BOLALAR UCHUN ZARARLI VIDEO VA TASVIR KONTENTINI ANIQLASHNING MULTIMODAL INTELLEKTUAL USULLARI

Toshboyeva Guljahon O'tkir qizi

Annotatsiya. Maqolada bolalarga mo'ljallangan onlayn video va tasvirlarda yashirin zararli kontentni aniqlash muammosi ko'rib chiqiladi. Zararli mazmun ko'pincha tasvir, ovoz va matnning birgalikdagi ta'sirida namoyon bo'lishi sababli, faqat bitta ma'lumot turini tahlil qiluvchi usullarning cheklanganligi asoslanadi. Vizual, audio va matnli oqimlarni e'tibor mexanizmi asosida birlashtiruvchi hamda natijani milliy qonunchilikdagi yosh toifalariga mos tartibli (ordinal) klassifikatsiya ko'rinishida beruvchi multimodal yondashuv taklif etiladi. Hisoblash xarajatlarini kamaytirish uchun kaskadli tahlil sxemasi ham ko'rib chiqiladi.

Kalit so'zlar: multimodal o'qitish, zararli kontent, bolalar xavfsizligi, kompyuter ko'rish, nutqni tanib olish, yosh toifalari, ordinal klassifikatsiya, video tahlil.

Аннотация. В статье рассматривается проблема выявления скрытого вредоносного контента в онлайн-видео и изображениях, ориентированных на детей. Поскольку вредоносный смысл часто возникает из совместного воздействия изображения, звука и текста, обосновывается ограниченность методов, анализирующих только один тип данных. Предлагается мультимодальный подход, объединяющий визуальный, аудио- и текстовый потоки на основе механизма внимания и выдающий результат в виде порядковой (ординальной) классификации, соответствующей возрастным категориям национального законодательства. Для снижения вычислительных затрат рассматривается каскадная схема анализа.

Ключевые слова: мультимодальное обучение, вредоносный контент, безопасность детей, компьютерное зрение, распознавание речи, возрастные категории, ординальная классификация, анализ видео.

Abstract. The paper addresses the problem of detecting hidden harmful content in online videos and images aimed at children. Since harmful meaning often emerges from the combined effect of imagery, sound and text, the limitations of methods that analyse only a single data type are substantiated. A multimodal approach is proposed that fuses visual, audio and text streams through an attention mechanism and produces its output as an ordinal classification aligned with the age categories defined in national legislation. A cascaded analysis scheme is also considered to reduce computational cost.

Keywords: multimodal learning, harmful content, child safety, computer vision, speech recognition, age ratings, ordinal classification, video analysis.

Kirish

Video platformalar bugungi kunda bolalarning bo'sh vaqt o'tkazishi va bilim olishining asosiy manbalaridan biriga aylangan. UNICEF ta'kidlaganidek, bolalar internet foydalanuvchilarining salmoqli

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

qismini tashkil etadi va raqamli muhit ular uchun ham imkoniyatlar, ham xavflar manbai hisoblanadi [1]. Ayniqsa, tashqi ko'rinishi bolalarga mo'ljallangan, ya'ni mashhur multfilm qahramonlari, yorqin muqova va beg'ubor sarlavhaga ega, ammo ichida zo'ravonlik, qo'rqituvchi sahnalar yoki yoshga mos bo'lmagan mazmun bo'lgan videolar jiddiy muammo sifatida e'tirof etilmoqda. Bunday videolar bo'yicha o'tkazilgan tadqiqot ularning YouTube platformasida keng tarqalganini hamda bola o'z yoshiga mos videoni tomosha qilayotganda ham tavsiya algoritmi orqali bir necha qadamda nomaqbul videoga o'tib qolish ehtimoli mavjudligini ko'rsatgan [2].

O'zbekistonda axborot mahsulotlari O'RQ-444-son "Bolalarni ularning sog'lig'iga zarar yetkazuvchi axborotdan himoya qilish to'g'risida"gi Qonun va uni amalga oshirish bo'yicha hukumat qarori asosida 0+, 7+, 12+, 16+ va 18+ yosh toifalariga ajratiladi, Internetga joylashtiriladigan axborot ham shu talablar doirasiga kiradi [3]. 2024-yil 14-oktabrdagi PQ-358-son qaror bilan tasdiqlangan Sun'iy intellekt texnologiyalarini 2030-yilga qadar rivojlantirish strategiyasida esa ijtimoiy sohada sun'iy intellekt yechimlarini joriy etish ustuvor yo'nalish sifatida belgilangan [4]. Biroq har daqiqada platformalarga yuklanayotgan videolarning yosh toifasini ekspertlar tomonidan qo'lda aniqlash amalda imkonsiz.

Mavjud avtomatik usullarning aksariyati faqat bitta ma'lumot turini — yoki tasvirni, yoki matnni tahlil qiladi. Holbuki, zararli ma'no ko'pincha bir necha modallikning birikmasidan kelib chiqadi: zararsiz tasvir unga qo'shilgan yozuv bilan haqoratga aylanishi, beg'ubor multfilm sahnasi esa qo'rqituvchi ovoz bilan bolani cho'chitishi mumkin. Multimodal memlar bo'yicha o'tkazilgan tadqiqotda faqat tasvir yoki faqat matnga tayangan modellar bunday holatlarda past natija ko'rsatgani, multimodal modellar esa inson darajasidan hali ancha orqada qolayotgani qayd etilgan [5]. O'zbek tilidagi kontent uchun muammo yanada murakkab, chunki mavjud modellarning aksariyati asosan ingliz tilidagi ma'lumotlarda o'qitilgan.

Mazkur tezisning maqsadi — bolalar uchun zararli video va tasvir kontentini aniqlashda vizual, audio va matnli ma'lumotlarni birgalikda tahlil qiluvchi multimodal intellektual usulni takomillashtirish yo'nalishlarini asoslashdan iborat.

Tadqiqot materiallari va usullari

Modalliklar va ulardagi xavf belgilari.

Taklif etilayotgan yondashuvda bitta video to'rtta ma'lumot oqimi sifatida qaraladi. Har bir oqim o'ziga xos xavf belgilarini o'z ichiga oladi va ularni aniqlash uchun turli sinfdagi modellar qo'llaniladi (1-jadval).

1-jadval

Video kontent modalliklari, ulardagi xavf belgilari va qo'llaniladigan modellar

№	Modallik	Bolalar uchun xavf belgilari	Qo'llaniladigan modellar
1	Vizual (kadrlar, muqova tasviri)	Zo'ravonlik va qon, qurol, qo'rqituvchi obrazlar, bolalar sevgan qahramonlar ishtirokidagi nomaqbul sahnalar	Vision Transformer (ViT) [6], CLIP [7], fazo-vaqt e'tiborli video modellari [8]

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

№	Modallik	Bolalar uchun xavf belgilari	Qo'llaniladigan modellar
2	Audio	Qichqiriq, o'q ovozi, portlash, qo'rqituvchi musiqa	AudioSet ontologiyasi asosidagi tovush hodisalari tasniflagichlari [9]
3	Nutq va matn (sarlavha, tavsif, izohlar, kadrdagi yozuvlar)	Haqorat va so'kinish, aldovchi sarlavhalar, zararli odatlar targ'iboti	Nutqni tanib olish (Whisper [10]) va o'zbek tiliga moslashtirilgan transformer modellari [11]
4	Metama'lumotlar	Kanal tarixi, teglar, sarlavha va mazmun o'rtasidagi nomuvofiqlik	Gradiyent busting, to'liq bog'langan neyron tarmoqlar

Kalit kadrlarni tanlash.

Videoning har bir kadrini og'ir neyron tarmoq orqali tahlil qilish katta hisoblash resurslarini talab qiladi. Shu sababli birinchi bosqichda sahna almashinuvini aniqlash orqali kalit kadrlar ajratib olinadi: qo'shni kadrlarning rang gistogrammalari orasidagi farq belgilangan chegaradan oshganda yangi sahna boshlangan deb hisoblanadi va har bir sahnadan bir nechta vakil kadr tanlanadi. Bu usul tahlil qilinadigan kadrlar sonini keskin kamaytiradi. Qisqa muddatli xavfli sahnalarni o'tkazib yubormaslik uchun esa sahnalar ichida ham siyrak tekis tanlov saqlanadi.

Nutqni tanib olish va matnli tahlil.

Audio yo'lakdagi nutq Whisper modeli [10] yordamida matnga o'giriladi; bu model ko'p tilli bo'lib, o'zbek tilini ham qo'llab-quvvatlaydi. Aniqlikni oshirish uchun modelni ochiq o'zbek nutq korpusi (USC) [12] ma'lumotlarida qo'shimcha o'qitish ko'zda tutiladi. Hosil bo'lgan matn video sarlavhasi, tavsifi va izohlar bilan birga o'zbek tiliga moslashtirilgan transformer modeli [11] orqali tahlil qilinadi. Shunday qilib, matnli zararli kontentni aniqlash uchun ishlab chiqilgan usullar video tahlilining tarkibiy qismiga aylanadi.

Modalliklarni birlashtirish.

Multimodal o'qitishda ma'lumotlarni birlashtirishning erta (belgilar darajasida), kech (qarorlar darajasida) va gibrid usullari farqlanadi [13]. Kech birlashtirish sodda va bitta modallik mavjud bo'lmaganda ham ishlay oladi, ammo modalliklar o'rtasidagi o'zaro ta'sirni hisobga olmaydi. Taklif etilayotgan yondashuvda har bir modallikning ahamiyati kiruvchi ma'lumotga qarab dinamik aniqlanadigan e'tibor mexanizmiga asoslangan birlashtirish qo'llaniladi:

$$\alpha_m = \exp(u^T \tanh(Wh_m)) / \Sigma_j \exp(u^T \tanh(Wh_j)), \quad (1)$$

$$z = \Sigma_m \alpha_m h_m, \quad (2)$$

bu yerda h_m — m-modallikning umumiy fazoga proyeksiyalangan vektori, W va u — o'qitiladigan parametrlar, α_m — m-modallikning shu video uchun ahamiyat koeffitsiyenti, z — videoning yagona multimodal vektori. Masalan, ovozsiz videoda audio oqimining vazni avtomatik ravishda kamayadi, yozuvlari ko'p bo'lgan memlarda esa matnli oqimning ta'siri ortadi. Koeffitsiyentlarning o'zi ham tushuntirish vositasi bo'lib xizmat qiladi: ular qarorga qaysi ma'lumot turi ko'proq ta'sir qilganini ko'rsatadi.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

Yosh toifalarini ordinal klassifikatsiya sifatida aniqlash.

Ko‘pchilik tizimlar kontentni “zararli” yoki “zararsiz” deb ikkilik tasniflaydi. Biroq qonunchilikdagi yosh toifalari tartiblangan: 16+ toifasidagi videoni 0+ deb belgilash uni 12+ deb belgilashdan ko‘ra ancha jiddiy xato hisoblanadi. Oddiy ko‘p sinfli klassifikatsiya bu tartibni hisobga olmaydi. Shu sababli masalani ordinal regressiya sifatida qo‘yish taklif etiladi. Rang izchilligini ta’minlovchi CORAL usulida [14] K ta toifa uchun K − 1 ta ikkilik masala yechiladi:

$$P(y > r_k \mid x) = \sigma(g(z) + b_k), \quad k = 1, \ldots, K-1, \tag{3}$$

bu yerda g(z) — barcha masalalar uchun umumiy bo‘lgan chiziqli funksiya, b_k — har bir chegara uchun alohida siljish, σ — sigmoid funksiya. Umumiy og‘irliklardan foydalanish ehtimolliklarning toifalar bo‘yicha izchil (monoton) kamayishini ta’minlaydi, yakuniy toifa esa 0,5 dan katta bo‘lgan ehtimolliklar soni bo‘yicha aniqlanadi. Yosh toifalari 0+, 7+, 12+, 16+ va 18+ bo‘lganida K = 5 bo‘ladi.

Kaskadli tahlil sxemasi.

Hisoblash xarajatlarini kamaytirish uchun ikki bosqichli kaskad qo‘llaniladi. Birinchi bosqichda yengil model (masalan, faqat muqova tasviri va metama’lumotlar asosida ishlovchi) videoni tez baholaydi. Agar model ishonchi yetarli darajada yuqori bo‘lsa, ya’ni

$$\max_k p_k \geq \tau, \tag{4}$$

shart bajarilsa (τ — ishonch chegarasi), natija shu bosqichda qabul qilinadi. Aks holda video to‘liq multimodal tahlilga yuboriladi. Bunday erta chiqish (early exit) g‘oyasi neyron tarmoqlarda hisoblash tezligini oshirishning samarali usuli sifatida ma’lum [15]. τ chegarasi aniqlik va tezlik o‘rtasidagi muvozanatni boshqaradi; bolalar xavfsizligi nuqtai nazaridan birinchi bosqichda zararli deb gumon qilingan videolar uchun chegarani yuqoriroq qo‘yish maqsadga muvofiq.

Tizimning umumiy ishlash algoritmi.

Yuqoridagi bosqichlarni birlashtirgan holda tizim quyidagi ketma-ketlikda ishlaydi:

1-qadam. Video kelib tushganda uning metama’lumotlari va muqova tasviri yengil model orqali baholanadi; (4) shart bajarilsa, natija darhol qabul qilinadi.

2-qadam. Aks holda videodan kalit kadrlar va audio yo‘lak ajratib olinadi.

3-qadam. Kadrlar vizual model, audio esa tovush hodisalari tasniflagichi orqali qayta ishlanadi; nutq matnga o‘giriladi va sarlavha, tavsif hamda izohlar bilan birga matnli modelga uzatiladi.

4-qadam. Modallik vektorlari (1)–(2) formulalar bo‘yicha birlashtiriladi va (3) formula asosida yosh toifasi aniqlanadi.

5-qadam. Aniqlangan toifa va tushuntirish (modallik koeffitsiyentlari, xavfli deb topilgan kadrlarning vaqt belgilari) asosida 2-jadvalda keltirilgan javob chorasi tanlanadi.

2-jadval

Yosh toifalari va tizimning tavsiya etilgan javob choralari

Toifa	Qonunchilikdagi mazmuni (qisqacha)	Tizimning tavsiya etilgan javobi
0+	Bola sog‘lig‘i va rivojlanishiga zarar yetkazmaydigan, ezgulikning yovuzlik ustidan g‘alabasi aks etgan kontent	Cheklovsiz ko‘rsatiladi
7+	Qo‘rquv uyg‘otmaydigan, naturalistik bo‘lmagan baxtsiz hodisa yoki kasallik tasvirlari, jinoyatning epizodik tasviri	Boshlang‘ich sinf o‘quvchilari uchun ruxsat etiladi

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

Toifa	Qonunchilikdagi mazmuni (qisqacha)	Tizimning tavsiya etilgan javobi
12+	Zo'ravonlik natural ko'rsatilmagan va qoralangan holda tasvirlangan kontent	12 yoshgacha bolalar uchun ota-ona tasdig'i so'raladi
16+	Giyohvand moddalar haqida ularning xavfi ta'kidlangan ma'lumot, uyatli so'zlar kirmaydigan haqoratli iboralar	16 yoshgacha bloklanadi, ota-onaga xabar beriladi
18+	O'ziga yoki boshqalarga zarar yetkazishga undash, zo'ravonlikni oqlash, uyatli so'zlar, pornografik mazmun	Bolalar uchun bloklanadi, mas'ul shaxsga hisobot yuboriladi

Natijalar va muhokama

Adabiyotlar tahlili va taklif etilgan arxitekturani nazariy asoslash quyidagi xulosalarni chiqarishga imkon beradi. Birinchidan, bolalarga mo'ljallangan zararli videolarni faqat bitta belgi asosida aniqlash yetarli emas. Bunday videolar bo'yicha o'tkazilgan tadqiqotda muqova tasviri, sarlavha, teglar va statistik ko'rsatkichlarni birgalikda tahlil qiluvchi chuqur o'qitish klassifikatori taklif etilgan [2], bu esa turli ma'lumot manbalarini birlashtirish zarurligini tasdiqlaydi. Taklif etilayotgan yondashuvda bu g'oya audio va nutq oqimlari bilan kengaytiriladi.

Ikkinchidan, CLIP kabi tasvir va matnni yagona fazoda ifodalovchi modellar [7] yangi xavf toifalarini qo'shimcha belgilangan ma'lumotlarsiz, faqat matnli tavsif orqali aniqlash (zero-shot) imkonini beradi. Bu xususiyat o'zbek tilidagi kontent uchun ayniqsa qimmatli, chunki mahalliy kontent bo'yicha belgilangan ma'lumotlar to'plamlari juda cheklangan. Shu bilan birga, bunday modellar asosan ingliz tilidagi ma'lumotlarda o'qitilganini hisobga olib, ularni mahalliy kontentda qo'shimcha moslashtirish talab etiladi.

Uchinchidan, ordinal yondashuv tizim xatolarining xarakterini o'zgartiradi. Ikkilik tasniflagich chegaraviy holatlarda keskin xato qiladi, ordinal model esa adashgan taqdirda ham odatda qo'shni toifani tanlaydi. Shu sababli baholash mezonlari sifatida makro-F_1 bilan birga o'rtacha absolyut xato (MAE) ham qo'llaniladi:

$$\mathrm{MAE} = (1/N)\,\Sigma_{i=1}^{N} |\hat{y}_i - y_i|, \qquad (5)$$

bu yerda y_i va $\hat{y}_i$ — i-videoning haqiqiy va bashorat qilingan yosh toifasi tartib raqami, N — videolar soni. MAE qiymati qanchalik kichik bo'lsa, model xatolari shunchalik "yumshoq", ya'ni haqiqiy toifaga yaqin bo'ladi.

To'rtinchidan, amaliy joriy etishda maxfiylik masalalariga alohida e'tibor qaratish lozim. Tizim bolaning qurilmasida yoki maktab tarmog'ida ishlaganda tahlil qilinayotgan kontent haqidagi ma'lumotlar faqat zarur hajmda saqlanishi, ota-onalar va o'qituvchilar uchun esa tizim qarorlari tushunarli ko'rinishda taqdim etilishi kerak. E'tibor koeffitsiyentlari (1) va yosh toifalari bo'yicha ehtimolliklar (3) bu maqsadda tabiiy tushuntirish vositasi bo'lib xizmat qiladi.

Beshinchidan, taklif etilayotgan tizim bloklash vositasi sifatida emas, balki yordamchi vosita sifatida loyihalashtiriladi: u video uchun tavsiya etiladigan yosh toifasini va xavf turini aniqlab, yakuniy qarorni ota-ona yoki o'qituvchiga qoldiradi. Bu yondashuv noto'g'ri bloklash holatlarining salbiy ta'sirini kamaytiradi va kattalar nazoratini saqlab qoladi.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

Xulosa

Bolalar uchun zararli video va tasvir kontentini aniqlash bitta ma'lumot turini tahlil qilish bilan cheklanib bo'lmaydigan murakkab masala hisoblanadi. O'tkazilgan tahlil asosida quyidagilar asoslandi:

1) vizual, audio, nutq va metama'lumotlar oqimlarini e'tibor mexanizmi asosida birlashtirish har bir video uchun eng muhim ma'lumot manbasini dinamik tanlash imkonini beradi;

2) natijani milliy qonunchilikdagi yosh toifalariga mos ordinal klassifikatsiya ko'rinishida shakllantirish tizimni huquqiy talablar bilan bevosita bog'laydi va qo'pol xatolar ehtimolini kamaytiradi;

3) kalit kadrlarni tanlash va kaskadli tahlil sxemasi tizimni cheklangan hisoblash resurslari sharoitida, real vaqtga yaqin rejimda qo'llash imkonini yaratadi.

Keyingi tadqiqotlarda o'zbek tilidagi bolalar video kontenti bo'yicha yosh toifalari belgilangan ma'lumotlar to'plamini yaratish, taklif etilgan arxitekturani eksperimental baholash va uni maktab hamda oila sharoitida sinovdan o'tkazish rejalashtirilgan.

Foydalanilgan adabiyotlar

1. UNICEF. The State of the World's Children 2017: Children in a Digital World. – New York: UNICEF, 2017.
2. Papadamou K., Papasavva A., Zannettou S., et al. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children // Proceedings of the International AAAI Conference on Web and Social Media (ICWSM). – 2020. – Vol. 14. – P. 522–533.
3. O'zbekiston Respublikasining "Bolalarni ularning sog'lig'iga zarar yetkazuvchi axborotdan himoya qilish to'g'risida"gi Qonuni. 08.09.2017, O'RQ-444-son. – URL: https://lex.uz/docs/3333797
4. O'zbekiston Respublikasi Prezidentining "Sun'iy intellekt texnologiyalarini 2030-yilga qadar rivojlantirish strategiyasini tasdiqlash to'g'risida"gi qarori. 14.10.2024, PQ-358-son. – URL: https://lex.uz/docs/7158604
5. Kiela D., Firooz H., Mohan A., et al. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes // Advances in Neural Information Processing Systems. – 2020. – Vol. 33.
6. Dosovitskiy A., Beyer L., Kolesnikov A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale // International Conference on Learning Representations (ICLR). – 2021.
7. Radford A., Kim J. W., Hallacy C., et al. Learning Transferable Visual Models From Natural Language Supervision // Proceedings of ICML. – 2021. – P. 8748–8763.
8. Bertasius G., Wang H., Torresani L. Is Space-Time Attention All You Need for Video Understanding? // Proceedings of ICML. – 2021.
9. Gemmeke J. F., Ellis D. P. W., Freedman D., et al. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events // Proceedings of IEEE ICASSP. – 2017. – P. 776–780.
10. Radford A., Kim J. W., Xu T., et al. Robust Speech Recognition via Large-Scale Weak Supervision // Proceedings of ICML. – 2023.
11. Kuriyozov E., Vilares D., Gómez-Rodríguez C. BERTbek: A Pretrained Language Model for Uzbek // Proceedings of the SIGUL 2024 Workshop (LREC-COLING 2024). – 2024.
12. Musaev M., Mussakhojayeva S., Khujayorov I., et al. USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments // Speech and Computer (SPECOM 2021). Lecture Notes in Computer Science. – Springer, 2021.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

13. Baltrušaitis T., Ahuja C., Morency L.-P. Multimodal Machine Learning: A Survey and Taxonomy // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 41, No. 2. – P. 423–443.
14. Cao W., Mirjalili V., Raschka S. Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation // Pattern Recognition Letters. – 2020. – Vol. 140. – P. 325–331.
15. Teerapittayanon S., McDanel B., Kung H. T. BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks // Proceedings of the 23rd International Conference on Pattern Recognition (ICPR). – 2016. – P. 2464–2469.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.