DOI 10.31891/2307-5732-2022-309-3-55-60
УДК 637.5.02

МЕЛЬНИКОВА Н. І.
Національний університет "Львівська Політехніка"
https://orcid.org/0000-0002-2114-3436
e-mail: melnykovanatalia@gmail.com
ПОБЕРЕЙКО П. Б.
Національний університет "Львівська Політехніка"
https://orcid.org/0000-0002-8884-1255
e-mail: pobereyko.petro26@gmail.com

ДОСЛІДЖЕННЯ МЕТОДІВ ПОШУКУ КЛЮЧОВИХ КАДРІВ У ВІДЕОПОТОЦІ З ВИКОРИСТАННЯМ НЕЙРОННИХ МЕРЕЖ ДЛЯ СИСТЕМ ПОШУКУ

У роботі викладено порівняльний аналіз сучасних досліджень в області аналізу відеоконтенту і на його основі встановлено, що ефективними методами аналізу цих даних є методи визначення ключових кадрів у відеопотоці. Особливо цінними є методи порівняння та пошуку збігів кадрів (фрагментів), а саме: методи пошуку послідовностей (виявлення об'єктів чи певних дій на кадрах); методи класифікації (визначення вмісту кадрів та розподілення їх до певних категорій); методи декодування кадрів (опис характеристик конкретного зображення) та методи виявлення аномалій у відеопотоці (пошук об'єктів, символів, які є унікальними властивостями фрагмента відносно інших). Особливо перспективними є методи засновані на технологіях машинного навчання, реалізація яких полягає у моделюванні тимчасових залежностей змінного діапазону з використанням загорткових нейронних мереж та функцій із спеціальними механізмами "уваги". Показано, що саме розвиток цих методів сприяє стрімкому розвитку інформаційних систем, за допомогою яких можна успішно здійснити аналіз відеоконтенту та розпізнати його оригінал.

Ключові слова: ключові кадри, нейронні мережі, навчання без вчителя, міра подібності

Nataliia MELNYKOVA, Petro POBEREIKO
Lviv Polytechnic National University

RESEARCH OF METHODS OF SEARCHING KEY FRAMES IN VIDEO FLOW WITH THE USE OF NEURAL NETWORKS FOR SEARCH SYSTEMS

The paper presents a comparative analysis of current research in the field of data analysis in the format of video content and regarding it, that effective methods of analysis of these data are methods of search keyframes in the video stream. The analysis shows that the application of a method of processing visual data is determined by the structure of this data. Therefore, in order to simplify their analysis, they were divided into the following categories: consistent comparison; global comparison, based on clustering, and those that use events or objects. Especially valuable are the methods of comparing and matching matches (fragments), namely: methods of sequence search (detection of objects or certain actions on frames); methods of classifications (determining the content of personnel and their distribution to certain categories); frame decoding methods (description of the characteristics of a particular image) and methods for detecting anomalies in the video stream (search for objects, characters that are unique properties of the fragment relative to others). It shows that the most optimal of the considered methods there are methods that are based on technologies of artificial intelligence and machine learning. And also shows the difference and efficiency of deep learning methods in relation to conventional methods. Particularly promising are the methods, the implementation of which is to model the temporal dependences of the variable range using convolutional neural networks and functions with special attention mechanisms. Methods that use an Actor-Critic model embedded in a Generative adversarial network have also demonstrated their effectiveness. It is shown that the development of these methods contributes to the rapid development of information systems with which you can successfully analyze video content and recognize its origin.

Keywords: keyframes, neural network, unsupervised learning, similarity measure, generative adversarial networks

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Зі стрімким розвитком Інтернету та сучасних технологій Інтернет мережі зображення та відео стали одним із найефективніших способів передачі даних. Це зумовлено тим, що людський мозок обробляє цю інформацію значно краще, ніж текстовий контент, наприклад, середньо статистична людина сприймає приблизно 80 % інформації через зір, а у випадку читання – лише 20 %. Тому знаходження важливої інформації у великих за обсягом текстах потребує значних затрат часу. Відео контент покликаний швидко та ефективно донести до людей те, що вони не бачать у текстових контентах. Тому сьогодні в Інтернет мережах поширено дуже багато відео контенту на найрізноманітнішу тематику, і їх одночасно можуть переглядати мільйони людей у різних місцях Землі. Окрім цього зазначимо, що на різних стрімінгових сервісах окрім відео оригінала, зазвичай знаходяться і його окремі фрагменти, які у багатьох випадках є спотвореними або пошкодженими. Тому, з огляду на зазначене, надзвичайно актуальним є дослідження присвячене удосконаленню відомих та розробленню нових методів аналізу відео контенту для пошуку відео оригінала, а не його спотворених чи пошкоджених копій та окремих фрагментів. Необхідність вирішення цієї проблеми спонукала багатьох науковців до розроблення різноманітних підходів та алгоритмів для проведення аналізу вмісту, класифікації, пошуку та узагальнення візуального контенту. Багатьом із них вдалося не лише побудувати ці алгоритми, їм вдалося також виявити і основні принципи на яких вони повинні базуватися. Зокрема, вони показали, що потужним інструментом для удосконалення відомих та побудови нових алгоритмів є методи штучного інтелекту та машинного навчання. Показали, що саме

розвиток цих методів сприяє стрімкому розвитку інформаційних систем за допомогою яких можна успішно здійснити аналіз відео контенту та розпізнати його оригінал. Виявили, що аналіз відео потоку доцільно проводити за даними результатів порівняння та пошуку збігів у послідовності кадрів (фрагментів). Подібністю кадрів можуть слугувати об'єкти на сцені чи палітра кольорів. Встановили, що в основі сучасних інформаційних системах аналізу відео контенту використовуються здебільшого так звані алгоритми для визначення ключових кадрів. І на основі аналізу способів їх практичного застосування довели, що вони дають змогу сформувати репрезентивну вибірку стислого огляду, яка забезпечує найбільш точне відображення відео змісту. Окрім цього, показали, що ефективне узагальнення відео із виділенням опорних кадрів значно полегшує перегляд і навігацію у великих колекціях відео в Інтернеті, що істотно збільшує залучення глядачів і споживання контенту, а також слугує основним кроком у системах пошуку за фрагментами.

Таким чином, для вирішення поставленої проблеми надзвичайно важливим є аналітичний огляд підходів та методів аналізу відео та фрагментів, які сформовані на методах штучного інтелекту та машинного навчання, а також пошук існуючих концепцій узагальнення візуального контенту. Для проведення такого огляду необхідно проаналізувати такі методи: пошук послідовностей (виявлення об'єктів чи певних дій на кадрах), класифікація (визначення вмісту кадрів та розподілення їх до певних категорій), декодування кадрів (опис характеристик конкретного зображення), виявлення аномалій (пошук об'єктів, символів, які є унікальними властивостями фрагмента відносно інших).

Аналіз досліджень та публікацій

Для дослідження методів пошуку ключових кадрів у відеопотоці проведемо аналіз функцій аналізу форм, кольорів та оптичного потоку.

Загалом методів для пошуку кадрів у відеопотоці є доволі багато. Усі вони мають важливе практичне значення. В одних випадках доцільно використовувати одні методи, а в інших випадках – інші. Застосування того чи іншого методу опрацювання візуальних даних визначається структурою цих даних. Тому з метою спрощення їх аналізу поділимо їх умовно на такі категорії: послідовне порівняння; глобальне порівняння, на основі кластеризації та ті які використовують події чи об'єкти.

Методи категорії послідовного порівняння використовуються в основному для вирішення задач визначення подібності між кадрами відеопотоку. Їх сутність полягає у порівнянні кожного наступного кадру із попереднім. Це порівняння доволі часто використовує кольорову гістограму, поскільки відмінність кольорів кадрів у відеофрагменті, що відображають певну подію, здебільшого є незначною (мінімальною). Однак, методи цієї категорії мають і деякі недоліків. Основним із них – це великі затрати часу на опрацювання кадрів відеопотоку та велика похибка у випадку оброблення даних із шумом.

Для стислого огляду сучасних методів, які найбільш наближені до вирішення задач, які повинна виконувати система узагальнення візуальних даних розділимо їх на два типи: звичайні методи та методи із використанням машинного навчання.

Звичайні методи базуються на певній цільовій функції та інструкціях, які не змінюються з часом. Переважна більшість цих функцій використовує в основному конвеєрну сегментацію. Зазвичай такі підходи виділяють характеристики SIFT та оптичний потік. За допомогою дескрипторів SIFT виділяються ключові точки та локальні характеристики кадрів. Для прикладу у роботі "Robust voting algoritm based on labels behavior for video copy detection" [1] виділяють опорні точки кадру (за допомогою детектора Харріса) і відстежують їхні позиції протягом усього відео. Після чого формують безмежно велику кількість траєкторій цих точок. Для знаходження подібностей використовують алгоритми нечіткого пошуку. Цей метод суттєво спрощує локалізацію нечітких дублікатів фрагментів та дає змогу провести узагальнення для відеопотоку. Проте він є дорогим відносно ресурсів для виділення ключових точок на зображеннях. А факт того, що траєкторії точок чутливі до руху камери, роблять алгоритм оптимальним лише у випадках пошуку точних копій відео. Іншим, не менш важливим типом звичайних підходів є методи, які базуються на основі кластеризації, характерною особливістю яких є те, що число кластерів, як правило, повинно задаватися апріорно. Вагомими, щодо удосконалення та розроблення нових методів цього типу, є роботи Tang H. [2], де для визначення ключового кадру у відеопотоці вперше запропоновано використовувати ентропію та щільність групування зображення для розпізнавання жестів руки. Не менш цінними є також і дослідження Vazquez R., який у роботі [3] запропонував знаходити ключові кадри за допомогою методу заснованого на спектральній кластеризації. Унікальність цього методу полягає у тому, що для його практичної реалізації не потрібно проводити обчислення міри подібності із виділенням спільних ознак для двох зображень. Актуальними є також і роботи Wang Y. [4], у яких синтезовано метод, ідея якого полягає у розрахунку матриці подібності та визначенні кластерів на її основі з подальшим пошуком і вилученням ключових кадрів. Характерною особливістю цього методу є те, що у ньому усунено обмеження на вибір одного кадру на кластер. Кількість кадрів, що відбираються з одного класу, залежить від продовження та складності змісту сцен. Порівняно з методами пошуку ключових кадрів на основі класифікації цей метод з обчислювальної точки зору є значно простішим. Основним його недоліком є те, що висновок про значущість кадрів робиться вихідним із положення, що на значущих сценах довше фокусується камера. Коли основні кадри відбираються з довгих послідовних кадрів у кластері, середній кадр кожної послідовності вважається значущим, що нагадує найбільш ранні підходи до пошуку ключових кадрів.

Для усунення обмежень та покращення ефективності звичайних методів розглянемо роботи, які присвячені синтезу моделей із використанням глибокого машинного навчання (багатошарові нейронні мережі). Одним із таких досліджень є робота Янга [5], де представлено двонаправлену короткочасну пам'ять Bi-LSTM, яку використовували налаштовані мережі Graph Attention Networks (GAT) для виділення опорних кадрів із відео. Ця мережа дозволила авторам перетворювати візуальні функції зображення у функції вищого рівня використовуючи механізм трансформації контекстний методів (CFT).

Наступною важливою науковою працею є робота Махасені [6], де вперше було застосовано генеративну змагальну мережу (GAN). Результатом дослідження була розроблена мережа підсумків для навчання мінімізації відстані між навчальними відео та розповсюдження їх узагальнень. Модель складалася з автокодера LSTM як підсумовувача та іншого LSTM як дискримінатор. Таким чином, підсумовувач LSTM був навчений вводити в оману дискримінатор, що змусило підсумовувач отримувати кращі зведення. Така модель продемонструвала високі результати під час експериментів.

У роботі Янг [7] описано особливості використання локальних функцій, зокрема функцій масштабно інваріантного перетворення ознак (SIFT) та функцій, отриманих від згорткової нейронної мережі (CNN). Окрім цього запропоновано інтегровану схему узгодження функції, яка інтегрує узгодження функцій SIFT і функцій CNN між зображеннями для виявлення часткових копій зображень. У цій схемі автори реалізували зіставлення функцій SIFT на основі моделі візуальних слів, щоб виявити потенційні дублікати пар регіонів між зображеннями, а потім зіставили характеристики CNN цих регіонів, витягнутих з згорткового шару мережі CNN для обчислення схожості образів. Також нещодавно у роботі Чжоу та ін. (2018) [8] використали мережу глибокого узагальнення за допомогою навчання з підкріпленням. Вони представили підхід DQSN на основі RL для узагальнення відео. На основі експерименту з навчанням без нагляду та з вчителем, їхня цільова функція дає можливість аналізувати семантику, використовуючи лише прості для отримання мітки на рівні відео.

Ще декілька важливих досліджень у цій області знань ґрунтуються на результатах аналізу просторово-часових зв'язків між частинами відео та використанні сучасних типів нейронних мереж, зокрема згорткових мереж. Важливими у цій області знань є роботи [9], авторами яких, з використанням згорткових нейронних мереж LSTM, синтезовано архітектурну модель декодера і кодера. Ця модель дає змогу змоделювати просторово-часовий зв'язок між фрагментами відео. Запропонований на її основі алгоритм ефективно генерує візуальну різноманітність ключових кадрів та механізми їх знаходження.

У 2019 році Юань опублікував роботу [10], в якій розробив метод який дозволяє навчати систему, і створювати нове представлення використовуючи стратегію злиття. Для оцінки серії послідовних кадрів автори застосовують функції втрат. У цьому ж році Елфекі [11] та автори робіт [12] побудували модель у якій вдало поєднали CNN і Gated Recurrent Units (один із типів RNN). Характерною особливістю цієї моделі є генерування векторів для оцінки рівня активності і важливості кожного кадру у відеопотоці.

Загалом автори досліджень вважають, що методи для узагальнення та пошуку ключових кадрів у відео є найбільш ефективними лише у випадках використання машинного навчання.

Формулювання цілей статті

Метою роботи є огляд та пошук ефективних підходів, які виконують узагальнення відео на основі машинного навчання для систем пошуку.

Виклад основного матеріалу

Порівняльний аналіз літературних джерел показав, що на сьогодні існує низка методів, які використовуються у сучасних дослідженнях для узагальнення відео і які на наш погляд є найбільш продуктивними у випадку їх застосування у системах пошуку за фрагментами.

Враховуючи сучасний стан предметної області ми переконані, що для систем пошуку відео за фрагментами найбільш оптимальними є методи, які використовують машинне навчання без вчителя. Вони дають змогу розробляти моделі, які є незалежним під час обробки даних від втручання користувачів.

Більшість підходів, які базуються на навчанні без вчителя використовують правило, згідно якого репрезентативна вибірка повинна допомогти користувачеві чи іншим компонентам системи зробити висновок відносно вихідного відеоконтенту. В цьому контексті методи зазвичай використовують GAN. Це узагальнення складається із селектора ключових кадрів (оцінюють важливість) та генератора (створює звіт). Особливістю застосування цих підходів є те, що навчання відбувається шляхом реконструкції відео на основі резюме [13]. Розглянемо концепцію цього підходу під час навчання, сутність якої наочно відображена на рис. 1. Зазвичай суматор складається із селектора ключових кадрів, який оцінює їх важливість та генератора, основним завданням якого є реконструкція відео. Реконструюванню відео разом із оригінальними даними в якості вхідних даних навчає дискримінатор (результатом якого є надання оцінки подібності). Процес навчання виглядає наступним чином: суматор намагається обманути дискримінатор під час того, як він намагається навчитись знаходити різницю між узагальненим звітом (ключовими кадрами) та оригінальним відео. Результатом навчання є стан дискримінатора, коли він не може знайти різницю (помилка класифікації приблизно однакова як для реконструйованого так і для оригінального відео).

Для вище продемонстрованої концепції описано процес реконструкції відео на основі резюме. Для того щоб модель виконувала зворотній процес, який в подальшому можна застосовувати, як компонент у

системах пошуку сучасні дослідження пропонують розширений підхід. У цьому випадку використовують пару дискримінаторів. Тоді селектор кадрів (двонаправлений LSTM) знаходить ключові кадри шляхом моделювання тимчасової залежності між кадрами. Далі на основі результатів селектора проводиться оцінка, яка складається із двох GAN [14]. Перший використовується для навчання відносно реконструкції відео на основі узагальнення, другий вчиться виконувати дії навпаки – від оригіналу до ключових кадрів. Розглянемо детальніше функціонування другого модуля GAN. На сьогодні запропоновано декілька варіантів процесу навчання. Одним з них є використання "актор-критик" моделі, яка розглядає задачу вибору опорних кадрів, як задачу генерації послідовностей. "Актор" та "Критик" [15] перебувають в стані постійного обміну результатами, назвемо це змаганням. Стратегія навчання дозволяє критику засвоїти ціннісну функцію і актору ефективно дізнаватися про політику вибору ключових фрагментів. Це допомагає корегувати вибір правильних значень для параметрів моделі. Високу ефективність демонструють також і GAN, які базуються на самоконтролі [16]. У цьому випадку генератор намагається передбачити оцінку важливості на рівні кадру для кожного кадру та створює зважені характеристики кадру на основі тимчасових уявлень необроблених характеристик кадру. Потім необроблені ознаки кадру та зважені ознаки кадру розглядаються як реальні та фальшиві вхідні дані для дискримінатора щоб провести порівняння. Слід врахувати що для захоплення тимчасових залежностей на великій відстані по всьому проміжку відеопослідовності використовується BiLSTM [17]. Недоліком цих концепцій досі залишається нестабільність процесу тренування [18], хоча і останні дослідження намагаються постійно розширювати та покращувати моделі GAN для мінімізації обмежень критеріїв оцінки кадрів.

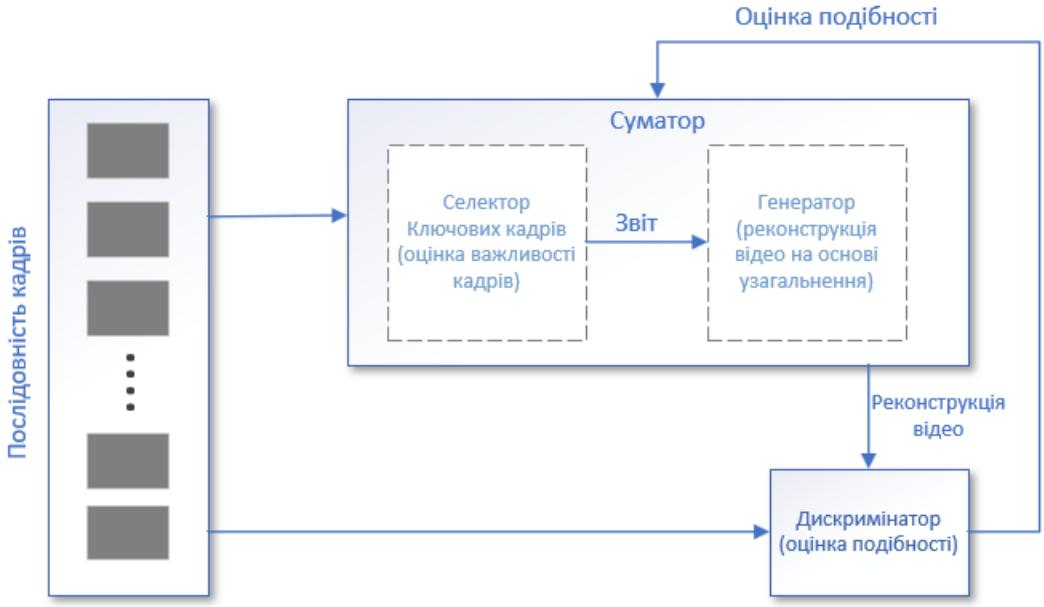

Рис. 1. Високорівневе представлення моделі навчання без вчителя для аналізу узагальнення

Загальна концепція навчання із підкріпленням базується на функції "винагород". Вона поділяється на декілька етапів. На основі вхідних даних у вигляді послідовності кадрів створюється звіт шляхом прогнозування оцінки важливості на рівні кожного кадру. Створений звіт надсилається наступному модулю, який відповідає за кількісну оцінку існуючих (наперед вибраних) характеристик за допомогою створених вручну функцій "винагород". Потім підраховані бали об'єднуються, щоб сформувати загальне значення "винагороди", яке використовує суматор для навчання. Одним із підходів є тренування суматора, таким чином, щоб він створював різноманітні та репрезентативні вибірки ключових кадрів з використанням нагороди за різноманітність. Ця нагорода (бали або коефіцієнт) вимірює відмінність між вибраними ключовими кадрами, а нагорода за репрезентативність обчислює відстань (що виражає візуальну схожість) між обраними кадрами з кадрів відео, що залишилися. Також для вирішення проблем LSTM відносно затухаючого (vanishing) та вибухаючого (exploding) градієнта використовують незалежні рекурентні нейронні мережі (IndRNN) [19] на основі функції активації – Leaky ReLU (Leaky Rectified Linear Unit) [20].

Загальна концепція аналізу руху ключових візуальних об'єктів базується на основі автокодувальника, де під час аналізу виконуються кроки, ціллю яких є пошук опорних об'єктів та траєкторій їх руху. Після виконання цих кроків створюються сегментовані фрагменти відео із рухом для кожного об'єкта. На наступному етапі використовується онлайн модель автокодувальника об'єкта (Stacked Sparse LSTM Auto-Encoder) [21] для запам'ятовування попередніх станів руху об'єкта шляхом постійного оновлення адаптованої мережі автокодувальника. Останній крок відповідає за реконструкцію фрагментів.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Викладено порівняльний аналіз методів опрацювання даних, поданих у форматі відеоконтенту і на його основі визначено, що найбільш оптимальними методами є методи засновані на алгоритмах штучного інтелекту та машинного навчання. Особливо перспективними є методи, практична реалізація яких полягає у моделюванні тимчасових залежностей змінного діапазону з використанням згорткових нейронних мереж та функцій із спеціальними механізмами "уваги".

Показано також, що для опрацювання "неконтрольованих" відео доцільно використовувати генеративні змагальні мережі у поєднанні із механізмами "уваги" та "актор-критик".

Виявлено, що спрощення процедури адаптації вимог до даних у різних доменах і сценаріїв їх застосування є можливим лише за умови вдосконалення моделей узагальнення, які навчаються без нагляду.

Література

1. Julien L., Olivier B., Valérie G., Nozha B. Robust voting algorithm based on labels of behavior for video copy detection. 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, October 23-27. 2006. DOI: https://doi.org/10.1145/1180639.1180826

2. Tang H., Liu H., Xiao W., Sebe N. Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion. Neurocomputing. 2019. Volume 331. P. 424-433. DOI: https://doi.org/10.1016/j.neucom.2018.11.038.

3. VáZquez–MartíN R., Bandera A. Spatio–temporal feature–based keyframe detection from video shots using spectral clustering. Pattern Recognit. Lett. 2013. Volume 34. P. 770–779.

4. Qu Z., Lin L., Gao T., Wang Y. An improved keyframe extraction method based on HSV color space. Journal of Software. 2013. Vol. 8, Iss. 7. P. 1751–1758.

5. Yang H., Wang B., Lin S., Wipf D., Guo M., Guo B. Unsupervised extraction of video highlights via robust recurrent auto–encoders. IEEE International Conference on Computer Vision, Santiago, Chile, 7-13 December 2015. P. 4633–4641.

6. Mahasseni B., Lam M., Todorovic S. Unsupervised video summarization with adversarial LSTM networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 21-26 July 2017. P. 1–10.

7. Jian M., Zhang S., Wu L., Zhang S., Wang X., He Y., Deep key frame extraction for sport training. Neurocomputing. 2019. Volume 328. P. 147–156.

8. Zhou K., Xiang T., Cavallaro A. Video Summarisation by Classification with Deep Reinforcement Learning. British Machine Vision Conf. (BMVC). 2018.

9. Lal S., Duggal S., Sreedevi I. Online video summarization: Predicting future to better summarize present. In IEEE Winter Conf. on Applications of Computer Vision (WACV). IEEE. 2019. P. 471–480.

10. Yuan Y., Li H., Wang Q. Spatiotemporal modeling for video summarization using convolutional recurrent neural network. IEEE Access. 2019. Volume 7. P. 676–685.

11. Elfeki M., Borji A. Video Summarization Via Actionness Ranking. IEEE Winter Conf. on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7-11 January 2019. P. 754–763.

12. Cho K., B. van Merrienboer, Gulcehre C., Bahdanau C., Bougares F., Schwenk H., Bengio Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, October 2014. P. 1724–1734.

13. Mahasseni B., Lam M., Todorovic S. Unsupervised Video Summarization with Adversarial LSTM Networks. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2017. P. 2982–2991

14. Apostolidis E., Metsai A., Adamantidou E., Mezaris V., Patras I. A stepwise, label–based approach for improving the adversarial training in unsupervised video summarization. In Proc. of the 1st Int. Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV '19). New York, NY, USA: ACM. 2019. P. 17–25.

15. Apostolidis E., Adamantidou E., Metsai A., Mezaris V., Patras I. AC–SUM–GAN: Connecting Actor–Critic and Generative Adversarial Networks for Unsupervised Video Summarization. IEEE Trans. on Circuits and Systems for Video Technology. 2020. P. 1.

16. He X., Hua Y., Song T., Zhang Z., Xue Z., Ma R., Robertson N., Guan H. Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks. In Proc. of the 27th ACM Int. Conf. on Multimedia (MM '19). New York, NY, USA: ACM. 2019. P. 2296–2304.

17. Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks. 2005. Volume 18. P. 5-6. DOI: https://doi.org/10.1016/j.neunet.2005.06.042.

18. Zhou K., Qiao Y. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity–Representativeness Reward. Proc. of the 2018 AAAI Conf. on Artificial Intelligence. 2018.

19. Li S., Li W., Cook C., Zhu C., Gao Y. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2018. P. 5457–5466

20. Wang L., Xiong Z., Wang Z., Qiao Y., Lin D., Tang X., Van Gool L. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in Europ. Conf. on Computer Vision (ECCV). Cham: Springer International Publishing. 2016. P. 20–36.

21. Zhang Y., Liang X., Zhang D., Tan M., Xing P. Unsupervised object–level video summarization with online motion auto–encoder. Pattern Recognition Letters 2018.