DOI: https://doi.org/10.15276/ict.02.2025.19
УДК 004.93

Контекстна класифікація відео з використанням VideoBERT та адаптерів DA-Ada

Новічонок Марія Сергіївна[1)]
Аспірантка каф. Інформатики
ORCID: https://orcid.org/0009-0007-1787-0546; mariia.novichonok@nure.ua
Машталір Сергій Володимирович[1)]
Д-р техніч. наук, професор каф. Інформатики
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Харківський національний університет радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна

АНОТАЦІЯ

У цій роботі запропоновано архітектуру для задачі контекстної класифікації відео, яка поєднує сильні сторони попередньо натренованого відео-мовного енкодера VideoBERT, адаптаційного модуля DA-Ada (Domain-Aware Adapter, доменно-орієнтований адаптер) та авто-регресивного трансформерного декодера. Основна мета полягає в побудові системи, здатної формувати текстові описи дій у відео з високим ступенем узагальнення до нових доменів. Архітектура розроблена з урахуванням вимог до масштабованості, гнучкої адаптації та зменшення витрат часу на дотренування моделі у майбутньому. Вхідне відео розбивається на послідовність кадрів, кожен кадр перетворюється у вектор ознак за допомогою ResNet-50, попередньо натренованого на ImageNet. Далі вектори кадрів проєктуються в простір візуальних токенів та передаються в модуль VideoBERT. Цей енкодер, побудований на основі трансформерної архітектури BERT, виконує контекстуалізацію ознак по всій відеопослідовності, моделюючи довготривалі часові залежності між кадрами. Усі параметри VideoBERT залишаються замороженими, що зменшує потребу в ресурсах при донавчанні. Після енкодингу кожне представлення передається в адаптаційний модуль DA-Ada, який складається з двох паралельних гілок: DIA (Domain-Invariant Adapter, доменно-інваріантний адаптер) та DSA (Domain-Specific Adapter, доменно-специфічний адаптер). DIA навчається фільтрувати загальні, інваріантні ознаки, характерні для більшості відео. DSA фокусується на виявленні ознак, притаманних певному домену (наприклад, побутові сцени, індустріальні об’єкти, тощо). Вихідні представлення обох адаптерів поєднуються за допомогою скалярного коефіцієнта, що визначає баланс між універсальністю і спеціалізацією. Результатом цього злиття є послідовність адаптованих векторів, яка подається до трансформера для генерації опису дій. Генерація здійснюється трансформерним декодером, який складається з шести шарів, що включають механізм самоуваги (self-attention) для роботи з частково сформованим текстом, механізм перехресної уваги (cross-attention) до відео-контексту, а також стандартні блоки з прямим розповсюдженням (feed-forward). Починаючи з токена <BOS>, декодер поетапно формує текстовий опис дії, завершуючи процес при генерації токена <EOS> або досягненні граничної довжини.

Запропонована архітектура забезпечує модульність, обмежену кількість параметрів, що підлягають донавчанню, та можливість використання і різних доменах. У подальшій роботі планується реалізація повного циклу навчання моделі на базі датасету Something-Something V2.

Ключові слова: контекстна класифікація відео; енкодер; декодер; нейронні мережі; адаптер; трансформер

Задача контекстної класифікації відео не втрачає своєї актуальності як завдяки можливостям застосування у різноманітних сферах, так і через зростання кількості відео, яку щодня генерує людство, і представляє інтерес для аналізу. Існуючі методи аналізу відео вже допомагають автоматизувати процеси у медицині, будівництві, виробництві, алгоритмах соціальних мереж тощо. Дана стаття є продовженням досліджень для отримання архітектури, яка може проаналізувати відеопослідовність певного виробництва чи установи, та надати детальний опис процесів, які виконують співробітники, що у подальшому стане основою для проведення реінжинірингу бізнес-аналітиком. У попередніх публікаціях вже було висвітлено особливості [1], які мають бути враховані під час вибору методів аналізу відео, а саме визначення структури виконуваних процесів, їх послідовності, типу взаємодії між людиною і предметом або іншою людиною. Також, оскільки замовник реінжинірингу може бути зацікавленою особою з будь-якої сфери, дуже важливим стає проблема мультидоменності майбутнього рішення. У реальних застосуваннях моделі комп’ютерного зору часто стикаються з розбіжністю між даними, на яких вони були навчені (source domain), і тими, з якими вони працюють під час розгортання (target domain). Це явище зветься доменним зсувом (domain shift), воно істотно знижує якість прогнозування [2]. Тому задача доменної адаптації та узагальнення залишається критично важливою для побудови систем, здатних ефективно працювати в умовах зміни домену. Саме цій проблемі приділено найбільшу

This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/deed.uk)

увагу у даній роботі. Запропоновано поєднання архітектури енкодеру VideoBERT [3] з набором адаптерів DA-Ada (Domain-Aware Adapter, доменно-орієнтований адаптер) [4], а також трансформерним декодером. Таке рішення дозволить легко дотреновувати модель на новому домені на невеликій кількості відео, або використовувати її одразу без додаткового вдосконалення для отримання гарних результатів на відеопослідовностях, які містять дії, не представлені у датасеті, який використано для тренування.

Огляд пов'язаних робіт

Генерація субтитрів

Для початку зупинимось на питанні генерації субтитрів – автоматичному створенні описів подій на відео природною мовою. Перші методи генерації описів відео ґрунтувалися на ручному або автоматичному виділенні ключових ознак (акторів, дій, об'єктів), після чого застосовувалися шаблони для формування речень. Такий підхід був обмеженим у виразності та масштабованості, оскільки не дозволяв поєднувати між собою довготривалі події або генерувати текст вільно, не орієнтуючись на задані шаблони. У 2015 році значущим стає публікація методу S2VT (Sequence to sequence – Video to Text, Послідовність за послідовністю – Відео до тексту) [5], з використанням «декодер-енкодер» архітектури. S2VT є першою моделлю, що напряму трансформує відео у текстовий опис без використання шаблонів. Оскільки в основі архітектури лежить двонаправлена LSTM-модель, S2VT навчається одночасно і на візуальних представленнях, і на описах до них. Sequence-to-sequence добре працює на відео різної довжини, розуміє послідовність дії у часі, генерує описи, схожі до тих, які може створити людина. Проте, модель не є адаптивною до інших доменів, і показує гарні результати на відео, схожих до тих, які було використано для тренування. Також, S2VT потребує вдосконалення для створення послідовних описів довготривалих сцен.

Мультимодальні трансформери

Трансформери є мультимодальними моделями, які поєднують зображення (або відео) та текст у спільному просторовому представленні. До найвідоміших належать CLIP, BLIP та Frozen [6]. Ці моделі зазвичай тренуються методом самонавчання (self-supervised) на парах *[зображення, текст]*, що дозволяє їм досягати гарних результатів на задачах класифікації та генерації описів. Наприклад, CLIP успішно переноситься на нові домени без донавчання, що зробило його стандартом у багатьох практичних застосуваннях. Однак, попри свою універсальність, ці моделі мають обмежене розуміння відео як послідовності пов'язаних між собою подій. CLIP та BLIP переважно працюють з окремими кадрами або короткими кліпами, і не враховують часовий контекст між фреймами. Frozen поєднує заморожені CLIP-подібні візуальні енкодери з мовною моделлю GPT, але також не має вираженого механізму для обробки динаміки відео. Усі ці моделі переважно оперують статичними сценами і не пристосовані для глибокого аналізу дій, послідовностей руху або причинно-наслідкових зв'язків у відео. Це обмеження стало однією з причин, чому в дослідженнях генерації описів дій фокус поступово змістився на моделі, які спеціально враховують темпоральну структуру відео, такі як VideoBERT, ClipBERT, ActionFormer та інші.

Доменна адаптація (Domain Apaptation, DA)

Однією з ключових проблем у генерації описів до відео є доменна спеціалізація моделей. Багато сучасних архітектур, включаючи VideoBERT та CLIP-похідні моделі, демонструють гарні результати на даних, схожих на тренувальні, однак суттєво втрачають якість при перенесенні на нові сцени, жанри або типи дій (наприклад, з побутових відео на індустріальні). Одним з ефективних рішень цієї проблеми є використання адаптерів (adapter modules) – невеликих параметризованих блоків, які вбудовуються у архітектуру трансформерів та дозволяють моделі пристосовуватись до нових умов без повного донавчання всієї мережі.

Серед відомих адаптерів можна виділити VL-адаптер та READ (Recurrent Adapter with Partial Alignment). Загалом, їх існує велика кількість модифікацій, але в основі лежить саме

здатність донавчати модель, змінюючи лише малий відсоток параметрів (наприклад, для VL-адаптер показав ефективність при донавчанні лише 3 % параметрів [7]).

У даній роботі увагу буде зосереджено на DA-Ada адаптері. Він складається з двох частин:

1) Domain-Invariant Adapter (DIA, доменно-інваріантний адаптер) – для універсальних ознак;

2) Domain-Specific Adapter (DSA, доменно-специфічний адаптер) – для локальних ознак, характерних для домену. Автори підходу зосередились саме на задачі зміщення домену, вирішення якої досягається шляхом розділення ознак на спільні для усіх доменів та специфічні, та поєднанням адаптерів всередині моделі шляхом додавання механізму уваги. DA-Ada так само дозволяє провести fine-tuning (тонкого налаштування) моделі, завдяки чому легше масштабується на нові домени.

VideoBERT та генерація описів до відео

Одним із ключових кроків у напрямку відео-мовного представлення стала модель VideoBERT, яка адаптувала архітектуру двонаправленого трансформера BERT до мультимодального контексту. На відміну від попередніх підходів, що покладалися на окрему обробку відео та тексту з подальшим злиттям, VideoBERT реалізує спільне просторове навчання відео- та текстових токенів у єдиній моделі. Модель навчається у режимі самонавчання, використовуючи задачу маскування токенів як для мовних, так і для візуальних послідовностей. Це дозволяє VideoBERT вивчати узагальнені семантичні подання без використання розмічених даних, які надалі можуть бути адаптовані до різних задач, таких як класифікація відео та генерація описів.

Автори статті [3] також продемонстрували застосування VideoBERT у задачі генерації описів відео шляхом інтеграції декодера на основі трансформера. У рамках цієї архітектури, VideoBERT слугує джерелом семантичних представлень відео, які потім передаються у декодер, навчений на парі *[відео, опис]*. Такий підхід дозволив здійснювати контекстно усвідомлену генерацію описів дій, зроблених природною мовою, базуючись на знаннях, здобутих під час попереднього самонавчання. Проте, незважаючи на продемонстровані переваги, ця архітектура все ще обмежена доменом тренування: модель демонструє найкращі результати на відео схожого типу, але має труднощі з узагальненням на нові сцени.

Запропонований метод класифікації відео

Архітектура досліджуваної моделі розроблена для вирішення задачі контекстної класифікації відеопослідовностей з можливістю підтримки різних доменів. На вхід системи подається відео, яке перетворюється на візуальні токени та послідовно обробляється енкодером VideoBERT, адаптером DA-Ada, шаром злиття ознак та трансформерем для декодингу. Розглянемо послідовно кожен модуль, а також набір даних, який обрано для навчання моделі.

Something-Something V2

Для навчання моделі обрано датасет Something-Something V2, який налічує 220,847 коротких відео з короткими послідовними пов’язаними діями (наприклад, приготування страви). Датасет розділено на тренувальний та тестовий набір розмірами 168,913 та 24,777 відповідно. Загалом, відео мають 174 унікальних типи дій у форматі «Робити [щось] з [чимось]», наприклад «Putting [something] onto [something]» [8].

Попередня обробка відео

Для початку необхідно перетворити відео у послідовності візуальних токенів фіксованої довжини. Вхідні відеофайли нарізаються на кадри з частотою 2 кадри на секунду з максимальною довжиною 16 кадрів на один відеофрагмент. Відео довшої тривалості обрізаються, коротші – доповнюються нульовими кадрами із відповідною маскою. Кожен кадр $f_i \in R^{H\times W\times 3}$ пропускається через попередньо навчену візуальну модель, яка виконує перетворення у вектор ознак. У даному досліджені обрано модель згорткової мережі ResNet-50, попередньо навченої на датасеті ImageNet. Вектор ознак має розмірність 2048 і отримується на виході з останнього згорткового шару моделі.

Кожен вектор ознак $z_i \in R^{2048}$ проектується у простір з розмірністю $\mathrm{D} = 768$ за допомогою лінійного шару:

$$\mathrm{x_i} = \mathrm{W_{proj}} \cdot \mathrm{z_i} + \mathrm{b_{proj}},\ \mathrm{x_i} \in \mathrm{R}^{768},$$

де $W_{\text{proj}} \in R^{768\times2048}$ – матриця ваг, яка виконує лінійне перетворення з простору ResNet-виходів у простір трансформеру, $b_{\text{proj}} \in R^{768}$ – вектор зміщення (bias), доданий після множення на ваги.

Отримана послідовність векторів $X = \{x_1, x_2, \dots, x_T\}$ формує базу для подальшої токенізації. До кожного токена $\mathrm{x_i}$ додається позиційне векторне представлення $\mathrm{p_i}$, що враховує порядок кадру в послідовності. Крім того, вводиться темпоральна маска $\mathrm{m} \in \{0,1\}^{\mathrm{T}}$, яка позначає справжні кадри та нульові. Ця маска використовується у self-attention модулях для уникнення обробки неінформативних позицій.

Фінальне представлення кожного кадру є:

$$x_i' = x_i + p_i,$$

де $x_i \in R^D$ – токен, отриманий після проєкції з ResNet, $p_i \in R^D$ – позиційне перетворення (embedding, ембеддінг), що враховує порядок кадру, x_i' – токен з позиційною інформацією. Вся послідовність подається до VideoBERT як $\mathrm{X}' = \{\mathrm{x}_1', \mathrm{x}_2', \dots, \mathrm{x}_\mathrm{T}'\}$ разом із темпоральною бінарною маскою m.

Цей етап забезпечує стабільне і структуроване представлення відеофрагмента у форматі, сумісному з трансформерною обробкою, та гарантує збереження часової інформації в межах послідовності.

Енкодер VideoBERT

Модуль VideoBERT слугує основним енкодером у запропонованій архітектурі. Його функція полягає в обробці послідовності візуальних токенів і формуванні контекстуалізованих представлень, які зберігають як локальні, так і глобальні часово-просторові залежності у відео. VideoBERT базується на трансформерній архітектурі BERT, адаптованій для обробки відеофрагментів.

У реалізації використовується 12-шаровий трансформер із багатоголовковим механізмом самоуваги (multi-head self-attention) та двошаровими блоками прямого розповсюдження.

Кожен шар включає наступні компоненти:

– механізм самоуваги для моделювання залежності між токенами відео, включаючи довготривалі зв'язки між кадрами;

– шари залишкового навчання та нормалізації для забезпечення стабільності градієнтів і полегшення оптимізації;

– feed-forward мережа для трансформації кожного токену окремо.

На вхід VideoBERT подається послідовність візуальних токенів $X' = \{x_1', x_2', \dots, x_T'\}$, а також темпоральна маска $m \in \{0,1\}^T$.

Модель обчислює контекстуалізоване представлення кожного токена:

$$H = \mathrm{VideoBERT}(X', m) = \{h_1, h_2, \dots, h_T\}, \quad h_i \in R^D,$$

де $H = \{h_1, h_2, \dots, h_T\}$ – вихідна послідовність векторів ознак, яка містить просторово-часовий контекст усіх токенів після обробки у VideoBERT, D – розмірність простору.

Ці вектори передаються до адаптаційного модуля DA-Ada для подальшого контекстного розділення ознак.

Адаптер DA-Ada

Для підвищення здатності моделі до узагальнення на нові домени, в архітектуру інтегрується модуль DA-Ada (Domain-Aware Adapter). DA-Ada – це адаптаційний блок, розроблений для селективного моделювання як універсальних, так і специфічних до домену ознак у трансформерному середовищі. DA-Ada розміщується після основного енкодера VideoBERT і модифікує його вихідні представлення перед їх подачею у декодер. Це дозволяє

зберігати знання з попереднього домену і зменшувати обсяг донавчання моделі при адаптації під інші домени.

DA-Ada складається з двох підкомпонентів – DIA та DSA. DIA реалізується як bottle-neck MLP (Multi-Layer Perceptron, мультишаровий персептрон), який обчислює узагальнені ознаки, стабільні незалежно від конкретного відео-домену.

Даний блок можна представити у вигляді:

$$\mathrm{DIA}(h_i) = W_2^{\mathrm{inv}} \cdot \sigma(W_1^{\mathrm{inv}} \cdot h_i + b_1^{\mathrm{inv}}) + b_2^{\mathrm{inv}}, \quad h_i \in R^D,$$

де $h_i \in R^D$ – вхідний вектор ознак для i-го токена з виходу VideoBERT, $W_1^{\mathrm{inv}} \in R^{d \times D}$ – матриця ваг першого шару адаптера, що стискає вектор у простір меншої розмірності d, $b_1^{\mathrm{inv}} \in R^d$ – вектор зміщення (bias) першого шару, $W_2^{\mathrm{inv}} \in R^{D \times d}$ – матриця ваг другого шару, яка повертає вектор у вихідний простір D, $b_2^{\mathrm{inv}} \in R^D$ – вектор зміщення другого шару та σ – активаційна функція ReLU.

DSA виділяє локальні або контекстуальні особливості, притаманні конкретному домену. Має таку ж архітектуру, як DIA, але з окремими параметрами:

$$\mathrm{DSA}(h_i) = W_2^{\mathrm{spec}} \cdot \sigma\left(W_1^{\mathrm{spec}} \cdot h_i + b_1^{\mathrm{spec}}\right) + b_2^{\mathrm{spec}},$$

де $h_i \in R^D$– вхідний вектор ознак для і-го токена з виходу VideoBERT, $W_1^{\mathrm{spec}} \in R^{d \times D}$ – матриця ваг першого шару адаптера, що стискає вектор у простір меншої розмірності d, $b_1^{\mathrm{spec}} \in R^d$ – вектор зміщення (bias) першого шару, $W_2^{\mathrm{spec}} \in R^{D \times d}$ – матриця ваг другого шару, яка повертає вектор у вихідний простір D, $b_2^{\mathrm{spec}} \in R^D$ – вектор зміщення другого шару та σ – активаційна функція ReLU.

Далі отримані ознаки передаються у шар злиття для формування фінального контексту для декодеру.

Шар злиття ознак

Обидва адаптери поєднуються на рівні кожного токена за допомогою навчуваного параметру злиття:

$$\widetilde{h_\iota} = \alpha \cdot \mathrm{DIA}(h_i) + (1 - \alpha) \cdot \mathrm{DSA}(h_i),$$

де $\widetilde{h_\iota} \in R^D$ – злитий адаптований вектор ознак. $\alpha \in [0,1]$ – це скаляр, що оптимізується в процесі навчання, $\mathrm{DIA}(h_i) \in R^D$ – вихід DIA адаптера, $\mathrm{DSA}(h_i) \in R^D$ – вихід DSA адаптера.

Цей етап є внутрішнім злиттям адаптерних ознак й формує адаптовану послідовність векторів $\widetilde{H} = \{\tilde{h}_1, \dots, \tilde{h}_T\}$, яка передається на наступний етап моделі.

Трансформер-декодер

Останнім етапом запропонованої архітектури є генерація текстового опису дії, що відбувається у відео. Для цього використовується трансформерний авто-регресивний декодер, який формує послідовність токенів природної мови на основі контексту, отриманого з адаптованих візуальних ознак.

До послідовності $\widetilde{H} = \tilde{h}_1, \dots, \tilde{h}_T$ додаються ембеддінги, які забезпечують впорядкованість у часовому вимірі, та темпоральна маска $m \in \{0,1\}^T$, що відсікає нульові неінформативні токени. Ця оброблена послідовність передається у модуль перехресної уваги кожного шару декодера як ключі та значення.

Декодер складається з 6 трансформерних шарів, кожен з яких включає:

1) маскований механізм самоуваги (masked self-attention) для автопідказки по частково сформованому тексту;
2) шар перехресної уваги до візуального контексту $\widetilde{H}$;
3) два лінійні шари прямого розповсюдження;
4) нормалізація та виключення (dropout) для підвищення стабільності.

Декодер працює у режимі авто-регресивного декодування, де кожен токен y_t генерується послідовно, з урахуванням усіх попередніх токенів $y_{<t}$ та контексту.

Задача генерації опису формулюється як максимізація ймовірності послідовності тексту $Y = \{y_1, y_2, \dots, y_N\}$ за умовою адаптованого візуального контексту $\widetilde{H}$:

$$P\left(Y \mid \widetilde{H}\right) = \prod_{t=1}^{N} P\left(y_t \mid y_{<t}, \widetilde{H}\right),$$

де $P\left(Y \mid \widetilde{H}\right)$ – ймовірність повного речення за умови контексту, $P\left(y_t \mid y_{<t}, \widetilde{H}\right)$ – умовна ймовірність токена в кроці.

Декодер реалізується як лінгвістичний трансформер з навчуваним словником, ініціалізованим випадковим чином та оптимізованим з використанням функцією витрат перехрестної ентропії (cross-entropy loss function).

На етапі застосуванні моделі генерація починається з токена <BOS> (begin-of-sentence, початок речення) і завершується, коли модель передбачає токен <EOS> (end-of-sentence, кінець речення) або досягає максимального ліміту довжини. На виході декодера формується фраза природною мовою, яка коротко описує дію у відео, наприклад: *«Pouring water in a pot».*

Висновок. У цій роботі запропоновано архітектуру для контекстної класифікації відеопослідовностей, що поєднує переваги попередньо натренованого енкодера VideoBERT з доменно-обізнаним адаптаційним модулем DA-Ada. Введення подвійної адаптерної структури дозволяє моделі розділяти універсальні та домено-специфічні ознаки, що є ключовим для узагальнення на нові типи сцен та контекстів, відсутні у тренувальних даних.

Архітектура доповнена авто-регресивним трансформерним декодером, який забезпечує точну генерацію текстових описів дій на основі адаптованих візуальних ознак.

На даному етапі представлена лише архітектурна частина моделі. У подальшій роботі буде реалізовано модель, проведено експериментальне навчання на базі датасету Something-Something V2 та виконано оцінювання якості контекстної класифікації.

СПИСОК ЛІТЕРАТУРИ

1. Новічонок М. С., Норматова Т. В. «Вибір методів машинного навчання для задач контекстної класифікації відео у бізнес аналізі». *Збірник наукових праць студентів, магістрантів та викладачів «Науковий пошук молодих дослідників»*. *Житомир*. 2025. С. 124–127. – URL: https://eprints.zu.edu.ua/44884/1/1.pdf (дата звернення: 08.09.2025).

2. Alijani S., Fayyad J., Najjaran H. “Vision transformers in domain adaptation and domain generalization: a study of robustness”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2404.04452.

3. Sun C., Myers A., Vondrick C., Murphy K., Schmid C. “Videobert: a joint model for video and language representation learning”. *Google Research*. 2019. DOI: https://doi.org/10.48550/arxiv.1904.01766.

4. Li H., Zhang R., Yao H., Zhang X., Hao Y., Song X., Li X., Zhao Y., Li L., Chen Y. “DA-Ada: learning domain-aware adapter for domain adaptive object detection”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2410.09004.

5. Venugopalan S., Rohrbach M., Donahue J., Mooney R., Darrell T., Saenko K. “Sequence to sequence – video to text”. *ICCV*. 2015. DOI: https://doi.org/10.1109/ICCV.2015.515.

6. Tsimpoukelli M., Menick J., Cabi S., Eslami S. M. A., Vinyals O., Hill F. “Multimodal few-shot learning with frozen language models”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2106.13884.

7. Sung Y.-L., Cho J., Bansal M. “VL-Adapter: parameter-efficient transfer learning for vision-and-language tasks”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2112.06825.

8. “Something-Something v.2 Dataset”. – URL: https://www.qualcomm.com/developer/software/something-something-v-2-dataset (дата звернення: 08.09.2025).

DOI: https://doi.org/10.15276/ict.02.2025.19
UDC 004.93

Video context classification by VideoBERT and DA-Ada adapters

Mariia S. Novichonok[1)]
PhD student of the Department of Informatics
ORCID: https://orcid.org/0009-0007-1787-0546; mariia.novichonok@nure.ua
Sergii V. Mashtalir[1)]
Doctor of Engineering Sciences, Professor of the Department of Informatics
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Kharkiv National University of Radio Electronics, 14, Nauky Ave. Kharkiv, 61166, Ukraine

ABSTRACT

This paper proposes an architecture for the task of contextual video classification that combines the strengths of the pre-trained video-speech encoder VideoBERT, the adaptive module DA-Ada (Domain-Aware Adapter), and the auto-regressive transformer decoder. The main goal is to build a system capable of generating text descriptions of actions in videos with a high degree of generalization to new domains. The architecture is designed taking into account the requirements for scalability, flexible adaptation, and reducing the time spent on pre-training the model in the future. The input video is divided into a sequence of frames, each frame is converted into a feature vector using ResNet-50 pre-trained on ImageNet. Then, the frame vectors are projected into the visual token space and passed to the VideoBERT module. This encoder, built on the basis of the transformer architecture BERT, performs contextualization of features across the entire video sequence, modeling long-term temporal dependencies between frames. All VideoBERT parameters remain frozen, which reduces the need for additional training resources. After encoding, each representation is passed to the DA-Ada adaptation module, which consists of two parallel branches: DIA (Domain-Invariant Adapter) and DSA (Domain-Specific Adapter). DIA learns to filter out common, invariant features that are characteristic of most videos. DSA focuses on detecting features that are inherent in a specific domain (e.g., household scenes, industrial objects, etc.). The output representations of both adapters are combined using a scalar coefficient that determines the balance between universality and specialization. The result of this fusion is a sequence of adapted vectors, which is fed to the transformer to generate action descriptions. The generation is carried out by a transformer decoder, which consists of six layers, including a self-attention mechanism for working with partially formed text, a cross-attention mechanism for the video context, as well as standard feed-forward blocks. Starting from the <BOS> token, the decoder gradually forms a text description of the action, completing the process when the <EOS> token is generated or the maximum length is reached.

The proposed architecture provides modularity, a limited number of parameters to be trained, and the possibility of using it in different domains. In further work, it is planned to implement a full cycle of model training based on the Something-Something V2 dataset.

Keywords: Video context classification; encoder; decoder; neural networks; adapter; transformer