بِسمِ اللهِ الرَّحمنِ الرَّحیم

بهبود چارچوب معلم-شاگرد بازآموزی ماسک‌شده برای تشخیص شیء تطبیق دامنه

(improvement on Masked Retraining Teacher-Student Framework for Domain Adaptive Object Detection)

دانشجو: امیررضا کاظم لو - سجاد قاضیانی

استاد درس: محمدعلی کیوان‌راد

سرفصل مطالب

۱ مقدمه

تشخیص اشیاء یکی از وظایف کلیدی در بینایی کامپیوتر است. با این حال، مدل‌های یادگیری عمیق زمانی که در محیطی متفاوت از داده‌های آموزشی خود مستقر می‌شوند، با افت عملکرد شدید مواجه می‌شوند که این پدیده به "شکاف دامنه" (**Domain Shift**) معروف است. برای غلبه بر این مشکل بدون نیاز به برچسب‌گذاری مجدد داده‌ها که فرآیندی پر هزینه است، از رویکرد انطباق دامنه بدون نظارت (**Unsupervised Domain Adaptation - UDA**) استفاده می‌شود. در حوزه تشخیص اشیاء، این رویکرد به عنوان **DAOD** شناخته می‌شود و بسیاری از روش‌های پیشرفته از چارچوب معلم-شاگرد (**Teacher-Student**) بهره می‌برند. در این چارچوب، یک مدل معلم برچسب‌های کاذب (**pseudo labels**) برای داده‌های بدون برچسب دامنه هدف تولید کرده و مدل شاگرد را آموزش می‌دهد. با این وجود، این روش‌ها با چالش جدی کیفیت پایین برچسب‌های کاذب روبرو هستند. این مشکل به دو صورت بروز می‌کند: تعداد محدود برچسب‌های تولید شده، و وجود پیش‌بینی‌های نادرست. پیش‌بینی‌های نادرست می‌توانند مدل شاگرد را گمراه کرده و آن را در یک بهینه محلی (**Local Optimum**) گرفتار کنند که منجر به نتایج ضعیف می‌شود.

این پروژه با هدف پیاده‌سازی و تحلیل چارچوب معلم-شاگرد بازآموزی ماسک‌شده (**MRT**) که در مقاله پایه **[1]** ارائه شده، انجام می‌شود. هدف اصلی این چارچوب، غلبه بر مشکل برچسب‌های کاذب بی‌کیفیت در وظیفه انطباق دامنه برای تشخیص اشیاء است.

۲ مقاله پایه (یا روش اصلی)

در مقاله پایه [۱]، یک چارچوب جدید به نام **MRT** برای انطباق دامنه بدون نظارت در تشخیص اشیاء ارائه شده است .این چارچوب بر روی یک مدل پایه معلم-شاگرد تطبیقی ساخته شده و دو نوآوری اصلی را برای بهبود کیفیت یادگیری معرفی می‌کند. معماری و روش اصلی آن به شرح زیر است:

- چارچوب پایه معلم-شاگرد تطبیقی :این ساختار شامل یک مدل معلم و یک مدل شاگرد است .مدل معلم تصاویر دامنه هدف با افزونگی ضعیف را دریافت کرده و برچسب‌های کاذب تولید می‌کند .سپس مدل شاگرد با استفاده از این برچسب‌ها (برای داده‌های هدف) و برچسب‌های واقعی (برای داده‌های منبع) آموزش می‌بیند .همچنین از

هم‌ترازی تخاصمی **(Adversarial Alignment)** برای کاهش شکاف بین ویژگی‌های دو دامنه در سطوح مختلف شبکه (ستون فقرات، رمزگذار و رمزگشا) استفاده می‌شود.

- شاخه خودرمزگذار ماسک‌شده **:(Masked Autoencoder - MAE Branch)** برای اینکه مدل شاگرد درک بهتری از ویژگی‌های دامنه هدف پیدا کند، یک شاخه **MAE** سفارشی به آن اضافه می‌شود .این شاخه به صورت تصادفی بخش‌هایی از نقشه‌های ویژگی تصاویر دامنه هدف را ماسک کرده و مدل را وادار می‌کند تا با استفاده از رمزگذار خود و یک رمزگشای کمکی، بخش‌های حذف‌شده را بازسازی کند .این وظیفه خودنظارتی به مدل کمک می‌کند تا حتی با تعداد محدودی برچسب کاذب، یادگیری کارآمدتری داشته باشد.
- مکانیزم بازآموزی انتخابی **:(Selective Retraining Mechanism)** برای جلوگیری از گرفتار شدن مدل در بهینه‌های محلی ناشی از برچسب‌های کاذب نادرست، از این مکانیزم استفاده می‌شود .به صورت دوره‌ای، بخشی از پارامترهای مدل شاگرد (مانند ستون فقرات و رمزگذار) با وزن‌هایی که توسط **MAE** بهبود یافته‌اند، مجدداً مقداردهی اولیه می‌شوند .این کار به مدل اجازه می‌دهد تا از بهینه محلی خارج شود، در حالی که مدل معلم دست‌نخورده باقی می‌ماند تا کیفیت تولید برچسب‌های کاذب افت نکند.
- هدف ما در پیاده سازی اولیه این مقاله و سپس اعمال ایده های جدید در زمینه ماسک کردن لایه های استخراجی از **back bone** است تا در عملکرد نهایی مدل در محیط **target** بهبود داشته باشیم**[2]**.

۳ مجموعه داده

در این پروژه معیار اصلی ارزیابی مدل ها و توسعه هایی که روی مدل اصلی صورت گرفته است بر مبنای دیتابیسی هست که جمع آوری شده است و دو بخش داده هایی که از بازی های کامپیوتری جمع آوری شده است که به عنوان محیط منبع به مدل و داده های واقعی از تانک و ماشین از سطح اینترنت به عنوان محیط هدف به آن داده خواهد شد.

برای این پروژه از مجموعه داده‌های استانداردی که در مقاله مرجع [۱] نیز استفاده شده است، هم بهره‌برداری خواهد شد که معیار اولیه برای عملکرد مدل بوده است ولی در نهایت معیار اصلی همان دیتابیس تانک-ماشین جمع آوری شده خواهد بود.

۱. **دامنه منبع** (Source Domain)**: مجموعه داده کشت که شامل ۲,۹۷۵ تصویر آموزشی از صحنه‌های شهری در شرایط آب و هوایی مناسب است.**

بخش آموزشی:(Training Set)

- **تعداد تصاویر :۲۲۲۰ تصویر**
- **تعداد حاشیه‌نویسی**(Annotation)**: ۴۰۵۱ جعبه برای ۲ دسته «تانک» و «خودرو»**
- **ابعاد تصاویر :تمامی تصاویر دارای ابعاد ثابت ۱۹۲۰ در ۱۰۸۰ پیکسل هستند.**
- **توزیع دسته‌ها:**
  - **تانک: ۳۹۸۵ نمونه (۹۹.۳۰٪)**
  - **خودرو: ۲۸ نمونه (۰.۷۰٪)**

بخش اعتبارسنجی:(Validation Set)

- **تعداد تصاویر :۵۵۶ تصویر و برای اعتبارسنجی مدل در دامنه منبع استفاده می‌شوند.**
- **تعداد**annotation**: ۱۱۸۷ شیء از دو دسته مذکور.**
- **ابعاد تصاویر :تمامی تصاویر دارای ابعاد ثابت ۱۹۲۰ در ۱۰۸۰ پیکسل هستند.**
- **توزیع دسته‌ها:**
  - **تانک: ۱۱۸۳ نمونه (۹۹.۶۶٪)**
  - **خودرو: ۴ نمونه (۰.۳۴٪)**

۲. **دامنه هدف** (Target Domain)**:**

- **تعداد تصاویر: ۲۸۷۰ تصویر که در فایل** target.json **قرار دارند.**
- **تعداد حاشیه‌نویسی: ۴۳۲۴ جعبه.**

- **ابعاد تصاویر: ابعاد تصاویر متغیر است (میانگین ۶۵۲ در ۴۷۲ پیکسل).**

برچسب‌گذاری: بخشی از این مجموعه داده (تقریبا ٪۵۰) دارای برچسب است تا در حین فرآیند آموزش برای اعتبارسنجی و ذخیره بهترین نقاط بازرسی (checkpoints) مدل بر اساس عملکرد در دامنه هدف استفاده شود. بخش دیگر آن بدون برچسب است.

۱. **بخش آزمون (Test Set):**

- **تعداد تصاویر :۶۴۱ تصویر و برای ارزیابی نهایی دقت مدل استفاده می‌شوند.**
- **تعداد حاشیه‌نویسی :۱۱۴۵ شیء از دو دسته**
- **ابعاد تصاویر :ابعاد تصاویر متغیر است (میانگین ۱۵۵۷ در ۹۷۲ پیکسل).**
- **توزیع دسته‌ها:**
  - **تانک: ۹۵۰ نمونه (٪۸۲.۹۷)**
  - **خودرو: ۱۹۵ نمونه (٪۱۷.۰۳)**
  -

مجموعه داده‌های دیگری که توسط مقاله اصلی استفاده شده اند مانند:

- **Cityscapes به BDD100k و Sim10k به Cityscapes نیز برای ارزیابی‌های جامع‌تر در نظر گرفته خواهند شد.**
- **دامنه منبع (Source Domain): مجموعه داده Cityscapesکه شامل ۲,۹۷۵ تصویر آموزشی از صحنه‌های شهری در شرایط آب و هوایی مناسب است.**
- **دامنه هدف(Target Domain): مجموعه داده Foggy Cityscapesکه با افزودن مه مصنوعی به تصاویر Cityscapesساخته شده است .در این پروژه از تصاویری با بالاترین چگالی مه (۰,۰۲) استفاده می‌شود تا شکاف دامنه به حداکثر برسد.**

این مجموعه داده به طور خاص برای ارزیابی توانایی مدل در انطباق از شرایط نوری مطلوب به شرایط نوری چالش‌برانگیز طراحی شده است .مجموعه داده‌های دیگری مانند Cityscapesبه BDD100k و Sim10kبه Cityscapes نیز برای ارزیابی‌های جامع‌تر در نظر گرفته خواهند شد.

٤ مراجع

[1] ***Zijing Zhao, Sitong Wei, Qingchao Chen, Dehui Li, Yifan Yang, Yuxin Peng, Yang Liu*****; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19039-19049**

[2] Siyuan Li, Luyuan Zhang, Zedong Wang, Di Wu, Lirong Wu, Zicheng Liu, Jun Xia, Cheng Tan, Yang Liu, Baigui Sun, Stan Z. Li; Masked Modeling for Self-supervised Representation Learning on Vision and Beyond; https://arxiv.org/abs/2401.00897