DOI: 10.35681/1560-9189.2023.25.2.300527

УДК 303.732.4

Н. В. Кузнєцова, С. С. Смірнов
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Проспект Берестейський, 37, 03056 Київ, Україна
e-mail: natalia-kpi@ukr.net, serhii.smirnov.sss@gmail.com

Узагальнена методологія розпізнавання мови жестів на відеопотоках на основі нейронних мереж і трансформерів

Представлено узагальнену методологію для розпізнавання мови жестів на відеопотоках, яка базується на спільному використанні нейронних мереж і трансформерів. Запропонована методологія використовує глибокі нейронні мережі для автоматичного виявлення та розпізнавання жестів у реальному часі. Для досягнення високої точності та швидкості обробки відеопотоків, використовуються трансформери — моделі штучного інтелекту, які ефективно моделюють довгострокові залежності в послідовностях даних. Запропонована методологія поєднує декілька областей знань, такі як комп'ютерний зір та обробка природної мови. Проаналізовано слабкі сторони запропонованої узагальненої методології, майбутні виклики щодо її впровадження та застосування до реальних даних.

Ключові слова: *системна методологія, системний аналіз, нейронні мережі, трансформери, мова жестів, розпізнавання, рекурентні нейронні мережі, згорткові нейронні мережі.*

Вступ

Розпізнавання жестової мови як способу комунікації людей, представлення та інтерпретація послідовності жестів у вигляді тексту або аудіо є одним із актуальних напрямів сучасних досліджень. Сьогодні у світі величезний обсяг відеоданих, а системи розпізнавання мови жестів на відеопотоках виявляються ключовим інструментом для прогресу та розвитку широкого спектра технологій. Ці системи ґрунтуються на аналізі візуальних даних і відкривають широкі можливості в різних галузях — від удосконалення засобів комунікації до розвитку майбутніх інтерфейсів. Системи розпізнавання мови жестів відіграють ключову роль у сприянні ефективній

взаємодії між людьми та технологіями, змінюючи спосіб, у який ми сприймаємо та взаємодіємо з довкіллям.

Сьогодні значну кількість досліджень зосереджено на розробці та вдосконаленні систем розпізнавання мови жестів на відеопотоках, а використання нейронних мереж і трансформерів є перспективним підходом для поліпшення їхньої якості. Завдяки поєднанню передових технологій машинного навчання та глибокого аналізу відеоданих, розробляються новаторські методи та алгоритми, що сприяють більш точному й ефективному розпізнаванню жестів у реальному часі.

Постановка завдання

У роботі виконано огляд існуючих методологічних підходів до розпізнавання мови жестів на відеопотоках, висвітлено основні принципи роботи комп'ютерних моделей і потенційні переваги для подальшого розвитку таких систем у майбутньому. Представлено основну ідею і основні етапи створення узагальненої системної методології використання нейронних мереж і трансформерів у процесі розпізнавання мови жестів на відеопотоках.

Основні етапи створення узагальненої методології

Базовий підхід до побудови узагальненої системної методології, що застосовується у процесі розробки систем розпізнавання мови жестів на відеопотоках, складається з таких етапів:

1) *збір відеоданих*:

а) визначення джерела даних і їхній збір;

б) фільтрація та обробка відеопотоків для використання в аналізі;

2) *попередня обробка та видобування ознак*:

а) попередня обробка кадрів відео для підготовки до подальшого аналізу;

б) видобування ознак або характеристик жестів з кадрів відеопотоку;

3) *навчання моделей*:

а) використання навчальних алгоритмів і даних для створення моделей розпізнавання жестів;

б) постійне покращення і оптимізація моделей шляхом навчання на нових даних;

4) *валідація та тестування*:

а) перевірка ефективності та точності системи на тестових даних;

б) валідація працездатності системи у різних умовах і середовищах;

5) *інтеграція і оптимізація*:

а) інтеграція системи в реальні додатки або пристрої;

б) адаптація системи для досягнення оптимальної продуктивності та швидкості реакції.

Ці етапи представляють загальну структуру розробки систем розпізнавання мови жестів на відеопотоках. Кожний етап включає певну кількість методів, методик і підходів, які дозволяють покращити ефективність і точність системи розпізнавання жестів.

Огляд існуючих підходів і методів

Інтенсивний розвиток відео технологій і величезний обсяг відеоданих системи розпізнавання мови жестів на відеопотоках стають ключовими для розвитку технологій. Ці системи засновані на різноманітних технологічних підходах, які дозволяють аналізувати та інтерпретувати жести, що відтворюються у відеопотоках, і реагувати на них. Розглянемо основі методи та моделі, які застосовуються в таких системах.

Нейронні мережі [2–4]. Одним із основних підходів є застосування нейронних мереж, зокрема згорткових нейронних мереж (CNN — convolutional neural network) і рекурентних нейронних мереж (RNN — recurrent neural network). Згорткові мережі застосовуються для виявлення патернів та ознак жестів на кадрах відеопотоку, в той час як рекурентні нейронні мережі дозволяють моделям аналізувати послідовність жестів у часі.

Трансформери [5, 6]. Механізми трансформерів виявляються досить ефективними у роботі з послідовностями даних у відеопотоках. Вони дозволяють виявити зв'язки між жестами в технологіях, які зазвичай використовуються в системі розпізнавання мови жестів, щоб видобути з повного відео текст мовою жестів. Потім система перекладу генерує розмовні переклади зі знакової мови жестів. Така методологія ще відома під назвою Gloss2Text. У роботі [5] представлено систему перекладу та покращення продуктивності завдяки використання мереж трансформерів і просторово-часових мереж Multi-Cue (STMC). Робота [6] представляє архітектуру на основі трансформера, яка одночасно вивчає безперервне розпізнавання та переклад мови жестів, навчаючись наскрізним способом, створюючи єдину уніфіковану архітектуру.

Комбіновані методи [7]. Деякі системи використовують комбінацію згорткових і рекурентних мереж, де згорткова мережа використовується для добування ознак із кадрів, а рекурентна нейронна мережа аналізує ці ознаки в послідовності для розпізнавання жестів.

Просторово-часові методи [8]. Аналіз руху та просторових змін між кадрами відеопотоку є важливим для розпізнавання жестів, оскільки дозволяє виявити динаміку жестів у часі.

Наведені вище підходи та методи можуть застосовуватись окремо або комбінуватися для створення ефективних систем розпізнавання мови жестів на відеопотоках, забезпечуючи високу точність і здатність реагувати в реальному часі.

Розробка систем розпізнавання мови жестів на відеопотоках включає не тільки технічні аспекти, але й питання дизайну, валідації, тестування та інтеграції. При цьому досі ще залишаються невирішеними проблеми точності, простоти та швидкості, тому потреба в нових підходах та адаптації існуючих методів є досить високою.

Побудова системи розпізнавання мови жестів

Конкретна методологія передбачає формалізацію постановки задачі та конкретизацію даних, з якими система буде працювати. Розв'язання задач розпізнавання мови жестів на відеопотоках полягає у створенні системи, яка здатна автоматично ідентифікувати та класифікувати жести, виконані людиною на відео. Основ-

на мета полягає в тому, щоби система могла розрізняти та класифікувати різні жести в реальному часі на основі відеопотоку, а також щоби система була робастною, тобто могла працювати за різних початкових умов, з різною швидкістю мовлення людини, за наявності та відсутності шумів на відеопотоці тощо.

За базу мови жестів A розглянемо деякий набір відео та методів їхньої обробки:

$$< A, f: A \to A >. \quad (1)$$

Відео представляють собою фрагменти, потоки, частини відео, цілі відео та їхні метадані, на яких люди розмовляють мовою жестів, та які можуть бути використані для подальшого тренування нейронних мереж, для їхньої валідації, оцінки якості системи чи використання в режимі реального часу для власне розпізнавання мови жестів конкретного співрозмовника у створеній системі. Дана база A задає спосіб зберігання відео для різних цілей, а також їхньої попередньої обробки. Ця база даних завідомо неповна, адже неможливо зібрати всі можливі знаки мови жестів, діалекти та жаргонізми для всіх мов світу, різні послідовності знаків і швидкість їхнього інтерпретування або показування. Саме тому першочергово ця база даних використовується для тренування майбутніх моделей системної методології і їхньої оцінки, аби на базі, наприклад, штучного інтелекту вирішити частково цю проблему неповноти.

Як приклад такої бази даних розглянемо набір даних LSA64 [9]. Датасет LSA64 (Argentinian Sign Language 64) є набором даних, спеціально створених для вивчення та розпізнавання жестів, які є характерними для Аргентинської мови жестів. Зібраний з урахуванням специфічних особливостей і варіацій жестів набір даних став ключовим ресурсом для досліджень у галузі комп'ютерного зору та машинного навчання. LSA64 містить послідовності відеокадрів або відеофрагментів, що відображають різні жести Аргентинської мови жестів (рис. 1 і рис. 2). Кожен жест у наборі даних відзначено анотацією, що дозволяє точно ідентифікувати та класифікувати жести з урахуванням їхнього значення та специфіки. Наразі в цій базі присутні 64 класи — слова з аргентинської мови жестів.

Під методами обробки відео $f(x)$ у цій базі слід розуміти алгоритми та підходи для перетворення суцільного відео або його частини на фрагменти, які можна буде подавати або в тренувальний процес нейронних мереж, або в режимі реального часу для задачі розпізнавання мови жестів. Наприклад, на вхід може подаватися суцільне відео x, яке, проходячи через деякий оператор f, буде ділитися на фрагменти: один фрагмент — одне речення, слово або буква в мові жестів залежно від подальшої задачі. Така операція може бути виконана алгоритмічно. Наприклад, речення можна розпізнавати за допомогою спеціальних символів. У мові жестів, а також при виконанні розпізнавання мови жестів, визначення кінця речення або слова може бути виконано через різноманітні жести, які сигналізують про завершення висловлювання. Одним із таких жестів є жест, що використовується для підсилення кінця висловлення. Наприклад, могутній жест з використанням руки або пальця, спрямований вниз і вперед від тіла, може сигналізувати про закінчення висловлення або слова. Деякі люди також використовують «жест пунктуації», де рука або палець мають певне рухове значення, що вказує на зупинку чи закінчення. Наприклад, підняття долоні вгору з послідовним швидким зниженням може інтерпретуватися як знак кінця висловлення чи речення.

Рис. 1. Приклад слова «зелений» з LSA64

Рис. 2. Приклад слова «колір» з LSA64

Також у рамках оператора f відбувається сегментація долонь, для звуження області уваги для наступних методів. Сегментація положень рук на відео для розпізнавання мови жестів є важливим етапом у процесі аналізу відеоданих, де можуть бути використані такі методи: фільтри Калмана, гістограма орієнтації градієнтів (HOG), що базуються на візуальних ознаках (наприклад, за кольором шкіри), сенсори глибини, моделі глибокого навчання (наприклад, Yolo) [14].

Для спрощення задачі сегментації на людей можуть бути одягнені кольорові рукавиці, так як показано на рис. 1, 2. Набір даних LSA64 містить сегментовані положення рук (рис. 3 та 4), виконані через використання декількох видів дескрипторів (radon transform, sift тощо) [10].

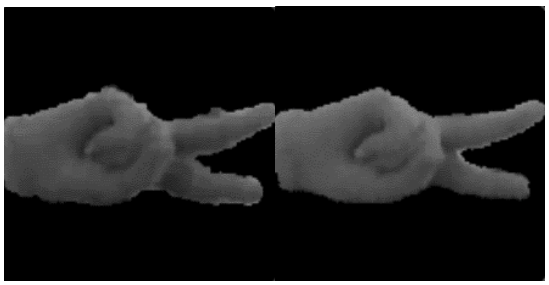

Рис. 3. Приклад сегментованих положень руки для слова «зелений» з LSA64

Тоді для узагальненої методології можемо визначити деяку базу знань щодо мови жестів разом з відео або фрагментами відео, що містять приклади речень, слів, знаків, а також оператор попередньої обробки цих відео для отримання сегментованих областей інтересу для подальшої роботи, яка є неповною та відкритою:

$$< A, f : A \to A >. \quad (2)$$

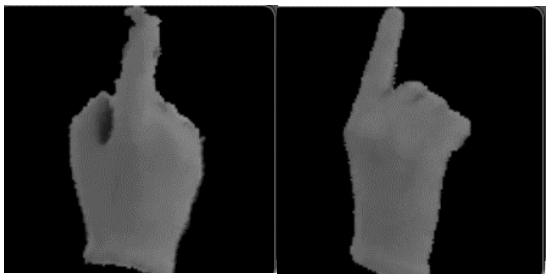

Рис. 4. Приклад сегментованих положень руки для слова «колір» з LSA64.

Подаючи на вхід частину відеопотоку $x \epsilon f(A)$, тобто символ або слово, застосовуємо деяку функцію розпізнавання слова на відеопотоці $g(x, a, n)$. Дана функція може бути представленою спеціальним перетворенням, що може приймати на вхід відеопотік і повертати ймовірності приналежності заданого фрагменту до певних класів, слів або символів з A, тобто

$$g(x, a, n \mid x \epsilon f(A), a \epsilon A, n \epsilon N) = \{a_n : y`_n \mid a_n \epsilon a, y`_n \epsilon [0,1]\}_{n \epsilon N}, \quad (3)$$

де a_n — n-не слово або символ (клас, на який розділяємо, тобто класифікуємо заданий відеопотік); $y`_n$ — ймовірність або впевненість, що це дійсно цей клас; n — кількість класів на виході, які складають топ n для заданого вхідного слова. Зазначимо, що n є наперед заданим параметром, який визначає кількість повернутих класів з їхніми ймовірностями, які модель визначає як найбільш імовірну відповідь до задачі розпізнавання. Альтернативним варіантом інтерпретування цього параметра може бути певна межа ймовірності, за якою функція g повертає тільки ті класи, які мають $y` > n \epsilon R$. Складність такого підходу полягає в тому, що модель може бути як певною мірою впевнена у своїй відповіді (тобто сума перших декількох найбільших імовірностей складає ймовірність більше за 0,95), так і невпевнена (тобто будь-яка ймовірність приналежності x до певного класу менша за 0,05). Буде важко підібрати ідеальний поріг для виходу функції g, а також за кожного x буде різна кількість імовірних класів на виході, що буде ускладнювати подальший процес аналізу на наступних кроках через стохастичність. Отже, зупиняємося на тому, що $n \in N$, яка задається наперед.

Дана функція $g(x, a, n)$ може бути представленою нейронною мережею. У методології, що пропонується, будемо застосувати мережі, що засновані на трансформерах, які були досліджені для даної задачі у статті [13]. Таким чином, на другому кроці маємо:

$$f(A) \text{э } x \rightarrow g(x, a, n) = \{a_n : y`_n \mid a_n \epsilon a, y`_n \epsilon [0,1]\}_{n \epsilon N}. \quad (4)$$

Як приклад розглянемо конкретну модель, що заснована на трансформерах, X-CLIP. Модель X-CLIP вперше описана у науковій статті [11], вона є мінімальним розширенням CLIP, спеціально призначеним для аналізу відео. Ця модель складається з кодера тексту, кодера крос-кадрового зображення, багатокадрової інтеграційної нейронної мережі з механізмом уваги та генератора підказок для відео. Автори роботи відзначають успіх контрастного попереднього навчання мови та зображення у створенні спільного візуально-текстового представлення на основі велико-

масштабних веб-даних, демонструючи вражаючу здатність до «zero-shot» для різних завдань обробки зображень.

Zero-shot learning (ZSL) — нульове коротке навчання — це підхід машинного навчання, який дозволяє моделям навчатися розпізнавати об'єкти чи поняття, на які вони не були навчені безпосередньо. У відеорозпізнаванні, zero-shot learning означає, що модель може розпізнавати та класифікувати об'єкти чи дії, які не були представлені у навчальному наборі даних. Це і відрізняє X-CLIP-модель від стандартної класифікаційної нейронної мережі.

Для досягнення zero-shot learning у відеорозпізнаванні модель зазвичай користується даними, які містять асоціації між класами або поняттями, але не містять безпосередньо відповідних візуальних прикладів для цих класів. Ці асоціації можуть бути представлені у формі текстових описів, атрибутів чи інших метаданих, які допомагають моделі розуміти та уявляти той чи інший клас чи дію.

У такому контексті модель навчається користуватися цими асоціаціями для виконання класифікації або розпізнавання нових об'єктів у відео, навіть якщо вона не мала можливості бачити конкретні приклади цих об'єктів у процесі навчання. Вона використовує знання, яке здобула під час тренування на інших, схожих об'єктах або діях, для узагальнення та розпізнавання нових.

Застосування zero-shot learning у відеорозпізнаванні може бути корисним у випадках, коли набір даних обмежений або коли потрібно розпізнати нові об'єкти, які не були доступні для тренування, розширюючи можливості моделі у вирішенні реальних завдань. Саме такий підхід дозволяє використовувати подібні нейронні мережі для задачі розпізнавання мови жестів, навіть якщо такі мережі були навчені на інших базах знань, або інших мовах жестів, або від бази знань A, яка є неповною системою знань щодо конкретної мови.

Отже, використовуючи подібну мережу, навіть не тренуючи її під конкретну задачу, можна під час формування висновку отримати ймовірності для класів, які модель не бачила раніше, але для яких вона може визначити необхідні патерни розпізнавання. Тож на вхід до подібної архітектури для нашої задачі слід надавати усі можливі класи з бази знань A, для яких модель і буде рахувати ймовірності.

Серед проблем або викликів використання тієї ж X-CLIP-моделі для задачі розпізнавання мови жестів, окрім як її оптимізація задля досягнення кращої якості та метрик роботи моделі, є робота зі словами або символами, які раніше не були відомі та представлені в базі знань A, наприклад, при роботі з діалектами або неологізмами. Такі проблеми можуть виникати і при живому спілкуванні між людьми. Для їхнього вирішення мозок зазвичай використовує складні механізми обробки природної мови для визначення нового почутого терміну або його ігнорування через новизну. Дана проблема є достатньо складною та неоднозначною і є предметом для подальших досліджень, тому на даний момент будемо приймати за задане, що база A містить уявлення про всі відомі на сьогодні слова/символи певної мови.

Чому не вистачає однієї моделі для розв'язання нашої задачі (якщо $n = 1$)? Для пояснення розглянемо метрики, які використовуються при оцінці та валідації моделей g. У задачах розпізнавання мови жестів часто використовують метрику top-K accuracy (точність top-K) замість звичайної точності (accuracy) або f1-score, які є популярними для класичних задач класифікації чи розпізнавання:

$$top - K\ accuracy = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \sum_{j=1}^{k} 1\left(\widehat{f_{i,J}} = y_i\right), \quad (5)$$

де $\widehat{f_{i,J}}$ — прогнозований клас для i-го зразка, що відповідає j-му найбільшому прогнозованому скору; y_i — відповідне істинне значення; k — кількість дозволених припущень; $1(x)$ — функція індикатора. Це пов'язано з декількома причинами.

1. *Множинність можливих класів*. У задачах розпізнавання мови жестів може існувати багато різних жестів, що ускладнює завдання класифікації. Замість простої класифікації на один клас, ми маємо десятки, сотні або навіть тисячі класів, кожен з яких може бути потенційно вірним прогнозом. У таких випадках метрика top-N accuracy вказує, наскільки часто правильний клас потрапляє серед перших N найбільш імовірних прогнозів моделі.

2. *Практична значимість*. У реальних сценаріях розпізнавання жестів часто важливо не тільки точно визначити правильний клас, але й здатність моделі представляти можливі варіанти, які можуть бути правильними. Top-N accuracy дозволяє оцінити, наскільки модель може «підказати» щодо можливого правильного варіанту, навіть якщо перший прогноз не є точним.

3. *Збалансованість*. Коли використовуються великі набори даних для розпізнавання жестів, кількість класів може бути нерівномірною. Деякі класи можуть мати більше прикладів для навчання, ніж інші, що може призвести до нерівного розподілу класів у валідаційному наборі. Top-N accuracy може бути більш робастною метрикою у випадку незбалансованих даних, оскільки вона оцінює правильність класифікації серед топ-N прогнозів, а не тільки першого.

Враховуючи ці фактори, top-K accuracy є більш інформативною метрикою для вимірювання продуктивності моделей у завданнях розпізнавання мови жестів. Дана функція дає представлення щодо того, що якби модель мала певний контекст щодо області або ситуації, в яких вона працює, вона могла би підлаштувати вихід задачі розпізнавання. Саме для використання контексту введемо до методології наступну функцію h, яка здатна працювати з виходом попереднього кроку та з контекстом.

На відміну від функції $g(x, a, n)$, яка працює в області комп'ютерного зору, функція h працює в області обробки природних мов. Для того, щоб використати контекст, ми накладаємо на наш попередній вихід наступну функцію:

$$h(w_t, \{a_n : y`_n \mid a_n \epsilon\, A, y`_n \epsilon\, [0,1]\}_{n \epsilon N}, \quad (6)$$

де w_t — речення з попередніх розпізнаних символів на момент часу t. Слід зауважити, що розглядається не весь попередній контекст, а починаючи з певного моменту часу $t_1 < t$. Під певним часом можна мати на увазі як час, наприклад, останні 10 секунд мовлення людини, або певну подію, наприклад, початe незакінчене речення. Вирізати початок можна, орієнтуючись на закінчення речення мовою жестів або великою паузою в мовленні, описаними у функції f — попередньої обробки відеопотоку для ділення потоку на речення. Тобто w_t — це початок певного речення, яке утворюється в момент часу t. У цей же час на вхід до системи розпізнавання подається символ, який треба розпізнати, аби логічно продовжити речення, сказане людиною перед камерою.

Виходом такої функції h, яка заснована на нейронних мережах і трансформерах відповідно до результатів статті [13], є слово або символ s, яке додаючи до

$w_t -> w_t + s = w_{t+1}$ створює вхідне речення на наступне слово для розпізнавання, тобто на момент часу $t + 1$.

Такий підхід дозволяє використати, по-перше, попередній контекст, в якому розмовляють співрозмовники. По-друге, використати ймовірності з попереднього кроку, аби врахувати впевненість в одних словах і невпевненість в інших і використати відповідні нейронні мережі (наприклад, нечіткі нейронні мережі тощо).

До складнощів застосування подібних нейронних мереж можна віднести випадки, коли речення, яке вимовляє людина, тільки починається. Тобто, $t = t_0$. У даній ситуації моделі буде важче оперувати вхідними ймовірностями, адже немає попереднього контексту ситуації. У таких окремих випадках можна або використовувати як попередній контекст попереднє речення, що розпізнавалося, w_t, де $t < t_0$, або ж навчити нейронну мережу як вхідний контекст, для розпізнавання поточного слова мовлення приймати сам факт початку речення. В контексті певної мови A можна отримати знання, з яких слів найчастіше починається речення, або точніше, з яких слів речення скоріш за все ніколи не починається. Наприклад, у німецькій мові речення не може починатися з присудку, якщо це не питання, а отже, коли це питання, то модель повинна більше орієнтуватися на присудки з результатів попередньої функції g, а в інших ситуаціях — навпаки, зменшувати ймовірність виходу таких присудків.

Як приклад такої функції h розглянемо трансформер Multiple Choice BERT. Multiple Choice BERT (MC-BERT) [12] — це варіація моделі BERT (Bidirec-tional Encoder Representations from Transformers), спеціально налаштована для вирішення задач з вибором із кількох варіантів відповіді. Ця модель призначена для розв'язання завдань, де потрібно вибрати правильну відповідь серед кількох альтернатив.

MC-BERT зазвичай використовується у завданнях тестування, де потрібно визначити правильну відповідь серед набору запитань і варіантів відповідей. Ця модель отримує на вхід текстові фрагменти: питання, можливі варіанти відповідей і контекст, які інкапсулюють інформацію, необхідну для визначення правильної відповіді. MC-BERT використовує механізми уваги та трансформерні шари для засвоєння контексту, які допомагають зрозуміти зв'язок між питанням, варіантами відповідей і контекстом.

Основна властивість MC-BERT — це здатність до контекстного розуміння тексту, що дозволяє йому ефективно працювати із запитаннями та альтернативами відповідей, враховуючи текстовий контекст. Вона може бути застосована у різних областях, де потрібно вибрати найкращий варіант серед запропонованих альтернатив, таких як тестування, опитування або різноманітні задачі, які передбачають вибір оптимальної відповіді серед багатьох варіантів.

Отже, зводячи все до однієї системи, отримаємо наступне представлення:

$< A, f: A \to A >$ — база знань щодо мови жесті;

T_0 — початок поточного речення, T_1 —закінчення поточного речення;

U — вхідний відеопотік речення;

$w_{T_0} = \varnothing$ — пусте речення або початок речення.

Вхідне невідоме речення для розпізнавання w_{T_1} — ?

Для $\forall t \in [T_0, T_1]$:

$u_t \in f(U)$ — вхідне сегментоване слово з фрагменту відео на час *t*;

$a \in A$ — всі відомі класи/символи/слова з бази знань

$$u_t, a \to g(u_t, a, n\epsilon N) = \{a_i : y`_i \mid a_i \epsilon\, a, y`_i \epsilon [0,1], \text{для } \forall i \epsilon [0,n]\},$$

де a_i — слово з набору слів a; $y`_i$ — ймовірність приналежності вхідного відео до i-го класу:

$$w_t, \{a_i : y`_i \mid \forall i \epsilon [0,n]\} \to h(w_t, \{a_i : y`_i \mid \forall i \epsilon [0,n]\}) = s,$$
$$s \to w_t + s = w_{t+1}.$$

Дана узагальнена методологія поєднує в собі декілька областей знань, такі як комп'ютерний зір та обробка природних мов, що дає можливість витягти більше знань з одного відеофрагменту для подальшої обробки і аналізу з метою ефективного розпізнавання мови жестів у режимі реального часу.

Висновки

Виконано огляд існуючих методологій, які використовуються в системах розпізнавання жестової мови. Завдяки цьому дослідженню стає очевидним, що, незважаючи на значний прогрес у цій галузі, такі проблеми як надійність, продуктивність у реальному часі та адаптація до різноманітних мов жестів, усе ще залишаються.

У відповідь на ці виклики запропоновано узагальнену методологію, яка базується на нейронних мережах і трансформерах для розпізнавання мови жестів. Використовуючи потужність цих передових технологій, запропоновано підхід, що спрямований на усунення існуючих обмежень, пропонуючи гнучкість між мовами жестів і покращені можливості розпізнавання в реальному часі. Використання нейронних мереж дозволяє розпізнавати складні образи та виділяти ознаки, тоді як нейронні мережі з механізмом уваги полегшують ефективне послідовне моделювання, більш повно фіксуючи залежності в жестах жестової мови.

Хоча запропонована нами узагальнена методологія формує базову концепцію створення та застосування в системах розпізнавання мови жестів, необхідні подальші дослідження та вдосконалення для оптимізації її продуктивності, масштабованості та застосовності в реальних сценаріях, а також наведення валідаційної методології для оцінки якості підходу системи як існуючих, так і запропонованої. Спільні зусилля міждисциплінарних команд, до складу яких входять лінгвісти, інженери та представники спільноти людей з вадами слуху, сприятимуть удосконаленню і адаптації цих систем для задоволення різноманітних потреб і нюансів різних мов жестів і контекстів користувачів.

Ця стаття є ідеєю і зміною парадигми шляхом інноваційної інтеграції нейронних мереж з механізмом уваги, які будуть каталізатором для подальшого розвитку технологій, сприяючи більшій інклюзивності та безперебійному спілкуванню людей, які покладаються на мову жестів як на основний засіб вираження.

1. Крак Ю.В., Барчукова Ю.В., Троценко Б.А. Побудова моделей дактилем для синтезу дактильної інформації. *Штучний інтелект*. 2011. № 3. С. 147–155.

2. Sri Lakshmi Murali R., Ramayya L.D., & Anil Santosh V. Sign Language Recognition System Using Convolutional Neural Network And Computer Vision. *International Journal of Engineering Innovations in Advanced Technology*. 2022. 4(4). 137.

3. Kadhim R.A., & Khamees M.A. Real-Time American Sign Language Recognition System using Convolutional Neural Network for Real Datasets. *TEM Journal*. 2020. 9(3). P. 937–943.

4. S. He. Research of a Sign Language Translation System Based on Deep Learning. 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Dublin, Ireland, 2019. P. 392–396. doi: 10.1109/AIAM48774.2019.00083.

5. Yin K. (2020, April 1). Sign Language Translation with Transformers. arXiv:2004.00588v.

6. Camgoz N.C., Koller O., Hadfield S., & Bowden R. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 10023–10033.

7. Masood S., Srivastava A., Thuwal H. C. and Ahmad M. Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. *In Intelligent Engineering Informatics.* Springer, 2018. P. 623–632.

8. Shivashankara S., & Srinath S. American Sign Language Recognition System: An Optimal Approach. *I.J. Image, Graphics and Signal Processing*. 2018. **8**. P. 18–30. https://doi.org/10.5815/ijigsp.2018.08.03.

9. Ronchetti F., Quiroga F., Estrebou C., Lanzarini L., and Rosete A. LSA64: A Dataset of Argentinian Sign Language. In XXII Congreso Argentino de Ciencias de la Computación (CACIC). 2016.

10. Ronchetti F., Quiroga F., Lanzarini L., & Estrebou C. Handshape Recognition for Argentinian Sign Language using ProbSom. *Journal of Computer Science and Technology*. 2016. 16(1). P. 1–5. ISSN 1666-6038.

11. Ni B., Peng H., Chen M., Zhang S., Meng G., Fu J., Xiang S., & Ling H. (2022). Expanding Language-Image Pretrained Models for General Video Recognition. Accepted by ECCV2022, Oral. arXiv preprint arXiv:2208.02816 [cs.CV]. https://doi.org/10.48550/arXiv.2208.02816.

12. Xu Z., Gong L., Ke G., He D., Zheng S., Wang L., Bian J., & Liu T.-Y. (2020). MC-BERT: Efficient Language Pre-Training via a Meta Controller. arXiv preprint arXiv:2006.05744 [cs.CL]. https://doi.org/10.48550/arXiv.2006.05744.

13. Nataliia Kuznietsova, Serhii Smirnov, Application of Vision Transformers and 3D Convolutional Neural Networks for Sign Language Cluster Recognition. CEUR Workshop Proceeding (ISSN 1613-0073). 2023. Vol. 3392, CMIS 2023. P. 151–163. http://ceur-ws.org/Vol-3392/

14. Manel BEN ABDALLAH, AMENI Sessi, MOHAMED Kallel, M.S.BOUHLEL. Different Techniques of Hand Segmentation in the Real Time. *International Journal of Computer Applications & Information Technology*. 2013, January. Vol. II, Issue I. ISSN: 2278-7720.

Надійшла до редакції 01.12.2023