# ԽՈՐԱՑՎԱԾ ՈՒՍՈՒՑՄԱՆ ԿԻՐԱՌՈՒՄԸ ՏԵՍԱՆՅՈՒԹԻ ՀԱՄԱՐ ԵՐԱԺՇՏՈՒԹՅԱՆ ԸՆՏՐՈՒԹՅԱՆ ԳՈՐԾԸՆԹԱՑՈՒՄ

**ՄԱՍՏՈՅԱՆ ԿԱՐԵՆ**
*ԳՊՀ բնագիտական ֆակուլտետի*
*համակարգչային ճարտարագիտության բաժնի*
*մագիստրատուրայի 2-րդ կուրսի ուսանող*
էլփոստ՝ kmastoyan@mail.ru

***Հոդվածում քննվում են տեսանյութի բովանդակության վերլուծության հիման վրա արհեստական բանականության միջոցով երաժշտության ընտրության իրականացման հնարավորությունները: Նկարագրված է, թե ինչպես կարելի է իրար կապել տեսանյութի գործողությունները, տեքստը և երաժշտությունը՝ օգտագործելով թվային դրոշմների տեխնոլոգիան (digital fingerprinting): Արդյունքում հնարավոր է ստեղծել համակարգ, որը հնարավորություն կտա տեսանյութի համար ընտրել երաժշտություն՝ հաշվի առնելով տեսանյութի բովանդակությունը և կատարվող գործողությունները: Նմանատիպ համակարգերը կարող են օգտագործվել ինչպես սիրողական մակարդակում սոցցանցերում տարածվող տեսանյութերի համար երաժշտություն ընտրելիս, այնպես էլ մասնագիտական մակարդակում տեսանյութերի խմբագրման ծրագրերում որպես հավելված, որը հնարավորություն կտա առավել հեշտ և արագ ընտրել համապատասխան երաժշտություն: Այսպիսի համակարգերը կարող են նաև, բացի հենց երաժշտության ավտոմատ ընտրությունից, ներկայացնել առաջարկություններ:***

***Բանալի բառեր՝*** *երաժշտության ընտրություն, խորը ուսուցում, տեսանյութից գործողությունների հայտնաբերում և ճանաչում:*

Ներկայումս որպես մասնագիտություն սկսել է մեծ տարածում գտնել վիդեոբլոգինգը, և գնալով ավելի շատ մարդիկ են սկսում զբաղվել այդ մասնագիտությամբ: Վիդեոբլոգերության կարևոր բաղկացուցիչ մասն է կազմում նկարված տեսանյութի խմբագրումը և համապատասխան երաժշտության ընտրությունը: Բացի դրանից՝ երաժշտությունն ուղեկցում է տեսանյութերին գրեթե ամենուր՝ սկսած գովազդերից մինչև ֆիլմեր: Ուստի երաժշտության ընտրության հարցը ինչպես սիրողական, այնպես էլ մասնա-

գիտական մակարդակում կարևոր և արդիական խնդիր է: Արհեստական բանականությամբ համակարգերը կարող են նպաստել նկարագրված խնդրի լուծման հեշտացմանը:

Ենթադրենք՝ ունենք մեր սեփական տեսահոլովակը, և որոշ տեսարաններում երաժշտության կարիք կա: Այս մեթոդը նկարագրում է, թե ինչպես կարելի է կապել գործողությունները, տեքստը և երաժշտությունը՝ օգտագործելով թվային դրոշմերի տեխնոլոգիան (digital fingerprinting):

Հոդվածում ներկայացված է մեթոդ, որի միջոցով կարելի է ընտրել տեսանյութի ֆոնային երաժշտություն:Նկարագրված մեթոդը թույլ է տալիս մեքենայական ուսուցման միջոցով ընտրել տեսանյութի երաժշտություն: Այս հայեցակարգը կարող է իրականացվել՝ օգտագործելով խորը ուսուցում (deep learning)՝ որպես մեքենայական ուսուցման մեթոդների ավելի լայն ընտանիքի մաս:

Իրականացված հետազոտությունները ցույց տվեցին, որ ներկայումս առկա տեխնիկական և ծրագրային միջոցները հնարավորություն են տալիս ամբողջությամբ իրականացնել առաջարկվող մեթոդը, պարզապես անհրաժեշտ է մի քանի ծառայություններ միավորել մեկ հավելվածում և ստանալ մեկ ընդհանուր համակարգ, որը հետագայում կարող է վերածվել նույնիսկ արտադրանքի: Այնուամենայնիվ, հավելվածը կլինի բավականին ռեսուրսատար՝ մեքենայական ուսուցման կիրառմամբ ու միաժամանակ մի քանի խնդիրների լուծումով պայմանավորված, ուստի հնարավոր է, որ, օրինակ, հեռախոսների հզորությունը չբավարարի ամբողջությամբ նմանատիպ խնդիրը լուծելու համար: Անհրաժեշտություն առաջացավ ուսումնասիրել նաև այնպիսի ամպային ծառայություններ, որոնք լուծում են երկու խնդիր համակարգի աշխատանքի համար.

1. տեսանյութի բովանդակությունից (Digital video fingerprinting), տեղի ունեցող գործողությունից (Action Recognition) կախված՝ որոշակի բանալի բառերի կամ digital fingerprint-ի առանձնացում (նմանատիպ ծառայություն մատուցողներ՝ Microsoft Azure computer vision services),
2. բանալի բառերի կամ digital fingerprint-ի առաձնացում երաժշտությունից (acoustic fingerprint) (նմանատիպ ծառայություն մատուցողներ՝ YouTube, ContentID, Shazam);

Այնուհետև ստացված սարդյունքները կարելի է համատեղել հավելվածում և առաջարկել գտնված համապատասխան մի քանի երաժշտության տարբերակներ:

**Երաժշտության ընտրության խնդիրը**

Տեսահոլովակի համար երաժշտություն ընտրելիս մենք բախվում ենք մի շարք խնդիրների, որոնցից մի քանիսը վերաբերում են դեդուկտիվ և ինդուկտիվ մտածողության համահունչությանը:

Առաջին հերթին մենք կսկսենք՝ նկարագրելով տեսարանը: Եկեք այս մտածելակերպը կոչենք «առաջ», երբ դիտում ենք տեսանյութ, այնուհետև վերլուծում և դասակարգում ենք այն, ինչ տեսնում ենք, փորձում ենք հասկանալ և նկարագրել առաջացած հույզերը: Դրանից հետո մենք պետք է հետադարձ մտածենք: Օգտագործելով մեր վերլուծության և զգացմունքների նկարագրության արդյունքները՝ մենք պետք է մեր մտքում գտնենք երաժշտության մի կտոր, որը հարմար է տվյալ տեսարանի համար:

Այսպիսով, երաժշտությունն ու գործողությունը ունեն ընդհանուր հայտարար, որը կարելի է արտահայտել բառերով, բայց դա այնքան էլ հեշտ չի լինի, եթե դրանում փորձ չունենանք: Երաժշտական ստեղծագործությունը նկարագրելու եղանակներից մեկն է հետևել ուղեցույցներին և վերլուծել երաժշտությունը՝ որպես արվեստի ձև: Դա շատ դժվար կլինի և որպես արդյունք՝ կարող է դառնալ շատ սուբյեկտիվ: Այնուամենայնիվ, եթե մտադիր ենք արհեստական բանականությունն օգտագործել երաժշտություն ընտրելու համար, ապա մեզ հարկավոր կլինի գտնել մեկ այլ ընդհանուր թվային հայտարար: Այն պետք է լինի բնօրինակ մեդիային համապատասխան գործառույթների փոքր քանակ, որոնք կարող են միմյանց հետ կապված լինել տվյալների բազայի հիմնական դաշտերի նման: Այս մեթոդը հիմնված է թվային տեսանյութի և ձայնային դրոշմների միջև ժամանակի հասկացության վրա:

**Գործողությունների ճանաչում**

Առաջին քայլում մենք կարող ենք լուծել գործողությունների ճանաչման խնդիրը: Այսօր մեքենայական ուսուցման համակարգերը կարող են հայտնաբերել և նույնականացնել տեսանյութում տեղի ունեցող գործողությունները՝ օգտագործելով տեսանյութի դրոշմները (video fingerprinting): Նմանատիպ գործողությունը կարելի է փորձարկել Microsoft Azure Computer Vision Services-ում:

Այս ոլորտը հայտնի է դարձել նաև Kaggle-ում[6]: Անցյալ տարի Google-ը մեկնարկեց մրցույթի առաջին փուլը՝ դասակարգման ալգորիթմներ մշակելու համար, որոնք ճշգրիտ կերպով հատկացնում են պիտակներ տեսանյութերի մակարդակով՝ օգտագործելով Youtube-ի նոր և կատարելագործված YT-8M V2 տվյալների հավաքածուն:

Գործողության ճանաչման նպատակը ֆիլմերի պատկերներից տեսարանների բացահայտումն է: Այս փուլի արդյունքում մենք կունենանք տեսանյութի թվային հետք և գործողության տեքստային նկարագրություն:

**Երաժշտության ընտրությունը**

Երկրորդ քայլում իրականացվում է տեսարանի համար լավագույն երաժշտության ընտրությունը: Բարեբախտաբար, շատ դեպքերում մենք գիտենք, թե որ երաժշտությունը պետք է համապատասխանի դինամիկ պատկերով ներկայացված պատմությանը: Մեզ օգնում է կյանքի փորձը: Ֆիլմերը հագեցած են հնչյունների և գործողությունների համադրությամբ: Մանկուց մենք գիտենք, թե որ հնչյուններն են համապատասխանում վախի, սիրո, երջանկության և այլնի ձգտմանը կամ հույզերին:

Ֆիլմարտադրությունը իր հայտնիության շնորհիվ սահմանում է ստանդարտ համապատասխան երաժշտության ընտրման գործընթացում: Այսպիսով, մենք կարող ենք օգտագործել ֆիլմերի սաունդթրեքները՝ որպես տեսանյութերի օրինակներ: Այս դեպքում Movieclips-ը կարող է օգնել մեզ, քանի որ այն YouTube-ում հասանելի ամենամեծ կինոցանցն է: Յուրաքանչյուր հոլովակ պարունակում է ֆիլմից կարճ տեսարան և դրա մասին հակիրճ տեքստ:

Այնուամենայնիվ, սաունդթրեքները չեն ընդգրկում ֆիլմի կամ տեսարանի ամբողջ նկարագրությունը: Երաժշտությունը հայտնվում է այնտեղ, որտեղ հեղինակը որոշել է գրավել դիտողի ուշադրությունը: Այսպիսով, հաջորդ խնդիրը երաժշտությունը մեկ այլ ձայնից կամ դրա բացակայությունից մեկուսացնելն է: Համակարգերը, ինչպիսիք են YouTube-ը, Content ID-ն կամ Shazam-ը, կարող են դա անել՝ օգտագործելով դրոշմների ակուստիկ տեխնոլոգիա:

Ձայնի որոնման համար կարևոր է ձայնագրությունից դրոշմ ստեղծելը[1]: Սովորական մոտեցումներից մեկն այն է, որ ստեղծվի ժամանակի հաճախականության գծապատկեր, որը կոչվում է սպեկտրոգրամ: Աուդիո ցանկացած կտոր կարող է փոխակերպվել սպեկտրոգրամի: Ձայնի յուրաքանչյուր կտոր ժամանակի ընթացքում բաժանվում է մի քանի հատվածի: Որոշ

դեպքերում հարակից հատվածները կիսում են ընդհանուր ժամանակային սահման, իսկ մյուս դեպքերում հարակից հատվածները կարող են համընկնել:

**Տեսանյութից գործողությունների հայտնաբերում և ճանաչում**

Գործողությունների հայտնաբերման և ճանաչման խնդիրը ներառում է տեսահոլովակներից տարբեր գործողությունների բացահայտում (2D պատկերների հաջորդականություն), որտեղ գործողությունը կարող է կատարվել կամ չկատարվել տեսանյութի ողջ ընթացքում: Թվում է, թե սա պարզապես պատկերի դասակարգման խնդիր է՝ յուրաքանչյուր հաջորդ պատկերի համար կանխատեսումների հետագա համախմբմամբ: Չնայած պատկերի դասակարգման (ImageNet) խորացված ուսուցման ճարտարապետությունների ապշեցուցիչ հաջողությանը՝ տեսադասակարգման և ներկայացման ճարտարապետության առաջընթացը դանդաղ է ընթանում[2]:

Մինչև խորացված ուսուցումը, գործողությունները ճանաչելու համակարգչային տեսողության ավանդական ալգորիթմների մեծ մասի բովանդակությունը կարելի է բաժանել հետևյալ 3 հիմնական քայլերի.

1. Դուրս են բերվում այն լոկալ բազմաչափ երևացող տարրերը(խիտ հետագծի ընտրանք), որոնք նկարագրում են տեսանյութում իրականացվող հիմնական բովանդակությունը:
2. Դուրս բերված հատկանիշները համակցված են ֆիքսված չափի տեսաշերտի նկարագրության մեջ: Այս քայլի իրականացման հայտնի տարբերակներից են հիերարխիկ կլաստերացման կամ k-means կլաստերավորման կիրառումը:
3. SVM կամ RF դասակարգիչները կարող են օգտագործվել վերջնական կանխատեսման համար:

Այս ալգորիթմներից խիտ հետագծի ընտրանքը (improved Dense Trajectories [3] (iDT)) արդիականներից է: Միաժամանակ 2013-ին գործողությունները ճանաչելու համար օգտագործվեցին նաև եռաչափ փաթութային ցանցեր[4]: Կան գործողությունների հայտնաբերման տարբեր մոտեցումներ, որոնցից են Single Stream Network և Two Stream Networks, որոնք իրենց հերթին բաժանվում են տարբեր տարատեսակների՝ LRCN, C3D, Conv3D & Attention, TwoStreamFusion, TSN, ActionVlad, HiddenTwoStream, I3D, T3D:

**Հայեցակարգ**

Ինչպես արդեն նշվեց, հայեցակարգը պարզ է: Այն ներկայացնում է տեսաֆիլմերի և երգի միջև կապը՝ որպես թվային տեսադրոշմի և ձայնադրոշմի ժամանակավոր համապատասխանություն:

Դասակարգման խնդրի լուծմանը միտված խորացված ուսուցման ալգորիթմը կարելի է բաժանել երկու փուլի.

1. Ուսուցում - Վերլուծել պիտակավորված ֆիլմերի տվյալների բազան (հայտնաբերե՛ք գործողություններ և հնչյուններ) և ստեղծել դրոշմերի տվյալների շտեմարան՝ պիտակավորված գործողություններով և համապատասխան աուդիո հետքերով:

2. Փորձարկում - Երաժշտական տվյալների շտեմարանից ընտրվում են հետքերի վերնագրերը: **Երբ մենք աշխատում ենք համակարգչային տեսողության նեյրոնային ցանցի հետ, պետք է հաշվի առնենք հավանականությունը:**

Համակարգում տվյալների մուտքագրման երկու տարբերակ կա: Ձայնը, տեսանյութը և տեքստը փոխկապակցված են մեդիա ֆայլերի կամ հոսքային տեսանյութերի ներսում:

1. Տեսանյութ

Մեքենայական ուսուցման համակարգերը կբացահայտեն տեսանյութում կատարվող գործողությունները, կստեղծեն տեսադրոշմներ, կփնտրեն նմանատիպ դրոշմներ ֆիլմերի տվյալների բազայում, արդյունքները կգնահատեն ըստ համընկման աստիճանի, արդյունքները կկապեն ֆիլմի սաունդթրեքների հետ և կվերադարձնեն սաունդթրեքների ցանկը հղումներով, օրինակ՝ Google Music կամ Apple iTunes:

2. Տեքստ

Տեսանյութի կամ դրա կարևոր մասի տեքստային նկարագրությամբ սաունդթրեքների հայտնաբերումը, որտեղ ցանկանում եք տեղադրել երգը, մի փոքր բարդացնում է հայեցակարգը: Ալգորիթմը պետք է լրացվի գործընթացով, որը փնտրում է գործողություն-նկարագրություն համընկնում:

**Կողմնակի ազդեցություն**

Գոյություն ունի ևս մեկ օգտակար տարբերակ՝ հիմնված այն անորոշության վրա, որը գալիս է համակարգի սովորած նոր տեսանյութի և տեսանյութի դրոշմների տարբերությունից:

Խորը ուսուցման դասակարգման մոդելները հաճախ արտադրում են նորմալացված միավորների վեկտորներ, որոնք պարտադիր չէ, որ արտա-

ցոլեն մոդելի անորոշությունը: Այն կարելի է ֆիքսել՝ օգտագործելով խորը ուսուցման բայեսյան մոտեցումները, որոնք գործնական շրջանակ են առաջարկում խորը ուսուցման մոդելների միջոցով անորոշությունը հասկանալու համար:

Այս անորոշությունը կարող է օգտագործվել որպես պոտենցիալ փոփոխությունների չափանիշ՝ ստեղծելու նոր աուդիո դրոշմ՝ օգտագործելով գեներացնող adversarial ցանց[5]: Մեկ այլ տարբերակ է՝ իրականացնել Ֆուրյեի հակադարձ արագ փոխակերպում (FFT), որը կարող է առաջացնել նոր յուրահատուկ երգ (երաժշտություն): FFT-ի պատճառած բնական կորուստը կարող է փոխհատուցվել նախկինում ստեղծված սաունդթրեքների ցանկից ամենահարմար երաժշտության ամենամոտ հատվածներով:

**Եզրակացություն**

Նկարագրված հայեցակարգը ցույց է տալիս, թե ինչպես ընտրել երաժշտություն տեսահոլովակի համար մեքենայական ուսուցման համակարգի միջոցով: Սա կարող է ավելի հեշտ դարձնել տեսահոլովակի կամ հոսքային մեդիայի համար մեզ անհրաժեշտ լավագույն երգի ընտրությունը: Այն կարող է նաև ապահովել երաժշտության ընտրանքներ՝ ավելի լավ ընտրության համար:

Հայեցակարգը ցույց է տալիս իր հզոր ներուժը՝ դառնալ այնպիսի հզոր ծառայությունների հիմնական հատկանիշ, ինչպիսիք են Youtube-ը, Google Photos-ը, Microsoft Azure-ը, Adobe Premiere Pro-ն, Corel VideoStudio-ն և Pinnacle Studio-ն: Գործողությունները և բառերը կարող են օգտագործվել երաժշտություն ստեղծելու համար: Հնարավոր է, որ այն այնքան գեղեցիկ չլինի, որքան մարդկային կոմպոզիցիան, բայց, ամենայն հավանականությամբ, լինի եզակի և կարող է խթանել հեղինակի ստեղծագործական կարողությունը:

Բացահայտված կողմնակի ազդեցությունը կարող է դառնալ մարդկային ստեղծագործության մեծ դաշտ: Սա կարող է հնարավորություն ստեղծել իմանալու, թե ինչպես է հնչում գործողությունը կամ պատմությունը:

**Օգտագործված գրականության ցանկ**

1. Sunil Lee and Chang D Yoo. Video fingerprinting based on centroids of gradients orientations. In ICASSP '06: Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 401_404, Washington, DC, USA, 2006.

2. S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. PAMI, 35(1):221– 231, 2013.
3. H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In CVPR. IEEE, 2011.
4. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901, 2013.
5. https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29
6. https://www.kaggle.com/c/youtube8m/code

## APPLYING DEEP LEARNING TO THE PROCESS OF CHOOSING THE RIGHT MUSIC

**MASTOYAN KAREN**
*2nd-year Master student of Computer Engineering Depertment*
*Faculty of Natural Sciences, GSU*
e-mail: kmastoyan@mail.ru

The article discusses the possibilities of choosing music based on the analysis of video content using artificial intelligence. It describes how video, text and music can be linked using digital fingerprinting technology. As a result, we can create a system that will allow us to choose music for a video, taking into account the actions performed in the video. Such systems can be used both in the choice of music for videos published on social networks at the amateur level, and at the professional level in video editing software as an add-on, which will allow us to select the appropriate music more easily and quickly. Such systems, in addition to automatically music selection, can make suggestions.

***Key words:*** *music selection, deep learning, action detection and recognition.*

# ПРИМЕНЕНИЕ УГЛУБЛЕННОГО ОБУЧЕНИЯ В ПРОЦЕССЕ ВЫБОРА МУЗЫКИ ДЛЯ ВИДЕО

**МАСТОЯН КАРЕН**

*Студент 2-го курса магистратуры отделения Компьютерная инженерия факультета естественных наук ГГУ*

В статье рассматриваются возможности выбора музыки на основе анализа видеоконтента с помощью искусственного интеллекта. Описывается, как можно связать видео, текст и музыку с помощью технологии цифровой отпечатки. В результате можно создать систему, которая позволит выбирать музыку для видео с учетом выполняемых действий в ролике. Такие системы можно использовать как при выборе музыки для видео, публикуемые в социальных сетях на любительском уровне, так и на профессиональном уровне в программном обеспечении для редактирования видео в качестве приложения, что позволит более легко и быстро выбирать подходящую музыку. Такие системы кроме автоматического выбора музыку могут сделать предложения.

***Ключевые слова:*** *выбор музыки, углублённое обучение, обнаружение и распознавание действий из видео.*

Հոդվածը ներկայացվել է խմբագրական խորհուրդ 28.08.2021թ.:
Հոդվածը գրախոսվել է 14.10.2021թ.: