Strategie Adaptacji Modeli Wizualno-Językowych do Zadań Docelowych

Niniejsza praca doktorska bada adaptacyjność modeli fundamentalnych do reprezentacji obrazu, do złożonych zadań docelowych, ze szczególnym naciskiem na podejścia wykorzystujące wewnętrzne możliwości modeli fundamentalnych przy minimalizacji kosztów ich modyfikacji, w tym dodatkowego treningu.

Nasze badania koncentrują się głównie na modelach wizualno-językowych (VLM), analizując, jak ich multimodalne reprezentacje obrazu i tekstu mogą być rozszerzone do zadań wymagające szczegółowej reprezentacji obrazu. Poprzez pięć powiązanych ze sobą prac, odpowiadamy na centralne pytanie: W jakim stopniu reprezentacje wizualne z modeli fundamentalnych mogą być wykorzystane do różnych zadań docelowych przy minimalnym nadzorze?

W pierwszej części doktoratu omawiamy metody adaptacji modelu CLIP do zadania semantycznej segmentacji z nieograniczonym słownikiem klas. Nasza pierwsza metoda, CLIP-DIY, demonstruje, możliwości adaptacji CLIP poprzez modyfikacje sposobu inferencji tym samym bez konieczności trenowania modelu. Następnie prezentujemy CLIP-DINOiser, metodę, która wzbogaca reprezentacje CLIP na poziomie pikseli poprzez wykorzystanie komplemetarnych umiejętności lokalizacji obiektów z reprezentacji uczonych z samonadzorem, takich jak DINO, za pomocą prostego modułu adaptacyjnego. W tej części pracy pokazujemy również jak poprawić skuteczność segmentacji poprzez starannie dobrane prompty tekstowe, które pozyskujemy poprzez analizę dużego korpusu danych wykorzystanych do trenowania VLM.

Wykraczając poza adaptację modelu, kolejna część doktoratu prezentuje metodę do ewaluacji reprezentacji wizualnych w złożonym zadaniu Visual Question Answering (VQA). Tak skonstruowany sposób ewaluacji pozwala dogłębnie zrozumieć skuteczność poszczególnych reprezenetacji wizualnych do zadań rozumowania na podstawie obrazu. W ostatniej części pracy pokazujemy praktyczne zastosowanie adaptacji VLM do zadania personalizacji wizualnych podsumowań na przykładzie problemu na platformie Booking.com, łącząc analizę reprezentacji wizualnej z analizową tekstowych recenzji użytkowników platformy.

Podsumowując, niniejsza praca pokazuje, że modele fundamentalne do reprezentacji obrazu mogą być efektywnie adaptowane do złożonych zadań wymagających szczegółowego rozumienia obrazu przy minimalnych wymaganiach obliczeniowych i anotacyjnych. Nasze obserwacje pogłębiają wiedzą dotyczącą rożnych sposobów reprezentacji obrazu jednocześnie dostarczając praktycznych rozwiązań do adaptacji modeli fundamentalnych w różnorodnych zastosowaniach.

Słowa kluczowe: metody adaptacji do zadań złożonych, reprezentacje wizualno-tekstowe,

segmentacja semantyczna z nieograniczonym słownikiem, uczenie nienadzorowane/ samonadzorowane