տրված նկարը: Տրված I_0 նկարի և c դասի համար հարկավոր է ռանգավորել I_0 -ի պիքսելները՝

ըստ նրանց կարևորության, $S_c(I_0)$ -ն հաշվելիս: $S_c(I)$ -ն մոտարկելով նրա առաջին կարգի Թեյլորի շարքի ֆունկցիայով I_0 նկարի մոտակայքում՝ $S_c(I) \approx w_c^T I + b_c$, ստանում ենք $w \approx \nabla_I(S_c)|_{I_0}$: $|w|$ -ն կարագրում է այն, թե ինչքանով պետք է փոփոխել պիքսելները՝ նպատակային արժեքը մաքսիմալացնելու համար:

[4]-ում առաջարկվել է պատկերավորման նոր մեթոդ՝ կարևորության հետընթաց մակարդակային շարժում (*Layer wise Relevance Propagation (LRP)*): Այն սահմանվում է որպես սահմանափակումների ամբողջություն և կամայական լուծում, որը բավարարում է այդ սահմանափակումները, համարվում է պատկերավորման այս մեթոդի մասնավոր դեպքը: Հիմնական նպատակն է՝ հասկանալ մուտքային տվյալների դերը $f(x)$ ցանցային ֆունկցիայի կանխատեսման մեջ: Դիտարկենք ֆիդբոքարդ ցանց՝ f , որի l -երորդ շերտի միավորները նշանակենք $\{z_d^{(l)}\}_{d=1}^{V(l)}$, որտեղ $V(l)$ -ը l -երորդ շերտի չափողականությունն է: $\{R_d^{(l)}\}_{d=1}^{V(l)}$ -ով նշանակենք l -երորդ շերտի կարևորության վեկտորը: Ունենալով $R_d^{(l+1)}$ կարևորության վեկտորը $(l+1)$ -րորդ շերտի համար՝ անհրաժեշտ է գտնել $R_d^{(l)}$ կարևորության վեկտորը l -երորդ շերտի համար այնպես, որ տեղի ունենան հետևյալ հավասարությունները.

$$f(x) = \dots = \sum_d R_d^{(l+1)} = \sum_d R_d^{(l)} = \dots = \sum_d R_d^{(1)}:$$

Ըստ դրանց՝ կարևորության արժեքը մի շերտից մյուսին անցնելիս պահպանվում է: Այսպիսի վերլուծությունը միակը չէ և չի երաշխավորում պատկերավորման լավ արդյունքներ: $R_{i \leftarrow j}^{(l,l+1)}$ -ն մեկնաբանվում է որպես հաղորդագրություն՝ i -երորդ միավորից դեպի j -երորդ միավորը: Միավորի կարևորությունը սահմանվում է որպես մուտքային հաղորդագրությունների գումար՝

$$R_i^{(l)} = \sum_{i \in \text{input}(k)} R_{i \leftarrow k}^{(l,l+1)}:$$

$R_{i \leftarrow k}^{(l,l+1)'}$ հաղորդագրությունը վերցրված է համեմատական՝ նրա ունեցած տեղային ազդեցությանը առաջընթաց շարժման ժամանակ $a_k^{(l+1)}$ միավորի հաշվարկման մեջ՝

$$R_{i \leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \frac{a_i w_{ik}}{\sum_h a_h w_{hk}}.$$

Պատկերավորման մեթոդները բնական լեզուների մշակման մեջ. Բնական լեզուների մշակման մեջ ներդրման մոդելները հիմնականում որպես մուտք ստանում են բառերի ներդրված վեկտորներ: Ներդրված վեկտորները ներկայացնում են բառի իմաստը վեկտորի միջոցով: Բազմաթիվ մոդելներով (օրինակ՝ Word2vec, Glove, StarSpace,...) կառուցվում են այդ վեկտորները, որոնք միավորող փիլիսո-

փայտությունն այն է, որ բառի իմաստը որոշվում է կոնտեքստից՝ այն բառերից, որոնց հետ միաժամանակ հանդիպում է տեքստում: Ներդրված վեկտորների մեջ ամփոփված է բառի իմաստը, սակայն կարիք կա ավելի խոր հասկանալ, թե նեյրոնային ցանցերը ինչպես են հասկանում այդ վեկտորների համադրումը:

Երբ ցանցի մուտքում ներդրված վեկտորներ են, ապա յուրաքանչյուր բառի կարևորությունը սահմանենք որպես այդ բառի ներդրված վեկտորի բոլոր տարրերի կարևորության արժեքների գումար՝

$$R^{(word)} = \sum_d R_d^{(word)}:$$

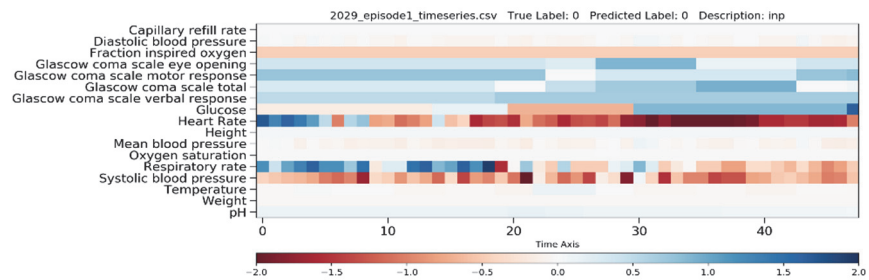
[5]-ում դիտարկված է մեկնաբանությունների տարբերակման խնդիրը, որտեղ անհրաժեշտ է տարբերակել դրական մեկնաբանությունները բացասականներից և պատկերավորել այդ խնդիրը լուծող ռեկուրենտ ցանցերի աշխատանքը: Տրված E ներդրված վեկտորների և c դասի համար նեյրոնային ցանցը (E, c) գույզին համապատասխանության մեջ է դնում $S_c(E)$ արժեքը: Նպատակն է պարզել, թե E -ի որ վեկտորներն են ամենաշատը ազդում $S_c(E)$ -ի վրա: Քանի որ խոր ցանցերում $S_c(E)$ -ն ոչ գծային է, այն կարելի է մոտարկել Թեյլորի առաջին կարգի վերլուծությամբ, որտեղից $w(e) = \frac{\partial S}{\partial e}$, որտեղ $e \in E$: Կարևորության քարտեզն էլ կլինի $|w(e)|$ -ն, որը նկարագրում է, թե ներդրված վեկտորների ամեն մի արժեք ինչպես է ազդում ցանցի վերջնական որոշման վրա:

Կարևորության հետընթաց մակարդակային շարժման մեթոդը միանգամից կիրառելի չէ ռեկուրենտ ցանցերը պատկերավորելիս, քանի որ ռեկուրենտ ցանցերի ճարտարապետության մեջ կան տարր առ տարր բազմապատկման գործողություններ: [6]-ում ներկայացված է վերոնշյալ ալգորիթմի ընդհանրացումը ռեկուրենտ ցանցերի համար: Դիցուք z_j -ն վերին շերտի որևէ նեյրոն է, որի արժեքը առաջընթաց շարժման ժամանակ հաշվարկվում է $z_j = z_g \cdot z_s$ բանաձևով, որտեղ z_g -ն համապատասխանում է հանգույցի արժեքին, իսկ z_s -ը կարճ կամ երկար կախվածություններին համապատասխանող ինֆորմացիայի արժեքն է: Այսպիսի արտադրյալային կապերը բնորոշ են LSTM և GRU տեսակի ցանցերին, որոնք ամենահաճախ օգտագործվող ռեկուրենտ ցանցերի տեսակներն են: Ենթադրենք R_j -ն z_j նեյրոնի կարևորության արժեքն է: Այս դեպքում ներքևի շերտի կարևորության արժեքը սահմանվում է հետևյալ կերպ՝ $R_g = 0$ և $R_s = R_j$:

MIMIC-III տվյալների հենքի բնութագիրը և խնդրի դրվածքը. *MIMIC-III (Medical Information Mart for Intensive Care III* [7])-ը ռեյացիոն, անվճար-հասանելի տվյալների հենք է, որը պարունակում է 40000 հիվանդների առողջական տվյալներ, ովքեր 2001-2012 թվականներին եղել են *Beth Israel Deaconess* բժշկական կենտրոնի վերակենդանացման բաժանմունքում: Տվյալների բազան ներա-

ռում է տեղեկություններ, ինչպիսիք են դեմոգրաֆիկ տվյալները, կատարված կենսական նշանների չափումները (1 ժամ միջոդադարով), լաբորատոր հետազոտության արդյունքները, կատարված բուժգործողությունները, ընդունած դեղերը և այլ տեղեկություններ:

Ստորև ներկայացված է տվյալների արդեն իսկ նախամշակված [1] նմուշի պատկերը (նկ. 1), որտեղ գունային ինտենսիվությունը փոխվում է [-2, 2] միջակայքում՝ -2-ը ներկված է մուգ դարչնագույն, իսկ 2-ը՝ կապույտ: Նկարի ձախ ու աջ մասերում մատնանշված են կենսական նշանների չափումների բժշկական տերմինները, օրինակ՝ *Fraction inspired oxygen* կամ *Mask for Glucose* (աղ. 1), իսկ մյուս առանցքը (*Time Axis*) բնութագրում է ժամանակը՝ 0-ից մինչև 47: *Mask for X* առանցքային բառով սկսվող նկարագրերն ունեն հատուկ նշանակություն. դրանք ցույց են տալիս, թե որ պահին է բժիշկը չափել *X* տեսակի նկարագիրը: Նկարագրերում կան ինչպես դիսկրետ, այնպես էլ անընդհատ մեծություններ: Ընդ որում՝ տվյալները մշակվել են հետևյալ օրինաչափությամբ. բժշկի մի չափումից մինչև մյուս չափումն ընկած ժամանակահատվածում նկարագրի արժեքը կրկնվում է ընդհուպ մինչև նոր չափման արժեքի գրանցումը:

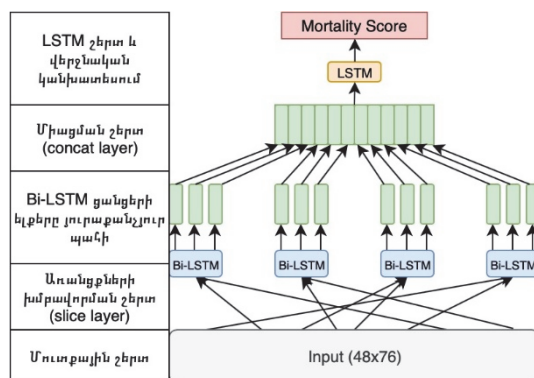


Նկ. 1. MIMIC տվյալների հենքի նախամշակված նմուշի պատկերը (նայել գունավոր տարբերակով): Պարզության համար *Mask* տեսակի նկարագրերը բաց են թողնվել

Աշխատանք [1]-ում ձևակերպված են չորս հիմնական խնդիրներ՝ *MIMIC-III* տվյալների հենքի հիման վրա: Սակայն այս աշխատանքում ուշադրություն կդարձնենք միայն մեկ՝ մարդու մահանալու հավանականությունը մոտարկելու խնդրի վրա (in-hospital mortality): Ներկայացված են մի շարք նեյրոնային մոդելներ, որոնք հասցեավորված են՝ մոտարկելու մարդու մահանալու հավանականությունը: Ներքևում համառոտ նկարագրված է այդ մոդելներից մեկը (Channel-wise LSTM), որի վրա էլ կիրառվել են պատկերավորման մեթոդները: Ի տարբերություն ստանդարտ LSTM-ի, որը միանգամից աշխատում է հաջորդական մուտքային տվյալներով, Channel-wise LSTM-ը իրարից անկախ մշակում է մուտքային տվյալների տարբեր առանցքները երկկողմանի LSTM-ներով, այնուհետև բոլոր LSTM-ների ելքերը միացվում են իրար (concatenation) և որպես մուտք փոխանցվում մեկ այլ LSTM ցանցի (նկ. 2):

Փոփոխականների նկարագրությունը

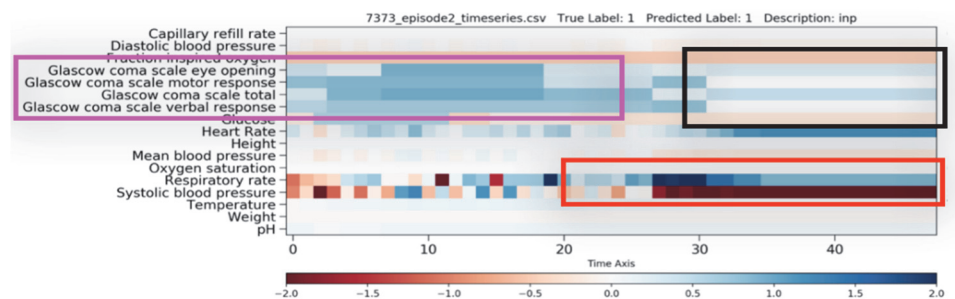
Գենասական չափումները	Նորմը	Տեսակը
Capillary refill rate	0.0	categorical
Diastolic blood pressure	59.0	continuous
Fraction inspired oxygen	0.21	continuous
Glasgow coma scale eye opening	4 spontaneously	categorical
Glasgow coma scale motor response	6 obeys commands	categorical
Glasgow coma scale total	15	categorical
Glasgow coma scale verbal response	5 oriented	categorical
Glucose	128.0	continuous
Heart Rate	86	continuous
Height	170.0	continuous
Mean blood pressure	77.0	continuous
Oxygen saturation	98.0	continuous
Respiratory rate	19	continuous
Systolic blood pressure	118.0	continuous
Temperature	36.6	continuous
Weight	81.0	continuous
pH	7.4	continuous



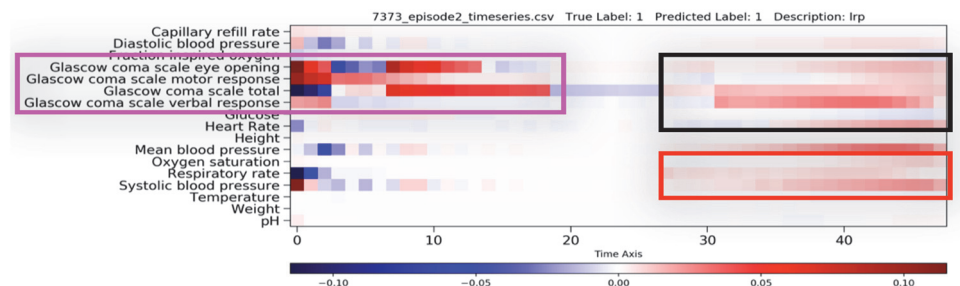
Նկ. 2. Channel-wise LSTM-ի ճարտարապետությունը

Ստացված արդյունքները. Յուրաքանչյուր օրինակի համար կառուցվում են գրադիենտային հիմք ունեցող և կարևորության հետընթաց մակարդակային շարժմամբ ստացված պատկերավորումները: Ի տարբերություն գրադիենտով պատկերավորման՝ *LRP*-ով ստացած պատկերների մեջ առկա են ինչպես կապույտ, այնպես էլ կարմիր վանդակներ: Կապույտները բնորոշում են բացասական ներգործություն վերջնական կանխատեսման վրա, ինչը հանգեցնում է նրան, որ վերջին շերտի սիգմաձև ֆունկցիան փոքր լինի 0,5-ից, իսկ կարմիրները՝ դրական:

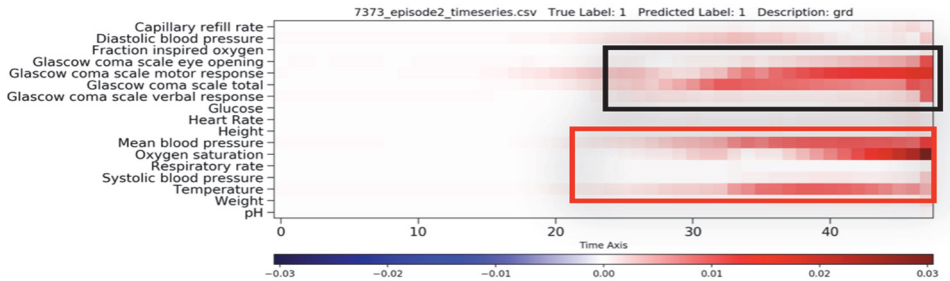
Քննարկենք մահվան ելք արձանագրած օրինակ, որի համար ցանցը ճիշտ կանխատեսում է կատարել (*True Positive*): Նկ. 3-ում պատկերված է այդպիսի օրինակ, իսկ նկ. 4...5-ում երկու մեթոդների միջոցով ստացված համապատասխան կարևորության պատկերներն են: Երկու պատկերավորումների մեջ մեծ կարևորություն է տրվել *Glasgow coma scale eye opening*, *Glasgow coma scale motor response*, *Glasgow coma scale total*, *Glasgow coma scale verbal response* նկարագրերին: Վերջին մի քանի ժամերին վերոնշյալ նկարագրերը կտրուկ փոփոխվել (վատթարացել) են, ինչի հիման վրա *Channel-wise LSTM* ցանցը կարողացել է ճիշտ որոշում կայացնել: Ի տարբերություն գրադիենտային պատկերավորման, որը հիմնականում կարևորում է նկարագրերի վերջին ժամերի արժեքները, կարևորության հետընթաց մակարդակային շարժմամբ ստացված պատկերավորումը ուշադրություն է դարձնում վերոնշյալ նկարագրերի ինչպես սկզբնական, այնպես էլ վերջնական չափումներին (նկ. 3...5-ում մանուշակագույն և սև ուղղանկյունները): Հատկանշական է նաև, որ կարևորության հիմնական մասը կենտրոնացված է *Glasgow coma scale eye opening*, *Glasgow coma scale motor response*, *Glasgow coma scale total*, *Glasgow coma scale verbal response* նկարագրերի վրա, որոնք բնութագրում են մարդու գիտակցության վիճակը, և դրանք ծառայում են որպես առաջնային զննման նկարագրեր, ինչը ուղղակիորեն կապված է մահացության հետ:



Նկ. 3. Մուտքային տվյալների օրինակ (նայել գունավոր տարբերակով). *True Positive*



Նկ. 4. Կարևորության հետընթաց տարածմամբ ստացված պատկերավորում. *True Positive*



Նկ. 5. Գրադիենտային հիմք ունեցող պատկերավորում (նայել գունավոր տարբերակով).
True Positive

Իրականացման մանրամասներ (Implementation Details). Իրականացումը կատարվել է Google Colab միջավայրում, որը հնարավորություն է ընձեռում անվճար օգտագործել համեմատաբար մեծ RAM ունեցող ապարատային ռեսուրսներ (CPU, GPU, TPU): Այն նաև տրամադրում է ինտերֆեյս՝ Python 3 ծրագրավորման լեզվով աշխատելու համար: Հիմնականում օգտագործվել են Keras և Numpy գրադարանները, որոնցից առաջինը նախատեսված է խոր ուսուցման մոդելներով, իսկ երկրորդը՝ կամայական չափողականություն ունեցող թենսորներով աշխատելու համար:

[1]-ում առաջարկված Channel-wise LSTM ցանցը իրականացվել էր Keras գրադարանով, որը ուղղակի հասանելիություն չի տրամադրում ռեկուրենտ ցանցերի հանգույցների արժեքներին, որոնք անհրաժեշտ են կարևորության հետընթաց մակարդակային շարժման ալգորիթմում: Այդ պատճառով օգտագործված ցանցերի դասերը վերասահմանվել (override) են, և ավելացվել է համապատասխան ֆունկցիոնալությունը:

Կարևորության հետընթաց մակարդակային շարժման ալգորիթմում α -ն ընտրվել է 0.01, իսկ δ -ի δ -ի այսպիսի ընտրության արդյունքում bias միավորներին կարևորություն չի տրվել:

Ի տարբերություն [5]-ում նկարագրված նախադասության սենտիմենտի կանխատեսման խնդրի, որտեղ դասերի քանակը հավասար էր 5-ի, մահվան կանխատեսման խնդրում դասերի քանակը հավասար է 2-ի, և սոֆթմաքսի փոխարեն կիրառվել է սիգմոիդ ելքային միավորը: Այս տարբերությունը պատճառ է հանդիսանում՝ վերանայելու վերջին շերտի կարևորության արժեքը, որպեսզի հնարավոր լինի ստանալ կարևորության պատկերներ՝ ըստ նպատակային դասի: Channel-wise LSTM ցանցի վերջին շերտը նկարագրվում է հետևյալ բանաձևով՝ $S = \sigma(W h_T^T + b)$, որտեղ h_T -ն LSTM ցանցի վերջնական բջջային վիճակն է, իսկ W, b - ն՝ վերջին լրիվ կապակցված շերտի կշիռները: Որպես h_T միավորի կարևորության վեկտոր վերցրվել է հետևյալը՝ $R(h_T) = Mask \odot W \odot h_T$, որտեղ $Mask$ -ը

բինար վեկտոր է, որը ցույց է տալիս, թե $W \odot h_T$ վեկտորի որ արժեքներն են նշանով համընկել ($Wh_T^T + b$)-ի նշանի հետ: Այլ կերպ ասած՝ $Mask = 1\{sign(W \odot h_T) = sign(S - S(0))\}$, որտեղ 1-ը ինդիկատոր ֆունկցիա է և ընդունում է 1 արժեք, եթե փակագծերում ներառած պայմանը ճիշտ է:

Եզրակացություն. Աշխատանքում փորձ է արվել՝ պատկերավորելու [1]-ում առաջարկված նեյրոնային ցանցերից մեկը՝ *channel-wise LSTM*-ը: Կիրառվել են երկու տեսակի պատկերավորման մեթոդներ. ցանցի որոշումները բացատրելու համար հետագա աշխատանքները ուղղված կլինեն ցանցի սխալ կայացրած որոշումները շտկելուն՝ օգտագործելով նկարագրերի կարևորության արժեքները:

Բոլոր կոդերը հասանելի են հետևյալ հասցեով.
<https://github.com/naymaraq/RNN-Visualization>.

ԳՐԱԿԱՆՈՒԹՅԱՆ ՑԱՆԿ

1. Multitask Learning and Benchmarking with Clinical Time Series Data / **H. Harutyunyan, H. Khachatrian, C.D. Kale, V.G. Steeg and A. Galstyan** // arXiv: 1703.07771.- Aug. 2019.
2. **Erhan D., Bengio Y., Courville A., and Vincent P.** Visualizing higher-layer features of a deep network: Technical Report 1341. - University of Montreal, Jun 2009.
3. **Simonyan K., Vedaldi A., Zisserman A.** Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps // ICLR. - 2014
4. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation/ **S. Bach, A. Binder, G. Montavon, et al.** – Universidad de Castilla-La Mancha, Spain, Jul. 2015.
5. **Jiwei L., Xinlei C., Hovy E., Jurafsky D.** Visualizing and understanding neural models in NLP // Proceedings of NAACL-HLT 2016.- San Diego, California, June 12-17, 2016.
6. **Arras L., Montavon G., Müller K., Samek W.** Explaining recurrent neural network predictions in sentiment analysis // Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017. - Copenhagen, Denmark, 8 September, 2017. –P. 159-168.
7. Mimic-iii, a freely accessible critical care database / **A.E. Johnson et al.** - Sci. data 3, 2016.

Երևանի պետական համալսարան: Նյութը ներկայացվել է խմբագրություն 11.02.2020:

Г.А. ХАЧАТРЯН, Д.С. КАРАМЯН, Т.С. КАРАМЯН

ОБЪЯСНЕНИЕ РЕШЕНИЙ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ ЧЕРЕЗ ВИЗУАЛИЗАЦИЮ

Современные нейронные сети могут аппроксимировать любую сложную функцию. Поэтому они могут с высокой эффективностью работать в задачах распознавания образов, обработки естественного языка и др. При этом неясно, каким образом нейронные сети осуществляют прогнозирование. Для этого необходимо понять, какие входные данные играют важную роль в принятии решений. В медицинской сфере важное значение имеет интерпретируемость нейронных сетей. Например, если нейронная сеть прогнозирует, что у человека имеется какое-то заболевание, то необходимо выяснить, какие входные характеристики оказывают большое влияние на прогноз.

В работе сделана попытка визуализировать работу нейронных сетей. В исследовании использовалась рекуррентная нейронная сеть, обученная по медицинской базе данных MIMIC-III, которая получает доступ к временным рядам клинических данных пациента в отделении реанимации и прогнозирует вероятность смерти в отделении реанимации. Были адаптированы два метода, разработанные для распознавания изображений и обработки естественного языка для нейронных сетей, обученных по клиническим данным, и визуализации работы этих сетей.

Ключевые слова: рекуррентные нейронные сети, градиентная визуализация, обратное слоевое распространение значимости, база данных MIMIC-III, клинические временные ряды.

H.H. KHACHATRYAN, D.S. QARAMYAN, T.S. QARAMYAN

EXPLAINING DECISIONS OF RECURRENT NEURAL NETWORKS THROUGH VISUALIZATION

Neural Networks can approximate any complex function, so they work very well in many tasks such as computer vision, natural language processing, etc. Despite their performance, it is unclear, how neural networks incorporate with input features and make decisions. In particular, we need to understand which input feature plays an important role in the decision-making process. In medicine, the interpretability of neural nets is very important. For instance, when NN predicts that a person has some disease, it is necessary to understand which input features have a great impact on the prediction.

In this work, an attempt has been made to shed light on the work of neural networks or, in other words, to visualize their work. The study recruited a recurrent neural network trained on the MIMIC-III medical database, which accesses the time series of clinical data of the patient in the resuscitation department and tries to predict the probability of death in the mentioned department. In this paper we proposed an activity that can manage to adapt two methods developed for image recognition and natural language processing for neural networks based on clinical data is proposed and visualize the work of these networks.

Keywords: recurrent neural networks, gradient based visualization, layer wise relevance propagation, MIMIC-III, clinical time series.