دانشکده مهندسی کامپیوتر

یادگیری بازنمایی علّی برای تعمیم خارج از توزیع

گزارش سمینار کارشناسی ارشد

در رشته مهندسی کامپیوتر – گرایش هوش مصنوعی و رباتیکز

نام دانشجو:

حسین رضائی

استاد راهنما:

دکتر عادل رحمانی

آذر ماه 1401


بسم الله الرحمن الرحيم

چکیده

الگوریتم های یادگیری ماشین معمولا روی فرض iid ساخته می شوند کـه داده هـای آمـوزش و آزمـون مستقل و به صورت یکسان توزیع شده اند. این فرض به این معنا اسـت کـه توزیـع داده هـای آمـوزش و داده های آزمون یکسان باشند. در دنیا واقعی به دلیل تغییرات توزیعی، این فرض به سختی برآورده می شـود کـه باعث کاهش شدید دقت این الگوریتم های کلاسیک یادگیری ماشین می شود. از طرف دیگر، الگوریتم هـای یادگیری ماشین غالبا از مدل های آماری برای مدل سازی وابستگی بین داده ها و برچسب هـا اسـتفاده مـی کنند که قصد یادگیری بازنمایی های مستقل از دامنه را دارد. با این وجود، مدل هـای آمـاری توصـیف هـای سطحی واقعیت هستند، زیرا آنها فقط به مدل سازی وابستگی به جای مکانیسم علّی ذاتی نیاز دارند. هنگامی که وابستگی با توزیع هدف تغییر می کند، مدل های آماری ممکن است در تعمـیم ناکـام باشـند. علّیـت، بـا تمرکز بر بازنمایی دانش ساختاری در مورد فرآیند تولید داده کـه اجـازه تغییـرات (interventions) را مـی دهد، می تواند به درک و رفع برخی محدودیت های روش های یادگیری ماشین فعلـی کمـک کنـد. علیـرغم موفقیت یادگیری آماری ، این مدل ها توصیفی نسبتا سطحی از واقعیت ارائه می دهند که تنها زمانی برقـرار است که شرایط آزمایشی ثابت باشد. در عوض، حوزه یادگیری علّی به دنبال مدل سازی اثر تغییرات توزیع بـا ترکیبی از یادگیری مبتنی بر داده و مفروضاتی است که قبلاً در توصیف آماری یک سیسـتم گنجانـده نشـده اند.

برای مشکل ناشی از تغییر توزیع، یعنی جایی که توزیع داده هـای آزمـون متفـاوت از داده هـای آمـوزش است، مسئله ی تعمیم خارج از توزیع مطرح می شود که در آن الگوریتم بتواند به خوبی عمل کنـد و تعمـیم خوبی روی داده های دیده نشده یعنی داده های آزمون داشته باشد. در این سمینار ما به بررسـی روش هـای مسئله ی تعمیم خارج از توزیع می پردازیم و همینطور یکی از روش های آن به نام علّیت که اخیرا به صورت گستره مورد توجه قرار گرفته است را با جزئیات دقیق تر بررسی می کنیم.

واژه های کلیدی: تعمیم خارج از توزیع، استنتاج علّـی، یـادگیری ثابـت، یـادگیری بازنمـایی، تعمـیم دامنـه

فهرست مطالب

عنوان **صفحه**

فهرست شکل ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فصل 1:

مقدمه

1-1- شرح مسأله

روش های یادگیری ماشین در بسیاری از زمینه ها مانند پردازش زبان طبیعی، بینایی کامپیوتر و ...، توانایی های زیاد خود را نشان داده اند. از طرفی، بسیاری از تحقیقات جدید، آسیب پذیری مدل های یادگیری ماشین را در مواجهه با داده ها با توزیع های مختلف، نشان داده اند. چنین مشکلی ناشی از نقض یک فرض اساسی است این فرض که داده های آموزش و آزمون به طور یکسان و مستقل توزیع شده اند، که بر اساس آن اکثر مدل های یادگیری موجود توسعه یافته اند. در بسیاری از موارد واقعی جایی که فرض iid[1] را به سختی می توان برآورده کرد، به ویژه آن دسته از کاربرد های پرمخاطره مانند مراقبت های بهداشتی، نظامی و رانندگی خودکار ، به جای تعمیم در توزیع آموزشی، توانایی تعمیم تحت تغییر توزیع اهمیت حیاتی تری دارد. بنابراین، بررسی تعمیم خارج از توزیع در هر دو حوزه دانشگاهی و صنعتی از اهمیت زیادی برخوردار است.

علیرغم اهمیت مسئله تعمیم خارج از توزیع ، روش های کلاسیک یادگیری با نظارت نمی توانند مستقیماً برای مقابله با مسئله تعمیم خارج از توزیع به کار گرفته شوند. از لحاظ نظری، یکی از اساسی ترین مفروضات یادگیری کلاسیک تحت نظارت، فرض iid[1] است که فرض می کند داده های آموزش و آزمون مستقل و به طور یکسان توزیع شده اند. از آنجایی که، تغییرات توزیعی در مسئله تعمیم خارج از توزیع اجتناب ناپذیر است، باعث می شود فرض iid[1] ارضا نشود و در نتیجه نظریه کلاسیک یادگیری غیرقابل اجر باشد. از نظر تجربی، روش های کلاسیک یادگیری با نظارت معمولاً با کمینه کردن خطا های آموزشی بهینه می شوند، که به صورت حریصانه تمام همبستگی های موجود در داده ها را برای پیش بینی بدست می آورند. اگرچه که در حالت iid[1] موثر است، اما زمانی که تغییرات توزیعی داریم باعث کاهش دقت می شود. زیرا همه همبستگی ها در توزیع های آزمایشی دیده نشده باقی نمی مانند. همانطور که در بسیاری از تحقیقات نشان داده شده است، زمانی که تغییرات توزیعی شدیدی داریم، مدل هایی که صرفاً با خطا های آموزشی بهینه شده اند، به طور چشمگیری عملکرد آن ها کاهش می یابد و گاهی حتی بدتر از حدس های تصادفی عمل می کنند. که این مسائل، ضرورت طراحی روش هایی برای مسئله تعمیم خارج از توزیع را نشان می دهند.

برای مقابله با مسئله تعمیم خارج از توزیع ، چندین مسئله ی مهم وجود دارد که باید حل شوند:

1) از آنجایی که داده های آموزشی و آزمایشی را می توان از توزیع های مختلف بدست آورد، نحوه توصیف رسمی تغییرات توزیعی هنوز یک مسئله باز است. در مقالات موجود که به موضوع

[1] independent and identically distributed

تعمیم خارج از توزیع پرداخته اند، راه های مختلفی را برای مدل سازی توزیع آزمون بالقوه اتخاذ می کنند. روش های تعمیم دامنه، عمدتاً بر سناریو های واقعی تمرکز می کنند و از داده های جمع آوری شده از حوزه های مختلف استفاده می کنند. روش های یادگیری علّی، توزیع های آموزش و آزمون را با ساختارهای علّی بیان می کنند و تغییرات توزیعی عمدتاً از تغییرات[1] یا عوامل مخدوش کننده[2] سرچشمه می گیرند. روش های یادگیری پایدار[3] تغییرات توزیعی را از طریق سو گیری انتخاب[4] معرفی می کنند.

2) نحوه طراحی یک الگوریتم که عملکرد آن برای مسئله تعمیم خارج از توزیع، مناسب باشد توجه های زیادی را به خود جلب کرده و تحقیقات زیادی روی آن انجام می شود. شاخه های زیادی از روش ها با تمرکز های تحقیقاتی مختلف، از جمله روش های یادگیری بازنمایی بدون نظارت، مدل های یادگیری با نظارت و روش های بهینه سازی وجود دارد.

3) ارزیابی عملکرد تعمیم خارج از توزیع روش های مختلف همچنان چالش برانگیز است و به مجموعه داده های خاصی با معیارهای ارزیابی مشخصی نیاز دارد، زیرا روش های کلاسیک که برای محیط های iid[5] هستند زمانی که تحت تغییرات توزیعی قرار می گیرند غیرقابل اجرا می شوند. بنابراین این موضوع، انگیزه تولید مجموعه داده ها و ارزیابی های مختلف را ایجاد می کند.

2-1- معرفی حوزه سمینار

در این گزارش، هدف ما ارائه یک بررسی سیستماتیک و جامع از تلاش های تحقیقاتی که در زمینه ی تعمیم خارج از توزیع انجام شده است می باشد که کل چرخه حیات مسئله تعمیم خارج از توزیع را پوشش می دهد. در ابتدا تعریف رسمی مشکل تعمیم خارج از توزیع را ارائه می کنیم. در مرحله بعد از میان روش های موجود، یادگیری علّی برای تعمیم دامنه را بررسی می کنیم و یکی از الگوریتم هایی که در این زمینه ارائه شده است به نام الگوریتم CIRL را مرور می کنیم. سپس روش های موجود دیگر را بر اساس موقعیت

[1] Interventions
[2] confounders
[3] Stable Learning
[4] selection bias
[5] independent and identically distributed

هایشان در کل مسیر یادگیری به سه بخش، یعنی، یادگیری بازنمـایی بـدون نظـارت، یـادگیری مـدل تحـت نظارت، و بهینه سازی، طبقه بندی می کنیم و روش های رایج برای هر دسته را به تفصیل مـورد بحـث قـرار می دهیم. در نهایت مجموعه داده های رایج و معیارهای ارزیابی را معرفی می کنیم.

3-1- ساختار گزارش

در فصل 2 ، تعریف مسئله تعمیم خارج از توزیع را ارائه می کنـیم و یکـی از روش هـای موجـود بـه نـام CIRL که در شاخه ی یادگیری علّی برای تعمیم خارج از توزیـع قـرار دارد را بیـان مـی کنـیم. در فصـل 3، روش های موجود دیگر را بر اساس موقعیت هایشان در کل مسیر یادگیری در سه بخـش ارائـه مـی کنـیم و مجموعه داده های رایج و معیارهای ارزیابی که برای مسئله ی تعمیم خـارج از توزیـع اسـتفاده مـی شـوند را معرفی می کنیم. و در فصل آخر نتیجه گیری نهایی از نتایج این تحقیق ارائه شده است، و برخی مسیر هـای مناسب برای تحقیق و توسعه بیشتر در این زمینه را پیشنهاد می دهیم.

فصل 2:

تعاريف و مفاهيم مبنايي

2-1- مقدمه

در سال های اخیر با پیچیده شدن مسائل یادگیری ماشین در دنیای واقعی، مسئله ی تعمیم خارج از توزیع، مشکلاتی را برای شبکه های عصبی عمیق مبتنی بر فرض iid[1] به وجود آورده است. به کارگیری مستقیم مدلی که روی دامنه ی مبدا آموزش دیده است به یک دامنه هدف دیده نشده با توزیع متفاوت، باعث کاهش عملکرد الگوریتم می شود که باید به آن توجه شود. برای مقابله با این موضوع مسئله ی تعمیم خارج از توزیع مطرح می شود که یکی از روش های آن، یادگیری بازنمایی علّی است.

به جز یادگیری بازنمایی علّی برخی از کارها از جمله فرا-یادگیری[2]، یادگیری بازنمایی ثابت[3] و تقویت دامنه[4] انجام شده است. اگرچه نتایج آن ها قابل قبول است، اما یک مشکل ذاتی در آن ها وجود دارد. این روش ها صرفاً سعی در جبران مشکلات ناشی از داده های خارج از توزیع و مدل سازی وابستگی آماری بین داده ها و برچسب ها بدون توضیح مکانیسم های علّی ذاتی دارند. در مقاله ی [1] بحث شده است که چنین اقداماتی ممکن است کافی نباشد، و تعمیم مناسب در خارج از فرض iid[1] مستلزم یادگیری وابستگی آماری صرف بین متغیر ها نیست، بلکه یک مدل علّی ذاتی نیاز است. به عنوان مثال، در یک مسئله طبقه بندی تصویر، به احتمال زیاد همه زرافه ها روی چمن هستند و وابستگی آماری بالایی را نشان می دهند که می تواند به راحتی مدل را گمراه کند و زمانی که پس زمینه در دامنه هدف متفاوت است، پیش بینی اشتباهی انجام دهد. از این گذشته، ویژگی های زرافه ها مانند سر، گردن و ... به جای پس زمینه، یک زرافه را «زرافه» می کند که باید به درستی یاد گرفته شوند و تعیین کننده برچسب هدف باشند.

2-2- تعریف مسئله

فرض کنید که χ فضای ویژگی ها، و y فضای برچسب ها باشد. یک مدل دارای پارامتر[5] به صورت $f_\theta: \chi \rightarrow y$ تعریف می شود که به عنوان یک تابع نگاشت از ویژگی های اصلی به برچسب با پارامتر قابل یادگیری θ عمل می کند. یک تابع زیان به صورت $l: y \times y \rightarrow \mathbb{R}$ تعریف می شود که فاصله بین برچسب پیش بینی شده و برچسب واقعی[6] را اندازه می گیرد. حال با توجه به این توضیحات می توانیم مسئله کلاسیک یادگیری با نظارت را تعریف کنیم.

---

[1] independent and identically distributed
[2] Meta-Learning
[3] Invariant Representation Learning
[4] Domain Augmentation
[5] parametric
[6] ground-truth

مسئله 1 (یادگیری با نظارت). یک مجموعه ایی از n نمونه آموزشی بـه فـرم $\{(x_1, y_1), \dots, (x_n, y_n)\}$ داده شده است که دارای توزیع آموزشی $P_{tr}(X, Y)$ هستند. یک مسئله ی یادگیری با نظارت این است که یک مدل بهینه ی f_θ^* را به گونه ایی پیدا کنیم که بتوانـد تعمـیم خـوبی روی داده هـای آزمـون کـه دارای توزیع $P_{te}(X, Y)$ هستند، داشته باشد:

$$f_\theta^* = \underset{f_\theta}{argmin}\, E_{X,Y\sim P_{te}}[l(f_\theta(X), Y)] \qquad (1)$$

2-2-1- یادگیری *iid*[1]

الگوریتم های یادگیری کلاسیک معمولاً فرض می کنند که نمونه های آموزشی و نمونه های آزمون هـر دو تحقق iid[1] از یک توزیع پایه مشترک هستند، که به معنای $P_{tr}(X, Y) = P_{te}(X, Y)$ اسـت. بـر اساس چنین فرضیه ای، ERM[2] که میانگین زیان نمونه های آموزشی را بـه حـداقل مـی رسـاند، مـی تواند یک مدل بهینه را به دست آورد که با موفقیت به توزیع آزمون تعمیم یابد. در واقع، ERM عبـارت زیر را کمینه می کند:

$$l_{ERM} = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i) \qquad (2)$$

2-2-2- تعمیم خارج از توزیع

ویژگی خوب ارائه شده توسط فرض iid[1] زمینه مناسبی را برای توسعه مـدل هـای یـادگیری زیـادی در چند دهه اخیر فراهم کرده است. با این حال، در موارد واقعی، توزیـع داده هـای آزمـون ممکـن اسـت از توزیع داده های آموزشی متفاوت باشد، یعنی $P_{tr}(X, Y) \neq P_{te}(X, Y)$. تغییر توزیـع مـی توانـد بـه دلایل زیادی مانند تغییرات زمانی/مکانی داده ها یا سوگیری[3] انتخاب نمونه در فرآینـد جمـع آوری داده ها باشد. در هر صورت، مسئله 1 را پیچیده تر از سناریوی یادگیری $i.i.d$ می کند. علاوه بر ایـن، توزیـع داده های آزمون که با آن مواجه می شویم ممکن است به دلیل ماهیت برنامه هایی مانند سناریو آنلایـن مبتنی بر جریان[4] که در آن داده های آزمون در آینده تولید می شوند ناشناخته باشد. بـه طـور خلاصـه، مسئله تعمیم خارج از توزیع را می توان به عنوان نمونه ای از مسئله یادگیری با نظارت تعریف کـرد کـه

[1] independent and identically distributed
[2] Empirical Risk Minimization
[3] *bias*
[4] *stream*

در آن توزیع آزمون $P_{te}(X,Y)$ نسبت به توزیع آموزشی $P_{tr}(X,Y)$ دچار تغییر[1] می شود و در طول مرحله آموزش ناشناخته باقی می ماند.

به طور کلی، مسئله تعمیم خارج از توزیع امکان پذیر نیست، مگر اینکه برخی از فرضیات را در مورد چگونگی تغییر توزیع داده های آزمون داشته باشیم. در میان تغییرات توزیعی مختلف، تغییر متغیر[2] متداول ترین نوع آن است و به طور گسترده مورد مطالعه قرار گرفته است. در واقع تغییر متغیر[3] به صورت $P_{tr}(Y|X) = P_{te}(Y|X)$ و $P_{tr}(X) \neq P_{te}(X)$ تعریف می شود، که به این معنی است که توزیع حاشیه ای X از فاز آموزش به فاز آزمون تغییر می کند در حالی که مکانیسم تولید برچسب بدون تغییر باقی می ماند. در این سمینار، تمرکز ما بر روی تغییر متغیر[3] است و کار های مختلفی را که در این زمینه انجام شده است، بررسی می کنیم.

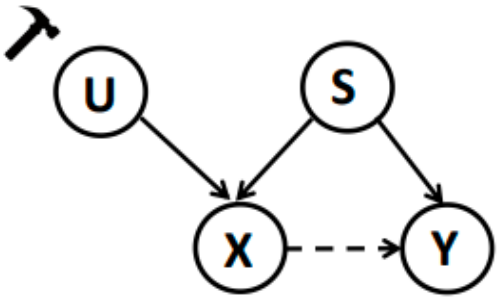
2-3- یادگیری بازنمایی علّی برای مسئله ی تعمیم خارج از توزیع

در این سمینار ما یک مدل علّی ساختاری[3] را برای مسئله تعمیم خارج از توزیع با هدف بدست آوردن مکانیزم علّی ذاتی بین داده ها و برچسب ها و همینطور بدست آوردن توانایی تعمیم بهتر، معرفی می کنیم. اطلاعات مربوط به هر دسته[4] در داده ها به عنوان عوامل علّی[5] در نظر گرفته می شود که رابطه آنها با برچسب، مستقل از دامنه است، به عنوان مثال، "شکل" در مسئله تشخیص رقم. در حالی که اطلاعات مستقل از دسته[6] به عنوان عوامل غیر علّی[7] در نظر گرفته می شود، که عموماً اطلاعات مربوط به دامنه است، به عنوان مثال، "سبک دست خط" در تشخیص رقم. هر داده خام X از ترکیبی از عوامل علّی S و عوامل غیرعلّی U ساخته شده است و فقط S ها به طور علّی بر برچسب دسته Y تأثیر می گذارد، همانطور که در شکل 1 نشان داده شده است. هدف استخراج عوامل علی S از ورودی خام X و سپس بازسازی مکانیسم های علّی ثابت[8] است، که می توان با کمک ایجاد تغییرات علّی[9] $P(Y|do(U),S)$ ، این کار را انجام داد. $do(\cdot)$ نشان دهنده ایجاد تغییرات بر روی متغیر ها است.

---

[1] *shift*
[2] covariate shift
[3] *SCM; structural causal model*
[4] *category-related information*
[5] *causal factors*
[6] *category*
[7] *non-causal factors*
[8] *invariant causal mechanisms*
[9] *causal intervention*

متأسفانه، ما نمی توانیم مستقیماً ورودی خام را به عنوان $X = f(S, U)$ بیان کنیم، زیرا عوامل علی/غیر علی معمولاً مشاهده نمی شوند و نمی توان آنها را فرمول بندی کرد که این موضوع استنتاج علّی را چالش برانگیز می کند.



شکل 1: SCM برای *Domain Generalization (Out-Of-Distribution Generalization)* پیکان های توپر نشان می دهد که گره والد باعث و عامل یک فرزند است در حالی که پیکان های خط تیره به معنای وجود وابستگی آماری است.

عوامل علّی S انتظار می رود بر اساس [1], [2], [3] سه ویژگی را برآورده کنند:

1) جدا و مجزا از عوامل غیر علّی U باشند.
2) فاکتورسازی S باید به صورت توام مستقل باشد.
3) از نظر علّی برای مسئله طبقه بندی $X \rightarrow Y$ حاوی تمام اطلاعات علّی کافی باشد.

ترکیب با U باعث می شود که S حاوی اطلاعات غیر علّی اساسی باشد، در حالی که فاکتورگیری وابسته توام، S را زائد می کند و منجر به از دست رفتن برخی از اطلاعات علّی اساسی می شود. در مقابل، عوامل علّی S، عواملی ایده آل هستند که همه الزامات را برآورده می کنند. با الهام از این، یک الگوریتم یادگیری بازنمایی الهام گرفته از علیت[1] پیشنهاد شده است که بازنمایی های آموخته شده را وادار می کند تا ویژگی های بالا را داشته باشند و سپس از هر بعد از بازنمایی ها برای تقلید از فاکتورسازی عوامل علّی که توانایی تعمیم قوی تری دارند، بهره برداری می کند.

به طور خلاصه، برای هر ورودی، ابتدا از یک ماژول ایجاد تغییر علّی[2] برای جداسازی عوامل علّی S از عوامل

[1] *CIRL; Causality Inspired Representation Learning*
[2] *causal intervention module*

غیرعلّی U با تولید داده های جدید با اطلاعات مرتبط با دامنه‌ی دچار تغییر شده[1] استفاده می‌کنند. داده های تولید شده دارای عوامل غیر علی متفاوت U هستند اما عوامل علی S در مقایسه با موارد اصلی یکسان هستند، بنابراین بازنمایی ها مجبور می شوند ثابت بمانند. علاوه بر این، آن ها یک ماژول فاکتورسازی را پیشنهاد می‌کنند که هر بُعد از بازنمایی ها را به طور توام مستقل می کند و سپس می‌توان از آن برای تقریب عوامل علی استفاده کرد. علاوه بر این، برای اینکه الگوریتم پیشنهادی از نظر علّی برای طبقه بندی کافی باشد، یک ماژول ماسک متخاصم[2] طراحی می‌کنند که به‌طور مکرر ابعادی را که حاوی اطلاعات علی نسبتاً کمتری هستند شناسایی می کند و آنها را مجبور می کند از طریق یادگیری خصمانه بین یک پوشش‌دهنده[3] و مولد بازنمایی، حاوی اطلاعات علّی بیشتر و جدید باشند.

1-3-2- متد

در این بخش، مطابق شکل 1، تعمیم دامنه را از نمای علّی با یک مدل علی ساختاری کلی در نظر می گیرند و نشان می دهند که مکانیسم های علّی ذاتی (که به عنوان توزیع های شرطی بیان می شوند) می توانند در صورت داشتن عوامل علّی، امکان پذیر باشند. با این حال، همانطور که در قسمت قبل گفته شد، بازیابی دقیق عوامل علّی دشوار است زیرا آنها غیر قابل مشاهده هستند. بنابراین، پیشنهاد می شود که بازنمایی های علّی را بر اساس ویژگی های عوامل علّی به عنوان یک تقلید، در حالی که توانایی تعمیم بهتری را دارا می باشند، یاد بگیریم.

1-1-3-2- تعمیم دامنه (تعمیم خارج از توزیع) از نگاه علّی

کارهایی که برای تعمیم دامنه انجام شده است بر مدل سازی وابستگی آماری بین ورودی های مشاهده شده و برچسب های مربوطه، یعنی $P(X, Y)$ متمرکز است که در دامنه ها مختلف، متفاوت در نظر گرفته می شود. برای به دست آوردن یک وابستگی ثابت، عموماً توزیع را وادار می کنند که به صورت حاشیه ای یا شرطی ثابت-دامنه[4] باشد. به عنوان مثال، فاصله بین دامنه ها را در $P(X)$ یا $P(X \mid Y)$ کمینه می کنند. با این حال، از آنجایی که وابستگی آماری نمی تواند مکانیسم علی ذاتی بین ورودی ها و برچسب ها را توضیح دهد، تمایل به تغییر با دامنه را دارد. بنابراین، وابستگی ثابت یاد گرفته شده در میان دامنه

---

[1] *perturbed domain-related information*
[2] *adversarial mask module*
[3] *masker*
[4] *domain-invariant marginally or conditionally*

های مبدا ممکن است همچنان در یک دامنه هدف دیده نشده به خوبی عمل نکنند. در همین حال، مکانیسم های علّی معمولاً در همه ی دامنه ها، پایدار و ثابت می مانند [1]. ابتدا ارتباط بین علّیت و وابستگی آماری را همانطور که رایشنباخ [2] در اصل 1 ادعا کرده است، بیان می کنیم.

اصل 1. اصل علت مشترک[1]: اگر دو متغییر قابل مشاهده X و Y از نظر آماری وابسته باشند، آنگاه یک متغیر S وجود دارد که به طور علّی بر هر دو تأثیر می گذارد و تمام وابستگی آن ها را با مستقل کردن آنها تحت شرط S، بیان می کند.

بر اساس اصل 1 آن ها مدل علّی ساختاری زیر را برای توصیف مسئله تعمیم دامنه، بیان می کنند:

$$X \coloneqq f(S, U, V_1), S \perp U \perp V_1,$$
$$Y \coloneqq h(S, V_2) = h(g(x), V_2), V_1 \perp V_2. \qquad (10)$$

که X و Y به ترتیب نشان دهنده ی تصاویر ورودی و برچسب مربوطه ی آن ها است. S نشان دهنده ی عامل های علّی است که به صورت علّی X و Y را یعنی اطلاعات مربوط به دسته[2] مانند "شکل" در مسئله ی تشخیص عدد است. درحالی که U نشان دهنده ی فاکتور های غیر علّی است که به صورت علّی فقط روی X اثر می گذارند. فاکتور های غیر علّی در واقع اطلاعات مرتبط با دامنه، مثل "سبک نوشتن" می باشند. V_1 و V_2، متغییر های نویز هستند که به صورت توام مستقل می باشند. f، h و g را می توان به عنوان توابع ساختاری ناشناخته در نظر گرفت. بنابراین، برای هر توزیع $P(X,Y)\epsilon\mathcal{P}$، اگر فاکتور های علّی S داده شده باشند یک توزیع شرطی کلی $P(Y|S)$ یعنی یک مکانیزم علّی ثابت وجود دارد. طبق صحبت فوق، اگر ما بتوانیم به فاکتور های علّی دسترسی پیدا کنیم آنگاه بدست آوردن مکانیزم علّی ایی که بتواند به خوبی خارج از فرض iid[3] با بهینه سازی h تعمیم یابد آسان است:

$$h^* = \underset{h}{argmin} E_P[l(h(g(X)), Y)] = \underset{h}{argmin} E_p[l(h(S), Y)], \qquad (11)$$

که $l(.,.)$ تابع زیان آنتروپی متقابل[4] است.

مشکلی که وجود دارد این است که ما فاکتور های علّی S را از قبل نداریم و تنها تصاویر خام X را داریم که

---

[1] *Common Cause Principle*
[2] *category-related information*
[3] independent and identically distributed
[4] *cross-entropy loss*

به صورت کلی بدون ساختار هستند. بازسازی مستقیم عوامل علّی و نیز مکانیسم ها به نوعی غیر عملی است، زیرا آنها غیرقابل مشاهده هستند و به صورت مشخص تعریف نشده اند. آنچه که واضح است این است که فاکتور های علّی باید از الزامات خاصی پیروی کنند. [1، 3] بیان می کنند که فاکتور های علّی باید به صورت توام مستقل باشند، همانطور که در اصل 2 بیان شده است.

اصل 2. اصل مکانیزم های علی مستقل[1]:

توزیع شرطی هر متغیر به شرط علل آن (یعنی مکانیسم آن) مکانیسم های دیگر را تحت تأثیر قرار نمی دهد.

از آنجایی که S در معادله (10) نشان دهنده مجموعه همه فاکتور های علّی $\{s_1, s_2, \dots, s_N\}$ است، این اصل به ما می گوید که:

1) ایجاد تغییر روی یک مکانیسم $P(s_i|PA_i)$ هیچ یک از مکانیسم های دیگر $P(s_j|PA_j)$، $i \neq j$ را تغییر نمی دهد (PA_i نشان دهنده ی پدر s_i در گراف علّی است) که می تواند به عنوان اطلاعات علّی در نظر گرفته شود که s_i حاوی است زیرا S پیش از این، گره ریشه است).
2) دانستن برخی مکانیسم های دیگر $P(s_i|PA_i)$ اطلاعاتی در مورد مکانیسم $P(s_j|PA_j)$ به ما نمی دهد. بنابراین، می توان توزیع توام فاکتور های علّی را به صورت شرطی بیان کرد که به فاکتورسازی علّی[2] اشاره دارد:

$$P(s_1, s_2, \dots, s_N) = \sum_{i=1}^{N} P(s_i|PA_i), \qquad (12)$$

بنابراین خاطر نشان می شویم که فاکتور های علّی S باید بر اساس اصل 1 و 2، سه ویژگی پایه را دارا باشند:

- فاکتور های علّی S باید جدا و مجزا از فاکتور های غیر علّی U باشند یعنی $S \perp U$. بنابراین انجام یک تغییر بر روی U نباید باعث تغییر در S شود.
- فاکتور سازی $s_1, s_2, \dots, s_N$ باید به صورت توام مستقل باشد و هیچکدام از آن ها حاوی اطلاعاتی از بقیه فاکتور ها نباشد.
- فاکتور های علّی S باید از نظر علّی برای مسئله ی طبقه بندی $X \rightarrow Y$ کافی باشند. یعنی حاوی

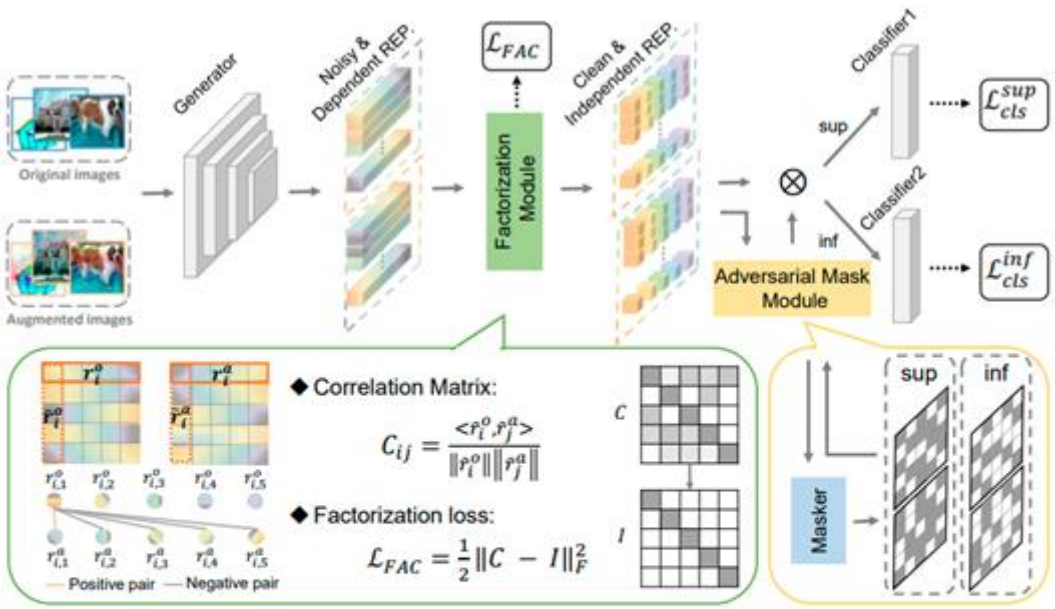
[1] *ICM; Independent Causal Mechanisms*
[2] *causal factorization*

اطلاعاتی باشند که بتواند تمام وابستگی های آماری را توضیح دهد.

بنابراین، به جای بازسازی مستقیم عوامل علّی، آن ها پیشنهاد به یادگیری بازنمایی های علّی به عنوان جایگزینی با وادار کردن آنها به داشتن ویژگی های یکسان با عوامل علی، می کنند. جزئیات این موضوع را در بخش بعد توضیح خواهیم داد.

2-1-3-2- یادگیری بازنمایی با الهام از علّیت[1]

در این بخش، الگوریتم یادگیری بازنمایی پیشنهادی الهام گرفته شده از علّیت *[4]* مورد بحث در بالا را نشان می دهیم که از سه ماژول تشکیل شده است: ماژول ایجاد تغییرات علّی[2]، ماژول فاکتور سازی علّی[3]، و ماژول ماسک متخاصم[4]. کل الگوریتم در شکل 3 نشان داده شده است.



شکل 2: چارچوب کلی الگوریتم *CIRL*. ابتدا تصاویر تقویت شده (*augmented images*) را به وسیله ماژول ایجاد تغییرات علّی با ایجاد تغییرات بر عوامل غیرعلّی تولید می کند. هر دو بازنمایی تصاویر اصلی و تقویت شده به ماژول فاکتور سازی ارسال می شوند، که یک تابع زیان فاکتور سازی را تحمیل می کند تا بازنمایی ها از فاکتور های غیر علّی جدا و مجزا و همچنین به طور توام مستقل شوند. در نهایت، ماژول ماسک متخاصم، رقابت را بین مولد و یک پوشش دهنده (*masker*) ایجاد می کند و بازنمایی ها را به طور علّی کافی برای طبقه بندی ارائه می دهد.

---

[1] *Causality Inspired Representation Learning*
[2] *causal intervention module*
[3] *causal factorization module*
[4] *adversarial mask module*

ماژول ایجاد تغییرات علّی[1]

در ابتدا هدف جدا سازی فاکتور های علّی S از ترکیب آن ها با فاکتور های غیر علّی U بـه وسـیله ی ایجـاد تغییرات علّی است. اگرچه فرم صریح استخراج کننده فاکتور های علّی $g(.)$ در عبارت (11) بـه صـورت کلی ناشناخته است اما با توجه به دانش قبلی می دانـیم کـه فـاکتور هـای علّـی S بایـد هنگـام ایجـاد تغییرات بر روی U، ثابت بمانند (یعنی $P(S \mid do(U))$). در حالی کـه از تحقیقـات مربـوط بـه تعمـیم دامنه (تعمیم خارج از توزیع) می دانیم که بعضی از اطلاعات مرتبط با دامنه نمـی تواننـد دسـته تصـویر ورودی را مشخص کنند که به عنوان فاکتور های غیر علّی در نظر گرفته مـی شـوند و توسـط بعضـی از تکنیک ها تسخیر می شوند. به طور مثال تبدیل فوریه یک ویژگی مناسب در این زمینه دارد: مولفه فـاز طیف فوریه معنای سطح بالای سیگنال اصلی را حفظ می کند، در حالی که مولفه دامنه حاوی اطلاعـات آماری سطح پایین است. بنابراین، روش $CIRL$ ایجاد تغییرات را بر روی U با برهم زدن اطلاعات دامنه انجام می دهد در حالی که اطلاعات فاز را بدون تغییر نگه می دارد. با توجه به یک تصویر ورودی اصـلی x^o، تبدیل فوریه آن را می توان به صورت زیر بیان کرد:

$$\mathcal{F}(x^o) = \mathcal{A}(x^o) \times e^{-j \times \mathcal{P}(x^o)}, \tag{13}$$

که $\mathcal{A}(x^o)$ و $\mathcal{P}(x^o)$ به ترتیب نشان دهنده ی مولفه ی دامنه و فاز است. تبدیل فوریـه ی $\mathcal{F}(.)$ و معکوس آن $\mathcal{F}^{-1}(.)$ می تواند به صورت موثر توسط الگوریتم FFT محاسـبه شـوند. سـپس اطلاعـات دامنه را از طریق درون یابی خطی بین طیف های دامنه تصویر اصـلی x^o و تصـویر $(x')^o$ کـه بـه طـور تصادفی از دامنه های مبدا دلخواه نمونه برداری شده است، تغییر می دهند:

$$\hat{\mathcal{A}}(x^o) = (1-\lambda)\mathcal{A}(x^o) + \lambda\mathcal{A}((x')^o), \tag{14}$$

که $\lambda \sim U(0,\eta)$ و η میزان تغییرات را کنترل می کند. سپس طیف هـای دامنـه تغییـر یافتـه را بـا مولفه فاز اصلی ترکیب می کنند تا تصویر تقویت شده x^a را با تبدیل فوریه معکوس ایجاد کنند:

$$\mathcal{F}(x^a) = \hat{\mathcal{A}}(x^o) \times e^{-j \times \mathcal{P}(x^o)}, \quad x^a = \mathcal{F}^{-1}\big(\mathcal{F}(x^a)\big). \tag{15}$$

[1] *causal intervention module*

مولـد بازنمـایی پیـاده سـازی شـده توسـط یـک مـدل CNN را بـا $\hat{g}(.)$ و بازنمـایی هـا را بـا $r = \hat{g}(x)\epsilon\mathbb{R}^{1\times N}$ نشان می دهیم که در آن N تعداد ابعاد است. برای شبیه سازی فاکتور های علّی که بـا ایجاد تغییرات در U ثابت می مانند، $\hat{g}$ را بهینه سازی می کنیم تا بازنمایی ها را در هـر بُعـد بـا ایجـاد تغییرات بالا بدون تغییر نگه دارند:

$$\max_{\hat{g}} \frac{1}{N}\sum_{i=1}^{N} COR(\tilde{r}_i^o, \tilde{r}_i^a), \qquad (16)$$

کـه $\tilde{r}_i^o$ و $\tilde{r}_i^a$ بـه ترتیـب نشـان دهنـده ی $Z\text{-}score$ نرمـال شـده i امـین سـتون $R^o = [(r_1^o)^T, \cdots, (r_B^o)^T]^T \in \mathbb{R}^{B\times N}$ و $R^a = [(r_1^a)^T, \cdots, (r_B^a)^T]^T$ اسـت. $B \in \mathbb{Z}_+$ انـدازه دسته ها[1]، $r_i^o = \hat{g}(x_i^o)$ و $r_i^a = \hat{g}(x_i^a)$ بـرای $i\epsilon\{1, \cdots, B\}$ اسـت. از یـک تـابع COR بـرای اندازه گیری همبستگی بازنمایی ها قبل و بعد از ایجاد تغییر استفاده می کنند. بنابراین، می توان اولـین مرحله شبیه سازی فاکتور های علّی S با بازنمایی های R را با مستقل کردن آنها از U درک کرد.

ماژول فاکتور سازی علّی[2]

همـانطور کـه در بخـش "تعمـیم دامنـه از نگـاه علّـی" گفتـه شـد کـه فـاکتور سـازی فـاکتور هـای علـی $s_1, s_2, \dots, s_N$ باید به صورت توام مستقل باشند طوری که هیچ یک از آنهـا حـاوی اطلاعـات دیگـران نیست. بنابراین، در الگوریتم $CIRL$ تلاش می کنند که هر دو بعد از بازنمایی هـا را مسـتقل از یکـدیگر کنند:

$$\min_{\hat{g}} \frac{1}{N(N-1)}\sum_{i\neq j} COR\left(\tilde{r}_i^o, \tilde{r}_j^a\right), i \neq j, \qquad (17)$$

برای کم کردن هزینه محاسبات محدودیت های درون R^a یا R^o را در نظـر نگرفتـه انـد. بـرای تحقـق اهداف بهینه سازی عبارات (16) و (17)، یک ماتریس همبستگی C به صورت زیر در نظر می گیرند:

---

[1] *batch size*
[2] *causal factorization module*

$$C_{ij} = \frac{< \tilde{r}_i^o, \tilde{r}_j^a >}{\|\tilde{r}_i^o\| \|\tilde{r}_j^a\|}, i, j \in 1,2,\cdots,N, \quad (18)$$

که $<\cdot>$ نشان دهنده ی عملگر ضرب داخلی است. بنابراین بُعد های یکسان R^a و R^o می توانند به عنوان جفت نمونه های مثبت در نظر گرفته شوند که باید همبستگی را به حداکثر برسانند. در حالی که بُعد های متفاوت می توانند به عنوان جفت نمونه های منفی در نظر گرفته شوند که باید همبستگی را به حدقل برسانند. بر این اساس، آن ها یک تابع زیان $\mathcal{L}_{Fac}$ در نظر می گیرند که به صورت زیر بیان می شود:

$$\mathcal{L}_{Fac} = \frac{1}{2}\|C - I\|_F^2. \quad (19)$$

معادله (19) می تواند عناصر قطری ماتریس همبستگی C را به صورت تقریبی 1 کند، به این معنی که بازنمایی های قبل و بعد از ایجاد تغییر نسبت به فاکتور های غیر علّی، ثابت هستند. این نشان می دهد که می توانیم به طور مؤثر فاکتور های علّی را از ترکیب آن ها با فاکتور های غیر علّی جدا کنیم. علاوه بر این، عناصر غیر قطری C را به 0 نزدیک می کند، یعنی ابعاد بازنمایی ها را به صورت توام مستقل می کند. بنابراین، با به حداقل رساندن $\mathcal{L}_{Fac}$، می توانیم بازنمایی های دارای نویز و وابسته را به بازنمایی هایی تمیز و مستقل تبدیل کنیم که دو ویژگی اول فاکتور های علّی ایده آل را برآورده می کنند.

ماژول ماسک متخاصم[1]

برای موفقیت در مسئله ی طبقه بندی $X \to Y$، بازنمایی ها باید به صورت علّی مناسب باشند به طوری که شامل همه ی اطلاعات کمک کننده باشند. سرراست ترین راه بکار گرفتن برچسب های با نظارت y در چند دامنه مبدا است:

$$\mathcal{L}_{cls} = l\big(\hat{h}\big(\hat{g}(x^o)\big), y\big) + l\big(\hat{h}\big(\hat{g}(x^a)\big), y\big) \quad (20)$$

که $\hat{h}$ طبقه بندی کننده است. با این حال، این راه ساده نمی تواند تضمین کند که هر بُعد از بازنمایی های

[1] *adversarial mask module*

یاد گرفته شده ما مهم است، یعنی حاوی اطلاعات علّی اساسی کافی برای طبقه بندی است. ممکن است ابعادی با ارزش پایین تری[1] وجود داشته باشد که اطلاعات علّی نسبتاً کمتری را دارا می باشند و در نتیجه سهم کوچکی در طبقه بندی دارند. بنابراین، نویسندگان روش *CIRL* پیشنهاد می کنند که این ابعاد را شناسایی کرده و آنها را برای مشارکت بیشتر به اجرا درآورد. از آنجایی که ابعاد نیز باید به کمک ماژول فاکتور سازی ارائه شده به صورت توام مستقل باشند، ابعاد با ارزش پایین تر[2] شناسایی شده به گونه ای ارائه می شوند که حاوی اطلاعات علّی بیشتر و جدید باشند که در سایر ابعاد گنجانده نشده است، که باعث می شود کل بازنمایی ها از جهت علّی غنی تر باشند.

بنابراین، برای تشخیص ابعاد با ارزش پایین تر، آن ها یک ماژول ماسک متخاصم طراحی می کنند. یک ماسک مبتنی بر شبکه عصبی می سازند که با $\widehat{w}$ نشان داده می شود تا سهم هر بعد را یاد بگیرد و ابعاد مربوط به بزرگترین نسبت $\kappa \in (0,1)$ به عنوان ابعاد با ارزش بالاتر[3] در نظر گرفته می شود. در حالی که بقیه به عنوان ابعاد با ارزش پایین تر در نظر گرفته می شوند:

$$m = Gumbel\_Softmax(\widehat{w}(r), \kappa N) \epsilon \mathbb{R}^N, \qquad (21)$$

که آن ها از ترفند مشتق پذیر رایج *Gumbel_Softmax* برای نمونه برداری از ماسکی با مقادیر κN نزدیک به 1 استفاده می کنند. با ضرب بازنمایی های یاد گرفته شده در ماسک های بدست آمده m و $1\text{-}m$ می توان به ترتیب ابعاد با ارزش بالاتر[4] و پایین تر[5] بازنمایی ها را بدست آورد. سپس، آنها را به دو طبقه بندی کننده مختلف $\hat{h}_1, \hat{h}_2$ وارد می کنند. معادله (20) را می توان به صورت زیر بازنویسی کرد:

$$\begin{aligned} \mathcal{L}_{cls}^{sup} &= l(\hat{h}_1(r^o \odot m^o), y) + l(\hat{h}_1(r^a \odot m^a), y), \\ \mathcal{L}_{cls}^{inf} &= l(\hat{h}_2(r^o \odot (1 - m^o)), y) + l(\hat{h}_2(r^a \odot (1 - m^a)), y), \end{aligned} \qquad (22)$$

آن ها ماسک کننده[6] را با کمینه کردن $\mathcal{L}_{cls}^{sup}$ و بیشینه کردن $\mathcal{L}_{cls}^{inf}$ ، بهینه می کنند. درحالی که مولد $\hat{g}$

[1] *inferior*
[2] *inferior dimensions*
[3] *superior dimensions*
[4] *superior dimensions*
[5] *inferior dimensions*
[6] *masker*

و طبقه بندی کننده های $\hat{h}_1, \hat{h}_2$ را با کمینه کردن دو تابع زیان با نظارت، بهینه می کنند.

ماژول ماسک متخاصم به طور دقیق می تواند ابعاد با ارزش پایین تر[2] را شناسایی کنند چون:

1) برای یک $\hat{h}_2$ بهینه شده با هدف کمینه کردن $\mathcal{L}_{cls}^{inf}$ بر اساس ابعاد ماسک شده[1] موجود، یادگیری m برای انتخاب ابعاد برای به حداکثر رساندن $\mathcal{L}_{cls}^{inf}$ می تواند ابعاد با ارزش پایین تر که مشارکت کمتری دارند را پیدا کند.
2) مجموعه ابعاد با ارزش بالاتر[1] و پایین تر[2] مکمل یکدیگر هستند به طوری که اگر یک بُعد به عنوان بُعد با ارزش بالاتر تلقی نشود، به عنوان یک بُعد با ارزش پایین تر تلقی می شود، بنابراین انتخاب ابعاد با ارزش بالاتر به انتخاب ابعاد با ارزش پایین تر کمک می کند.

علاوه بر این، در مقایسه با بهینه سازی معادله (20)، ماژول ماسک متخاصم همراه با ماژول فاکتورسازی علّی می تواند به تولید بازنمایی هایی از جهت علّی کافی تر کمک کند، زیرا با بهینه سازی $\hat{g}$ برای کمینه کردن $\mathcal{L}_{cls}^{inf}$ و $\mathcal{L}_{Fac}$، ابعاد با ارزش پایین تر مجبور می شوند اطلاعات علّی بیشتری را حمل کنند و مستقل از ابعاد با ارزش بالاتر موجود، باشند. در نهایت، بازنمایی های آموخته شده با «جایگزینی» مکرر بازنمایی های با ارزش پایین تر به عنوان بازنمایی های با ارزش بالاتر جدید، به حالتی که از جهت علّی کافی تر باشند نزدیک می شوند.

هدف کلی بهینه سازی $CIRL$ به صورت زیر خلاصه می شود:

$$\min_{\hat{g},\hat{h}_1,\hat{h}_2} \mathcal{L}_{cls}^{sup} + \mathcal{L}_{cls}^{inf} + \tau\mathcal{L}_{Fac}\,, \quad \min_{\hat{w}} \mathcal{L}_{cls}^{sup} - \mathcal{L}_{cls}^{inf}\,, \tag{23}$$

که در آن τ یک پارامتر برای تنظیم شدت تاثیر $\mathcal{L}_{Fac}$ است. توجه داشته باشید که کل بازنمایی r و طبقه بندی کننده $\hat{h}_1$ در طول استنتاج[2] استفاده می شوند.

2-4- نتیجه‌گیری

در این فصل ابتدا، مسئله تعمیم خارج از توزیع را به صورت کلی بیان کردیم و ارتباط و تفاوت آن با مسئله یادگیری با فرض iid را نشان دادیم. سپس، دو اصل اساسی از یادگیری علّی را بیان کردیم و بر اساس آن یک الگوریتم یادگیری بازنمایی الهام گرفته شده از علّیت به نام CIRL را معرفی کردیم.

---

[1] *masked*
[2] *inference*

فصل 3:

مروري بر كارهاي مرتبط

3-1- مقدمه

برای مقابله با چالش های ناشی از تغییرات توزیعی که ناشناخته است، تلاش های زیادی در تعمیم خارج از توزیع انجام شده است. تکنیک های استفاده شده از جمله علّیت، یادگیری بازنمایی، تکنیک های مبتنی بر ساختار و مبتنی بر بهینه سازی، بسیار متفاوت از هم هستند. در این سمینار سعی شده است که به تفصیل این تکنیک ها توضیح داده شوند.

به طور کلی، مسئله یادگیری تحت نظارت تعریف شده در معادله 1 را می توان به سه جزء به صورت نسبتا مستقل تقسیم کرد:

1) بازنمایی ویژگی های X (به عنوان مثال $g(X)$).
2) تابع نگاشت $f_\theta(X)$ از ویژگی های X (یا $g(X)$) به برچسب Y، که عموماً به عنوان مدل یا سوگیری استقرایی[1] نیز شناخته می شود.
3) فرمول بندی هدف بهینه سازی.

بنابراین، روش های موجود را بر اساس موقعیت آنها در کل مسیر یادگیری به سه بخش دسته بندی می کنیم:

1) **یادگیری بازنمایی بدون نظارت برای تعمیم خارج از توزیع:** شامل یادگیری بازنمایی تفکیک شده[2] و یادگیری بازنمایی علّی می باشد که از تکنیک های یادگیری بازنمایی بدون نظارت (مثل *variational Bayes*) برای جاسازی[3] کردن دانش قبلی در فرایند یادگیری، استفاده می کند.
2) **یادگیری مدل با نظارت برای تعمیم خارج از توزیع:** شامل یادگیری علّی، یادگیری پایدار[4] و تعمیم دامنه می باشد که معماری مدل های مختلف و استراتژی های یادگیری را برای دستیابی به تعمیم خارج از توزیع طراحی می کند.
3) **بهینه سازی برای تعمیم خارج از توزیع:** شامل بهینه سازی مقاوم توزیعی[5] و بهینه سازی مبتنی بر عدم تغییر[6] که به صورت مستقیم تعمیم خارج از توزیع را فرموله می کند و بهینه

---

[1] *inductive bias*
[2] Disentangled Representation Learning
[3] *embed*
[4] *stable learning*
[5] *Distributionally Robust Optimization*
[6] *Invariance-Based Optimization*

سازی را با تضمین های نظری برای تعمیم خارج از توزیع در نظر می گیرد.

3-2- یادگیری بازنمایی بدون نظارت

در این بخش، روش هایی که بر روی یادگیری بازنمایی بدون نظارت تمرکز دارند را بررسی می‌کنیم، که می‌توان آن‌ها را عمدتاً به دو شاخه تقسیم کرد، یعنی یادگیری بازنمایی تفکیک شده[1] و یادگیری بازنمایی علّی. این روش ها از دانش قبلی انسان برای طراحی رویه و مراحل یادگیری بازنمایی استفاده می کنند، که به بازنمایی یادگرفته شده، ویژگی های خاصی می بخشد که به طور بالقوه برای تعمیم خارج از توزیع مفید هستند.

3-2-1- یادگیری بازنمایی تفکیک شده

هدف یادگیری بازنمایی تفکیک شده، یادگیری بازنمایی هایی است که در آن فاکتور های متمایز و دارای اطلاعات مفید از تغییرات در داده ها، از هم تفکیک شده باشند [5]، که به عنوان یک ویژگی از بازنمایی خوب در نظر گرفته می شود و به طور بالقوه از تعمیم خارج از توزیع بهره می برد. کارهایی که برای انجام تفکیک سازی با روش های مبتنی بر VAE[2] انجام شده اند، هم بر تفسیر پذیری و هم بر پراکندگی[3] تأکید می کنند، که پراکندگی به معنای تغییرات کوچک توزیع است که معمولاً خود را به صورت پراکنده یا محلی در فاکتور گیری تفکیک شده[4] نشان می دهند.

علیرغم موفقیت یادگیری بازنمایی تفکیک شده، لوکاتلو و همکاران [5] برخی از مفروضات رایج یادگیری بازنمایی تفکیک شده بدون نظارت را به چالش می کشد (به عنوان مثال، استقلال عوامل پنهان[5]). همچنین این پرسش را مطرح می کند که آیا تفکیک سازی می تواند عملکردهای *downstream task* را بهبود بخشد، و الهام بخش کارهای بعدی برای در نظر گرفتن *downstream task* ها، از جمله عملکرد تعمیم خارج از توزیع می باشد.

با این حال، اینکه آیا بازنمایی تفکیک شده به نفع تعمیم خارج از توزیع است، بحث برانگیز است. محققان برخی آزمایش های برون یابی کمی را انجام دادند و دریافتند که بازنمایی تفکیک شده آموخته شده به داده های دیده نشده تعمیم نمی یابد. در مجموع، مزیت بازنمایی تفکیک شده در وظایف خارج از توزیع

---

[1] disentangled
[2] Variational Autoencoders
[3] sparsity
[4] disentangled factorization
[5] latent factor

هنوز به تحقیق و بحث بیشتری نیاز دارد.

3-2-2- یادگیری بازنمایی علّی

مشابه یادگیری بازنمایی تفکیک شده ، هدف یادگیری بازنمایی علّی، یادگیری متغیر ها در گراف علّی به یک روش بدون نظارت یا نیمه نظارتی[1] است. علاوه بر این، بازنمایی علّی را می توان به عنوان هدف نهایی روش تفکیک شده در نظر گرفت، که تعریف غیررسمی بازنمایی تفکیک شده را از نظر تفسیر پذیری و پراکندگی برآورده می کند. با بازنمایی علّی یادگرفته شده، می توان به فرآیند تولید داده های پنهان[2] پی برد، که می تواند به پایداری[3] در برابر تغییرات توزیعی ناشی از مداخلات[4] کمک کند.

در سناریو های واقعی که مشاهدات به جای داده های ساختار یافته در قالب تصاویر یا جملات شکل گرفته اند، اطلاعات انتزاعی سطح بالا باید از داده های سطح پایین استخراج شود، و در مقالات موجود پیشنهاد می شود که فاکتورسازی علّی[5] را از طریق تفکیک سازی بدست آورد.

3-3- یادگیری مدل با نظارت برای تعمیم خارج از توزیع

جدای از یادگیری بازنمایی ها صرفاً به صورت بدون نظارت، شاخه هایی از کارهای انجام شده وجود دارد که اطلاعات نظارت شده را برای طراحی معماری های مدل مختلف و استراتژی های یادگیری مربوطه در نظر می گیرد. در این بخش، روش هایی را بررسی می کنیم که بر یادگیری مدل انتها به انتها[6] برای دستیابی به توانایی تعمیم خارج از توزیع تمرکز می کنند، از جمله تعمیم دامنه، یادگیری علّی، یادگیری ثابت[7] و یادگیری پایدار[8].

---

[1] *semi-supervised*
[2] *the latent data generation process*
[3] *resist*
[4] *interventions*
[5] *causal factorization*
[6] *end-to-end*
[7] *invariant learning*
[8] *stable learning*

3-3-1- تعمیم دامنه

هدف تعمیم دامنه[1]، یادگیری مدل از طریق ترکیب داده ها از چندین دامنه مختلف است که به خوبی در دامنه هدف دیده نشده تعمیم یابد که عمدتاً بر مسائل طبقه بندی مربوط به بینایی کامپیوتر تمرکز دارد، زیرا پیش بینی ها تحت تاثیر بر اختلال[2] در تصاویر (مانند سبک، پس زمینه، نور، چرخش و غیره) هستند. روش های تعمیم دامنه را به سه شاخه، یعنی یادگیری بازنمایی، استراتژی آموزش و تقویت داده ها تقسیم می کنیم که به اختصار آن ها را معرفی خواهم کرد.

3-3-1-1 یادگیری بازنمایی برای تعمیم دامنه

یادگیری بازنمایی یک بخش مهم در تعمیم دامنه است. مقاله ی [6] از نظر تئوری یا تجربی ثابت می کند که اگر بازنمایی ها در هنگام تغییر دامنه ثابت بمانند، بازنمایی ها قابل انتقال[3] و مقاوم[4] به دامنه های مختلف هستند. روش هایی که تلاش می کنند بازنمایی های ثابت را در میان دامنه های مختلف یاد بگیرند، عمدتاً به سه دسته تقسیم می شوند:

1. یادگیری متخاصم دامنه[5]
2. هم ترازی دامنه[6]
3. روش های مبتنی بر هسته.

یادگیری متخاصم دامنه: محققان یک شبکه عصبی متخاصم دامنه[7] را برای تطبیق دامنه[8] پیشنهاد کرده اند. *DANN* با بهینه سازی توام ویژگی های پایه، بازنمایی هایی را می آموزد که با تغییر دامنه، متمایز و ثابت هستند. همچنین *DANN* دارای یک پیش بینی کننده برچسب است که برچسب های کلاس را پیش بینی می کند که هم در مرحله آموزش و هم در مرحله استنتاج[9] استفاده می شود. همینطور *DANN* دارای یک

---

[1] DG; Domain Generalization
[2] disturbance
[3] transferable
[4] robust
[5] domain adversarial learning
[6] domain alignment
[7] DANN
[8] domain adaptation
[9] inference phase

طبقه بندی دامنه است که بین دامنه های مبدا و هـدف در طـول آمـوزش تمـایز قائـل مـی شـود. در واقـع بازنمایی ها به گونه ای آموزش داده می شوند که طبقه بندی کننده دامنه را گیج کنند تا ویژگی های ثابـت دامنه یاد گرفته شود. لی و همکاران [7] این ایده را برای تعمیم دامنه استفاده کرده اند. و گنگ و همکـاران [7] آموزش متخاصم را بیشتر به فضای چندگانه[1] گسترش داده اند. لی و همکاران [7] پیشنهاد به یادگیری شبکه های متخاصم کلاس خاص[2] از طریق یک شبکه متخاصم ثابت شرطی[3] می کنند.

هم ترازی دامنه[4]: برخی از مقاله ها پیشنهاد به یادگیری بازنمایی های ثابت دامنـه[5] از طریـق هـم تـرازی[6] ویژگی ها می کنند. معطیان و همکاران [7] پیشنهاد به یادگیری هم ترازی معنایی[7] بین دامنه های مختلف با به حداقل رساندن فاصله بین نمونه ها از دامنه های مختلف اما کلاس یکسان و به حداکثر رسـاندن فاصـله بین نمونه ها از دامنه ها و کلاس های مختلف می کنند. برخی از مقالـه هـا بـا بـه حـداقل رسـاندن فاصـله حداکثر میانگین اختلاف[8] ، فاصله *Wasserstein* و همبستگی مرتبـه دوم، واگرایـی توزیـع ویژگـی هـا را برای تطبیق دامنه یا تعمیم دامنه به حداقل می رسانند.

روش های مبتنی بر هسته: روش های مبتنی بر هسته نیز به طور گسترده در تعمـیم دامنـه اسـتفاده مـی شوند. بلانچارد و همکاران [7] ابتدا مسئله تعمیم دامنه را از طریق روش های مبتنی بر هسـته بررسـی مـی کنند و پیشنهاد به یادگیری یک هسته ثابت دامنه[9] با داده های آموزشی می کنند. موانـدت و همکـاران [7] یک روش مبتنی بر هسته کلاسیک برای تعمیم دامنه به نـام تجزیـه و تحلیـل مؤلفـه هـای ثابـت دامنـه[10] پیشنهاد می کنند که واریانس توزیعی بین نمونه ها از دامنـه هـای مبـدا را بـه حـداقل مـی رسـاند. گـان و همکاران [7] *DICA* را با تنظیم صفت[11] گسترش می دهند.

---

[1] manifold space
[2] class-specific adversarial networks
[3] CIAN- conditional invariant adversarial network
[4] domain alignment
[5] domain invariant representations
[6] alignment
[7] semantic alignment
[8] MMD-maximum mean discrepancy
[9] domain-invariant kernel
[10] DICA- Domain-Invariant Component Analysis
[11] attribute regularization

2-1-3-3 استراتژی یادگیری

با توجه به اینکه تعمیم دامنه بیشتر بر مسائل طبقه بندی مربوط بـه بینـایی کـامپیوتری تمرکـز مـی کنـد، برخی از مقالات، استراتژی های یادگیری را برای افزایش توانایی تعمیم مدل های عمیق بـر روی داده هـای تصویری به کار می گیرند که می توان آنها را به چهار دسته تقسیم کرد: فرا یـادگیری[1]، یـادگیری گروهـی[2]، تعمیم دامنه بدون نظارت/نیمه-نظارتی[3] و مابقی روش ها.

فرا یادگیری:

فرا یادگیری از طریق یک الگوی یادگیری جایگزین، تجربه و دانش را در طول دوره های یادگیری متعدد بـه دست می آورد. فین و همکاران [7] فرایادگیری مدل-آگنوستیک[4] را برای انطباق دامنه[5] پیشنهاد می کنند، که مفهوم «قسمت ها[6]» را در مرحله آموزش معرفی می کند و تا حد زیادی بر تحقیقات فرا یـادگیری بـرای تعمیم دامنه تأثیر می گذارد. ایده اصلی فرا یادگیری برای تعمیم دامنه این است که دامنه هـای مبـدا را بـه فرا آموزش[7] و فرا آزمون[8] تقسیم کنیم، جایی که تابع زیان فرا آموزش[9] و فرا آزمون[10] به طور همزمان بهینه می شوند.

یادگیری گروهی:

به طور معمول، روش های مبتنی بر یادگیری گروهی، مجموعه ای از چندین مدل خاص را برای دامنه هـای مبدا مختلف یاد می گیرند تا توانایی تعمیم را بهبود بخشند. برخـی از مقالـه هـا، زیـر شـبکه هـای خـاص-دامنه[11] را برای دامنه های مبدا مختلف، همراه با یک طبقه بندی کننده واحد یا چنـدین سـر طبقـه بنـدی کننده خاص-دامنه در نظر می گیرند. برخی دیگر از نرمال سازی دسته ای خاص-دامنـه بـرای دامنـه هـای مختلف استفاده می کنند تا نرمال سازی بهتری را بیاموزند.

---

[1] meta learning
[2] ensemble learning
[3] unsupervised/semi-supervised DG
[4] MAML; model-agnostic meta-learning
[5] Domain Adaptation
[6] episodes
[7] meta-train
[8] meta-test
[9] meta-train
[10] meta-test
[11] domain-specific

تعمیم دامنه بدون نظارت/نیمه-نظارتی:

اخیراً، با الهام از انطباق[1] دامنه بدون نظارت و نیمه نظارت، برخی از مقاله ها پیشنهاد می کنند که توانایی تعمیم مدل را با یادگیری بدون نظارت یا نیمه نظارت افزایش دهند. ژانگ و همکاران [7] آموزش بدون نظارت نامربوط دامنه[2] را برای مقابله با تغییرات توزیع بین دامنه های مبدا و دامنه های هدف پیشنهاد می کنند. آنها نمونه های منفی معتبر را برای هر نمونه صف داده شده با توجه به شباهت بین دامنه های مختلف انتخاب می کنند تا بازنمایی های نامربوط دامنه[3] را یاد بگیرند.

مابقی روش ها:

کارلوچی و همکاران با الهام از روش یادگیری خود نظارتی[4] [7] یک مسئله پازل خودآموز را با مسئله طبقه بندی ترکیب می کنند تا بازنمایی های مقاوم[5] را بیاموزند. ریو و همکاران [7] نمونه برداری از نمونه های مثبت و منفی را از طریق جنگل تصادفی[6] پیشنهاد می کنند. لی و همکاران [7] به طور متناوب لایه های کانولوشن و طبقه بندی کننده را آموزش می دهند. هوانگ و همکاران [7] یک الگوریتم حذف تصادفی خود چالش برانگیز[7] را برای جلوگیری از بیش برازش مدل به دامنه های مبدا معرفی می کنند.

3-1-3-3 تقویت داده

تقویت داده ها روشی رایج و موثر در یادگیری عمیق به ویژه در بینایی کامپیوتری است. توانایی تعمیم مدل های عمیق تا حد زیادی به ناهمگونی داده های موجود بستگی دارد. بنابراین ناهمگونی ناشی از افزایش داده ها می تواند از بیش برازش جلوگیری کند و توانایی تعمیم را بهبود بخشد. روش های افزایش داده ها برای تعمیم دامنه را می توان به تقویت مبتنی بر تصادفی بودن[8] ، تقویت مبتنی بر گرادیان[9] و تقویت مبتنی بر

---

[1] adaptation
[2] DIUL; Domain-Irrelevant Unsupervised Learning
[3] domain-irrelevant representations
[4] self-supervised
[5] robust
[6] random forest
[7] self-challenging dropout algorithm
[8] randomization based augmentation
[9] gradient-based augmentation

تولید[1] تقسیم کرد[8] ,[9] ,[10].

2-3-3- یادگیری ثابت و علّی[2]

در مقایسه با تعمیم دامنه که به طور معمول مسائل بینایی را هدف قرار می دهد ، یادگیری علّی و یادگیری ثابت ، از موضوع استنتاج علّی ناشی می شود و مسئله تعمیم خارج از توزیع را به روشی اصولی تر بررسی می کند، که هدف آن کشف متغیرهای علّی برای پیش بینی است و اخیراً کاربردی تر شده است.
برای روش های یادگیری علّی، اغلب فرض می شود که ناهمگونی داده و رابطه علّی در داخل داده ها وجود دارد. به طور خاص، یادگیری علّی فرض می کند که فرد به داده ها از چندین محیط دسترسی دارد.

1-2-3-3- روشهای مبتنی بر استنتاج علّی

ابتدا روش های مربوط به استنتاج علّی را بررسی می کنیم که سعی می کند متغیرهای علّی را از داده های ناهمگن به دست آورد.
به وضوح مشخص است که یک استاندارد مناسب برای شناسایی اثر علّی یک متغیر، انجام آزمایش های تصادفی سازی شده مانند آزمایش A/B است، اما آزمایش های کاملاً تصادفی معمولاً گران هستند و حتی در کاربردهای واقعی غیرممکن هستند. از آنجایی که استنتاج علّی[3] یا یادگیری ساختاری علّی[4] بسیار بلندپروازانه است، این تکنیک های استنتاج باید به عنوان "حقیقت پایه[5]" در نظر گرفته شوند، اما لزوماً نمی توانند در عمل محقق شوند. بنابراین، طراحی چنین تکنیک هایی که از نظر "توضیح علّی" نسبت به رگرسیون استاندارد[6] یا چارچوب طبقه بندی[7] غنی تر هستند و همچنین می توانند نوعی تغییر ناپذیری را در بین محیط ها به دست آورند، عملی تر است. به دنبال چنین شهودی، رشته‌ای از روش‌ها [11] با بهره بردن از ناهمگونی درون داده‌ها (به عنوان مثال، محیط های متعدد) توسعه یافته اند.

---

[1] generation based augmentation
[2] Causal & Invariant Learning
[3] causal inference
[4] causal structural learning
[5] ground truth
[6] standard regression
[7] classification framework

پیترز و همکاران [7] در ابتدا سعی کردند که این واقعیت را بررسی کنند که «عدم تغییر[1]» تا حدی می تواند ساختار علّی را تحت شرایط لازم استنتاج کند و پیش بینی علّی ثابت[2] را پیشنهاد کردند. به طور خاص، آنها از این واقعیت استفاده می کنند که هنگام در نظر گرفتن همه علل مستقیم یک متغیر هدف، توزیع شرطی هدف به شرط علل مستقیم، هنگام تداخل با همه متغیرهای دیگر در مدل به جز خود هدف، تغییر نخواهد کرد.

اگرچه ICP اولین تلاش برای ارتباط دادن روش ثابت[3] به علّیت است، اما محدودیت های متعددی دارد. ابتدایی ترین آنها، الزامات سختگیرانه برای ناهمگنی[4] است زیرا قدرت ICP به شدت به کیفیت محیط های موجود (یا اغتشاشات[5]) بستگی دارد. اگر زیرجمعیت های آشفته موجود[6] کافی نباشد یا حتی یک محیط واحد وجود نداشته باشد، کارایی ICP از بین خواهد رفت.

2-2-3-3- یادگیری ثابت

طبق روش های مبتنی بر استنتاج علّی، روش های یادگیری ثابت، که با حداقل سازی ریسک ثابت[7] مشخص می شوند، مکانیسم های علّی پنهان[8] را هدف قرار می دهند و ICP را به حالت کاربردی تر و عمومی تر گسترش می دهند. در حالی که روش های پیش بینی علّی، سطح متغیر خام[9] را در نظر می گیرند.

بر اساس IRM که با بازنمایی $\Phi(X)$ به گونه ای است که $E[Y|\Phi(X)]$ ثابت می ماند، کارهای بعدی تغییراتی را در این راستا به صورت قاعده مندتر از فرض عدم تغییر IRM[10] پیشنهاد کرده اند که منجر به جایگزین های مشابه می شود. آهوجا و همکاران [7] از تئوری بازی ها در این مسئله استفاده کرده اند و طبقه بندی کننده خطی در IRM را با مجموعه ای از طبقه بندی کننده ها از محیط های مختلف جایگزین می کنند. جین و همکاران [7] تنظیم کننده IRM[11] را با تاثر پیش بینی کننده[1] جایگزین می کنند و

---

[1] invariance
[2] ICP- Invariant Causal Prediction
[3] invariance
[4] heterogeneity
[5] perturbations
[6] available perturbed subpopulations
[7] IRM- invariant risk minimization
[8] latent causal mechanisms
[9] raw variable level
[10] IRM's invariance assumption
[11] regularizer of IRM

محدودیت های سختگیرانه تری را بر $\Phi(X)$ تحمیل می کنند. کروگ و همکاران [7] پیشنهاد می کنند که واریانس ریسک ها در بین محیط ها جریمه شوند، در حالی که Xie و همکاران. [7] تقریباً همان هدف را مطرح می کنند اما جریمه اصلی را با جذر واریانس جایگزین می کنند. ماهاجان و همکاران [7] یک قاعده ساز کنتراست[2] را معرفی می کند که بازنمایی اشیاء مشابه را در بین محیط ها مطابقت می دهد. و کریگر و همکاران مشکل از دست رفتن برچسب های محیطی IRM را هدف قرار می دهند و استنتاج محیطی برای یادگیری ثابت[3] را برای یادگیری محیط هایی که جریمه IRM را به حداکثر می رسانند، افزایش می دهند. کل الگوریتم دو مرحله ای است که ابتدا محیط ها را طبق یک مدل مرجع مغرضانه[4] تولید می کند و سپس یادگیری ثابت را با محیط های یادگیری انجام می دهد.

در حالی که نتایج در IRM امیدوارکننده به نظر می رسد، روزنفلد و همکاران [7] به برخی از مشکلات IRM در مسائل طبقه بندی اشاره می کنند. در حالت خطی، آن ها شرایط ساده‌ای را ارائه می کنند که در آن راه‌حل بهینه اغلب نمی تواند پیش بینی کننده ثابت بهینه[5] را بازیابی کند. به طور خاص، روزنفلد و همکاران [7] نشان می دهند که یک راه حل امکان پذیر وجود دارد که فقط از ویژگی های محیطی استفاده می کند و در عین حال عملکرد بهتری نسبت به پیش بینی کننده ثابت بهینه[6] در همه محیط ها دارد. در یک حالت غیر خطی، آنها نشان می دهند که IRM می تواند به طور فاجعه باری شکست بخورد، مگر اینکه داده های آزمون به اندازه کافی شبیه توزیع آموزش باشند.

3-3-3- یادگیری پایدار[7]

در مقایسه با تعمیم دامنه و یادگیری علّی، یادگیری پایدار راه دیگری را برای ترکیب استنتاج علّی با یادگیری ماشین ایجاد می کند، که به طور قابل توجهی الزامات محیط های متعدد را تسهیل می کند.

مسئله 2 (تنظیمات یادگیری پایدار): با توجه به داده های آموزشی $D^e = (X^e, Y^e)$ از یک محیط $e \in supp(\varepsilon_{all})$، هدف یادگیری پایدار، یادگیری یک مدل پیش بینی با عملکرد یکنواخت مناسب در

---

[1] predictive regret
[2] contrastive regularizer
[3] EIIL- Environment Inference for Invariant Learning
[4] a biased reference model
[5] optimal invariant predictor
[6] the optimal invariant predictor
[7] Stable Learning

هر محیط ممکن در $supp(\varepsilon_{all})$ است.

برای حل چنین مشکل دشواری، با کمک استراتژی های متعادل کننده متغیر[1]، شن و همکاران [12] پیشنهاد می کنند که همه متغیرها را به عنوان رفتار[2] در نظر بگیریم و مجموعه ای از وزن های نمونه سراسری را یاد بگیریم که می تواند سوگیری مخدوش‌کننده[3] را برای همه رفتار های بالقوه[4] از توزیع داده ها حذف کند. آنها یک تابع زیان متعادل سراسری[5] را به دست می آورند که می تواند به راحتی به مسائل استاندارد یادگیری ماشین به عنوان تنظیم کننده[6] متصل شود، همانطور که در معادله 3 نشان داده شده است:

$$\sum_{j=1}^{p}\left\|\frac{X_{-j}^{T}\cdot\left(W\odot I_j\right)}{W^{T}\cdot I_j}-\frac{X_{-j}^{T}\cdot\left(W\odot\left(1-I_j\right)\right)}{W^{T}\cdot\left(1-I_j\right)}\right\|_2^2, \tag{3}$$

که W نشان دهنده ی وزن های نمونه و $\left\|\frac{X_{-j}^{T}\cdot(W\odot I_j)}{W^{T}\cdot I_j}-\frac{X_{-j}^{T}\cdot\left(W\odot(1-I_j)\right)}{W^{T}\cdot(1-I_j)}\right\|_2^2$ نشان دهنده زیان تعادل مخدوش کننده[7] هنگام تنظیم ویژگی j به عنوان متغیر رفتار[8] است، و X_{-j} ویژگی های باقی مانده (یعنی مخدوش کننده ها) به جز ستون j ام است. I_j به معنای ستون jام I است و I_{ij} به وضعیت رفتار واحد i هنگام تنظیم ویژگی jبه عنوان متغیر رفتار اشاره دارد. با به حداقل رساندن زیان تعادل سراسری[9]، سوگیری مخدوش کننده[10] را می توان در مقیاس سراسری حذف کرد.

علاوه بر این، کوانگ و همکاران. [7] بازنمایی ویژگی بدون نظارت را در مرحله تعادل سراسری[11] با رمزگذار های خودکار[12] [7] ترکیب می کنند و تنظیم کننده اصلی[13] را به نسخه "عمیق" به عنوان معادله 4 تغییر

---

[1] variable balancing strategies
[2] treatment
[3] confounding bias
[4] potential treatments
[5] global balancing loss
[6] regularizer
[7] loss of confounder balancing
[8] treatment variable
[9] global balancing loss
[10] the confounding bias
[11] the global balancing stage
[12] auto-encoders
[13] original regularizer

می دهند:

$$\sum_{j=1}^{p}\left\|\frac{\emptyset\left(X_{\cdot,-j}\right)^{T}\cdot\left(W\odot X_{\cdot j}\right)}{W^{T}\cdot X_{\cdot,j}}-\frac{\emptyset\left(X_{\cdot,-j}\right)^{T}\cdot\left(W\odot(1-X_{\cdot,j})\right)}{W^{T}\cdot(1-X_{\cdot,j})}\right\|_{2}^{2}, \qquad (4)$$

روش های فوق به ویژگی های باینری محدود می شوند، زیرا مسائل مطرح شده در مورد استنتاج علّی در محدوده رفتار باینری هستند. هنگامی که متغیر رفتار[1] قطعی یا پیوسته است، روش های متعادل سازی قدیمی[2] دیگر امکان پذیر نیستند، زیرا سطح رفتار[3] می تواند بی نهایت بزرگ باشد. برای کاهش چنین محدودیت هایی و پرداختن به مشکل رفتار پیوسته[4]، کوانگ و همکاران. [7] پیشنهاد به یادگیری مجموعه ای از وزن های نمونه می کنند تا توزیع وزنی رفتار و مخدوش کننده[5] بتواند شرایط استقلال را برآورده کند، که با این واقعیت مطابقت دارد که اگر رفتار و مخدوش کننده[7] مستقل باشند، اثر رفتار را می توان به طور دقیق تخمین زد.

3-4- بهینه سازی برای تعمیم خارج از توزیع

برای پرداختن به مسئله تعمیم خارج از توزیع، جدا از یادگیری بازنمایی بدون نظارت و مدل های یادگیری انتها به انتها، روش های بهینه سازی با ضمانت های نظری، اخیراً به صورت گسترده مورد توجه قرار گرفته اند که هم به صورت آگنوستیک مدل[6] و هم آگنوستیک ساختار داده[7] می باشند. در این بخش ابتدا هدف این روش های بهینه سازی خارج از توزیع را معرفی می کنیم و سپس روش های بهینه سازی از جمله بهینه سازی مقاوم توزیعی[8] و بهینه سازی مبتنی بر ثابت[9] را بررسی می کنیم.

به منظور پرداختن به مسئله تعمیم خارج از توزیع از منظر بهینه سازی، مسئله تعمیم خارج از توزیع برای کنترل خطای پیش بینی بدترین حالت در بین ε_{all} فرموله شده است که به شکل زیر می باشد:

---

[1] treatment variable
[2] traditional balancing methods
[3] treatment level
[4] the continuous treatment problem
[5] treatment and confounder
[6] model agnostic
[7] data structure agnostic
[8] Distributionally Robust Optimization
[9] Invariant-Based Optimization

$$\underset{f}{argmin} \underset{e\in supp(\varepsilon_{all})}{max} \mathcal{L}(f|e) \quad (5)$$

که در آن ε_{all} متغیر تصادفی در همه محیط های ممکن است، و برای همه $e \in supp(\varepsilon_{all})$، توزیع داده ها و برچسب $P^e(X,Y)$ می تواند کاملاً متفاوت از توزیع آموزشی $P_{tr}(X,Y)$ باشد. $\mathcal{L}(f|e) = \mathbb{E}[l(f(X),Y)|e] = \mathbb{E}^e[l(f(X^e),Y^e)]$ ریسک پیش بینی کننده f در محیط e و $l(.,.): y \times y \to \mathbb{R}^+$ تابع زیان است. به طور شهودی، هدف روش های بهینه سازی تضمین عملکرد بدترین حالت تحت تغییرات توزیعی است.

3-4-1- بهینه سازی مقاوم توزیعی

بهینه سازی مقاوم توزیعی[1]، مستقیماً مسئله تعمیم خارج از توزیع را با بهینه سازی خطای بدترین حالت در یک مجموعه توزیع عدم قطعیت[2] حل می کند تا از مدل در برابر تغییرات توزیعی بالقوه در مجموعه عدم قطعیت محافظت شود. اغلب توسط شرایط لحظه ای یا پشتیبانی[3] ، واگرایی-f [4] و فاصله ***Wasserstein***، محدود می شود. تابع هدف روش های DRO را می توان به صورت زیر خلاصه کرد:

$$\underset{f}{argmin} \underset{Q\in\mathcal{P}(P_{tr})}{sup} \mathbb{E}_{X,Y\sim Q}[l(f(X),Y)] \quad (6)$$

که در آن $\mathcal{P}(P_{tr})$ مجموعه توزیعی است که در پیرامون توزیع آموزشی P_{tr} قرار دارد و $l(.,.): y \times y \to \mathbb{R}^+$ تابع زیان است. روش های مختلف DRO انواع مختلفی از محدودیت ها را برای فرمول بندی مجموعه توزیع $\mathcal{P}(P_{tr})$ و الگوریتم بهینه سازی مختلف اتخاذ می کنند.

3-4-2- بهینه سازی مبتنی بر عدم تغییر[5]

برخلاف روش های DRO که مستقیماً برای بدترین حالت بدون هیچ فرضی اضافی بهینه سازی را انجام می دهند، روش های بهینه سازی مبتنی بر عدم تغییر خاصیت عدم تغییر را در داده ها فرض می کنند و از

[1] DRO-Distributionally Robust Optimization
[2] uncertainty distribution set
[3] moment or support conditions
[4] f-divergence
[5] Invariance-Based Optimization

محیط های متعدد برای یافتن چنین تغییر ناپذیری برای تعمیم تحت تغییرات توزیعی استفاده می کنند. چانگ و همکاران [7] و کویاما و همکاران [7] بازنمایی عدم تغییر مورد نظر را با استفاده از تئوری اطلاعات، فرموله کرده اند و پیشنهاد به پیدا کردن پیش بینی کننده تغییر ناپذیر حداکثری[1] در محیط های یادگیری برای کنترل بدترین خطا در معادله 5 می کنند.

3-5- مجموعه داده ها و معیارهای ارزیابی

برای توسعه و کمک به تحقیقاتی که در راستای تعمیم خارج از توزیع انجام می شوند، ارزیابی عملکرد تعمیم خارج از توزیع الگوریتم های مختلف به صورت منطقی و دقیق، از اهمیت زیادی برخوردار است. ارزیابی یک الگوریتم معمولاً از دو بخش مجزا به اسم مجموعه داده ها و معیارهای ارزیابی تشکیل شده است. در این بخش، مجموعه داده ها و معیارهای ارزیابی را که معمولاً به عنوان معیار خارج از توزیع[2] در تحقیقات و مقالات موجود استفاده می شوند را ارائه می کنیم.

3-5-1- مجموعه داده ها[3]

مجموعه داده ها را می توان بر اساس معیارهای مختلفی طبقه بندی کرد، برای مثال، داده های ساختگی و داده های دنیای واقعی، داده های ویژگی خام و داده های پیچیده. زمینه های تحقیقاتی مختلف از انواع مختلف مجموعه داده ها استفاده می کنند، برای مثال، یادگیری آماری ای که به صورت مرسوم استفاده می شود اغلب از داده های ساختگی[4] استفاده می کند. بینایی کامپیوتر اغلب از داده های تصویری در دنیای واقعی استفاده می کند. در مورد مسئله تعمیم خارج از توزیع، متفاوت از وظایف یادگیری ماشین سنتی که مبتنی بر فرض $i.i.d$ هستند، لازم است که تغییرات توزیعی را برای شبیه سازی توزیع داده های آزمون که ناشناخته هستند به کار بگیریم یا ایجاد کنیم، تا آزمایش شود که آیا یک الگوریتم می تواند به توزیع های دیده نشده تعمیم یابد یا خیر. بنابراین، مطابق با کارهای اخیر در مورد تعمیم خارج از توزیع ، استفاده از داده های مصنوعی و ساده و داده های دنیای واقعی و پیچیده برای تأیید اثربخشی روش تعمیم خارج از توزیع ضروری می باشد.

---

[1] MIP- Maximal Invariant Predictor
[2] OOD benchmarks
[3] Datasets
[4] synthetic data

1-1-5-3- داده های ساختگی

داده های ساختگی[1] ابزار مهمی برای شبیه سازی تغییرات توزیعی قابل توضیح و کنترل هستند. همانطور که اوبین و همکاران [13] دریافتند که روش های تعمیم خارج از توزیع اخیر در برخی از مسائل خطی ساده کم بُعد ضعیف عمل می کنند، لازم است روش های تعمیم خارج از توزیع را روی چنین داده های ساده اما چالش برانگیزی آزمایش کنیم، که می تواند نشان دهد آیا و تا چه حد یک الگوریتم می تواند در برابر نوع خاصی از تغییرات توزیعی مقاوم باشد[2].

به طور کلی، سه مکانیسم برای شبیه سازی تغییرات توزیعی در محیط ها وجود دارد، به نام های سوگیری انتخاب[3]، سوگیری مخدوش‌کننده[4] و اثر ضد علّی[5]، که با آن ها می توان انواع خاصی از تغییرات توزیعی را به درجات مختلف شبیه سازی کرد و به وضوح اثر واقعی الگوریتم را توجیه کرد. در این مکانیسم ها، *covariate* های X به دو گروه $X = [S, V]^T$، مربوط به بخش های ثابت و متغیر داخل داده ها تقسیم می شوند. و فرض بر این است که $P(Y|S)$ در سراسر محیط ها ثابت می ماند، و $P(Y|V)$ با مکانیسم های مختلف دچار تغییر می شود[6]، که تغییرات توزیعی را به همراه دارد.

الف) سوگیری انتخاب[7]

کوانگ و همکاران [7] یک مکانیسم سوگیری انتخاب را برای معرفی تغییرات توزیعی پیشنهاد می کنند. در الگوریتم آن ها، $P(Y|V)$ توسط سوگیری انتخاب دچار تغییر می شود.

ب) سوگیری مخدوش‌کننده[8]

سوگیری مخدوش کننده نیز یکی از رایج ترین ابزارهای تغییرات توزیعی است. در مقایسه با سوگیری انتخاب، در این روش، ویژگی متغیر V به هدف Y با در نظر گرفتن مخدوش کننده مشاهده نشده C[9] مربوط می شود.

---

[1] Synthetic data
[2] resist
[3] selection bias
[4] confounding bias
[5] anti-causal effect
[6] perturbed
[7] selection bias
[8] confounding bias
[9] unobserved confounder C

ج) اثر ضد علّی[1]

علاوه بر مکانیسم های فوق، آریوفسکی و همکاران [7] و لیو و همکاران [7] یک مکانیسم ضد علت را برای تغییر $P(Y|V)$ معرفی می کنند.

2-1-5-3- داده های واقعی[2]

داده های مصنوعی به راحتی می توانند تغییرات توزیعی را با درجات مختلف ایجاد کنند، همچنین می توانند یک الگوریتم را به صورت کامل بررسی کنند که آیا می تواند به توزیع های دیده نشده تعمیم یابد یا خیر. با این حال، داده های مصنوعی نسبتاً ساده هستند و تولید داده های پیچیده (مثلاً تصاویر) دشوار است. علاوه بر این، اینکه آیا یک الگوریتم می تواند مسائل تغییر توزیعی در دنیای واقعی را حل کند نیز یک معیار مهم برای ارزیابی یک روش تعمیم خارج از توزیع است. بنابراین، لازم است که مجموعه داده های دنیای واقعی را در نظر بگیریم. در اینجا، ما معروف ترین مجموعه داده های دنیای واقعی مورد استفاده در تحقیقات خارج از توزیع را بررسی می کنیم. از جمله مجموعه داده های تصویری و سایر فرم های داده (به عنوان مثال، داده های جدولی، داده های زبان). خلاصه ای از این مجموعه داده ها در جدول 1 آمده است.

Image Data Set	ImageNet-Variant [165], [166], [167]	Colored MNIST [2]	MNIST-R [169]	Waterbirds [153]	Camelyon17 [173]	VLCS [172]	PACS [170]
# Domains	-	3	6	2	5	4	4
# Categories	-	2	10	2	2	5	7
# Examples	-	-	6,000	4,800	45w	2,800	10w
Shift Type	Adversarial Policy	Color	Angle	Background	Hospital	Data Source	Style
Image Type	Mixed Type	Digits	Digits	Birds	Tissue Slides	Real Objects	Mixed Type

Image Data Set	Office-Home [171]	DomainNet [174]	iWildCam [175]	FMoW [178]	PovertyMap [179]	NICO [177]
# Domains	4	6	323	16 × 5	23 × 2	188
# Categories	65	345	182	62	Real Value	19
# Examples	15w	50w	20w	50w	2w	2.5w
Shift Type	Style	Style	Location	Time, Location	Country, Urban/Rural	Background, Attribute, Action, View and Co-occurring Object
Image Type	Mixed Type	Mixed Type	Real Animals	Satellite	Satellite	Real Objects

جدول 1:مجموعه داده های تصویری که معمولاً برای تعمیم *OOD* استفاده می شوند. نوع تغییر (*Shift type*) نشان دهنده نوع تغییرات توزیعی است و نوع مختلط (*mixed type*) در نوع تصویر (*image type*) به این معنی است که هم تصاویر واقعی و هم غیر واقعی وجود دارد.

مجموعه داده های تصویری

با توسعه سریع بینایی کامپیوتر، تعدادی از مجموعه داده های تصویری ایجاد شده است. ما آن ها را با توجه

[1] anti-causal effect
[2] Real-World Data

به میزان شبیه سازی تغییرات توزیعی به سه دسته طبقه بندی می کنیم:

1. داده های تبدیل مصنوعی[1]
2. داده های رام نشده ثابت[2]
3. داده های رام نشده قابل کنترل[3]

الف) داده های تبدیل مصنوعی[1]

اگرچه بیشتر مجموعه داده های تصویر برای تعمیم خارج از توزیع تولید نمی شوند، اما محققان آن ها را با برخی تبدیل های مصنوعی تغییر می دهند تا تغییرات توزیعی را شبیه سازی کنند. رایج ترین آن ها، شامل انواع *ImageNet* (به عنوان مثال *ImageNet-A*، *ImageNet-C*، *ImageNet-R*) است که خط مشی انتخاب داده یا تغییرات[4] خاصی را برای تولید داده های آزمایشی با تغییرات توزیعی اتخاذ می کنند. بقیه مجموعه داده ها روی *MNIST* کار کرده اند که انواع *MNIST* را به وجود آورده اند (مانند *MNIST* رنگی، *MNIST-R*)، که محیط های مختلف را با رنگ آمیزی یا چرخش تصاویر اصلی شبیه سازی می کنند. از آنجایی که این مجموعه داده ها به خوبی طراحی شده اند، آنها را برای مطالعه اولیه و تأیید اثربخشی الگوریتم های تعمیم خارج از توزیع استفاده می کنند.

ب) داده های رام نشده ثابت[5]

چند مجموعه داده برای اعتبار سنجی[6] تعمیم خارج از توزیع ساخته شده اند که عمدتا از پس زمینه ها یا محیط های دنیای واقعی استفاده می کنند. *PACS* و *Office-Home* که به طور گسترده در تعمیم دامنه استفاده می شود، از سبک تصویر (به عنوان مثال، هنری، کارتونی) برای متمایز کردن دامنه ها/توزیع ها استفاده می کنند، و *VLCS* داده ها را از چهار منبع مستقل جمع آوری می کند. علاوه بر این، *Camelyon17* حاوی اسلایدهای بافت نمونه برداری شده است که روی آنها پس-پردازش[7] انجام شده و از بیمارستان های مختلف جمع آوری شده است.*DomainNet* *PACS* را در مقیاسی بسیار بزرگتر، متشکل از حوزه ها و دسته های بیشتری گسترش می دهد.

---

[1] synthetic transformation data
[2] fixed wild data
[3] controllable wild data
[4] perturbations
[5] fixed wild data
[6] validation
[7] post-processed

ج) داده های رام نشده قابل کنترل[1]

اخیراً، مجموعه داده ای وجود دارد که روش های انعطاف پذیرتر و قابل کنترل تری را برای شبیه سازی تغییرات توزیعی امکان پذیر می سازد که با $NICO$ مشخص می شود. روش $NICO$ به طور قابل قبولی زمینه های بصری[2] را با تنظیمات مختلف از جمله پس زمینه، ویژگی[3]، نما[4] و غیره انتخاب می کند و محیط های مختلفی را تولید می کند. این روش می تواند انواع مختلفی از تغییرات توزیعی واقعی با زمینه های متنوع را شبیه سازی کند، و با اندازه نمونه یکسان در هر زمینه، می‌توان درجات مختلفی از تغییرات توزیعی را به راحتی به دست آورد. این دو ویژگی باعث می شود روش $NICO$ تنظیمات انعطاف پذیر تغییرات توزیعی را دارا باشد.

2-5-3- معیار های ارزیابی

علاوه بر مجموعه داده ها، معیارهای ارزیابی نیز برای ارزیابی مناسب الگوریتم های تعمیم خارج از توزیع مهم هستند که باید در نظر گرفته شوند. در مقایسه با مسئله ی iid که در آن تنها یک توزیع آزمون در نظر گرفته می شود، اغلب، توزیع های آزمون متعددی در مسائل تعمیم خارج از توزیع وجود دارند تا توزیع های دیده نشده را بهتر به تصویر بکشند. علاوه بر این، Ye و همکاران [14] به طور تجربی دریافتند که دقت آزمون روی یک محیط واحد، ارزیابی الگوریتم ها را در یک سناریوی خارج از توزیع به درستی انجام نمی دهد. بنابراین، برای ارزیابی مناسب الگوریتم های تعمیم خارج از توزیع ، اطلاعات آماری بیشتری را باید در مورد دقت محیط های آزمون مختلف در نظر گرفت. در اینجا، ما سه معیار ارزیابی، از جمله دقت متوسط، دقت در بدترین حالت، و انحراف معیار دقت را بررسی می کنیم. برای راحتی، دقت مدل در K محیط آزمون را به ترتیب $\{acc_1, \cdots, acc_k, \}$ فرض می کنیم.

الف) دقت متوسط

دقت متوسط $\overline{Acc}$ روی توزیع های آزمون، ساده ترین راه برای ارزیابی اثربخشی الگوریتم های خارج از توزیع است که معمولاً در مقاله های تعمیم خارج از توزیع استفاده می شود و به صورت زیر محاسبه می شود:

---

[1] controllable wild data
[2] visual contexts
[3] attribute
[4] view

$$\overline{Acc} = \frac{1}{k}\sum_{k=1}^{k} acc_k, \tag{7}$$

دقت متوسط، عملکرد کلی را در میان توزیع های آزمون اندازه‌گیری می کند، اما نمی تواند نوسانات عملکرد یک الگوریتم را توصیف کند. علاوه بر این، میانگین دقت به طور یکسان با همه توزیع های آزمون بدون در نظر گرفتن ویژگی هر یک، رفتار می کند که ممکن است از توزیع هایی که مکررا رخ می دهند فریب بخورد.

ب) دقت در بدترین حالت

این دقت به طور گسترده در مقالات DRO استفاده می شود. Acc_{worst} به عنوان دقت در بدترین حالت روی توزیع های آزمون تعریف می شود:

$$Acc_{worst} = \min_{k\in[k]} acc_k \tag{8}$$

دقت در بدترین حالت، قابلیت اطمینان یک الگوریتم را نشان می دهد.

ج) انحراف معیار[1]

انحراف معیار دقت روی توزیع های آزمون (Acc_{std})، تغییرات عملکرد در توزیع های مختلف را اندازه گیری می کند که به صورت زیر تعریف می شود:

$$Acc_{std} = \sqrt{\frac{1}{k-1}\sum_{k=1}^{k}(acc_k - \overline{Acc})^2} \tag{9}$$

این متریک حساسیت یک الگوریتم را اندازه‌گیری می کند و استحکام[2] و پایداری[3] الگوریتم را نشان می دهد

[1] STD; Standard Deviation
[2] robustness
[3] stability

که این موضوع برای یک الگوریتم تعمیم خارج از توزیع مهم است.

3-6- نتیجه‌گیری

مسئله تعمیم خارج از توزیع اخیراً بسیار مورد توجه قرار گرفته است و همانطور که می دانیم بـرای اسـتقرار الگوریتم های یادگیری ماشین حیاتی است. در این فصل، تعریف، شاخه های اصلی روش ها، و مجموعه داده ها و معیارهای ارزیابی مسئله تعمیم خارج از توزیع را بررسی کردیم. بر اساس تجزیه و تحلیل، ما بـا چنـدین چالش بالقوه روبرو شدیم که در فصل بعد آن ها را مطرح خواهیم کرد.

فصل 4:

نتيجه‌گيري و كارهاي آينده

4-1- نیاز به محیط های متعدد

اکثر روش های تعمیم خارج از توزیع به چندین محیط آموزشی نیاز دارند. با این حال، مجموعه داده های مدرن اغلب با ادغام داده ها از منابع متعدد بدون حفظ برچسب های منبع جمع آوری می شوند، که به شدت استقرار روش های تعمیم خارج از توزیع را در سناریو های واقعی محدود می کند. بنابراین، عملی تر و واقع بینانه تر است که ما فقط به یک محیط آموزشی با ناهمگونی نهفته دسترسی داشته باشیم. اخیراً، اگرچه برخی از مقالات [15] وجود دارند که سعی می کنند از ناهمگنی پنهان استفاده کنند و نیاز برای محیط های متعدد را کاهش دهند، اما نحوه جست و جو و استفاده مناسب از ناهمگونی پنهان درون داده ها برای استقرار روش های تعمیم خارج از توزیع حیاتی است. و باید به عنوان ادامه ی کار های آینده در نظر گرفته شود.

4-2- ارزیابی های منطقی

اگرچه معیارهای ارزیابی از جمله داده های آزمون و مکانیسم های انتخاب مدل، برای الگوریتم های یادگیری ماشین کلاسیک تحت فرض iid به خوبی توسعه داده شده اند اما آنها را نمی توان مستقیماً در سناریو های خارج از توزیع استفاده کرد. از آنجایی که توزیع آزمون ناشناخته و متفاوت از آموزش است، نحوه طراحی تنظیمات تجربی منطقی و واقعی همچنان یک مشکل چالش برانگیز است. علاوه بر این، مکانیسم انتخاب مدل نیز مهم است، زیرا انتخاب داده های اعتبارسنجی در سناریو های خارج از توزیع مهم می باشد. تحقیقات نشان می دهند که الگوریتم های تعمیم دامنه بدون استراتژی انتخاب مدل ناقص هستند. همچنین بیان می کنند که اثرات واقعی بسیاری از روش های تعمیم دامنه ضعیف است که نشان می دهد معیارهای ارزیابی موجود برای اعتبارسنجی الگوریتم های تعمیم خارج از توزیع ناکافی هستند. بنابراین، ضروری است که معیارهای ارزیابی معقول تری برای تعمیم خارج از توزیع ایجاد شوند.

4-3- معرفی موضوع مورد نظر برای پایان‌نامه

به نظر بنده مدل های موجود هنوز به خوبی بازنمایی های مناسبی برای تعمیم خارج از توزیع بدست نمی آوردند. بنده تصمیم دارم همچون مقاله ی CIRL که در فصل دوم به آن پرداخته شد یک SCM مناسب تر برای یادگیری بازنمایی های ثابت از طریق یادگیری علّی ارائه دهم که بتواند تعمیم مناسبی روی دامنه هدف دیده نشده داشته باشد.

مراجع

[1] J. Peters, D. Janzing, and B Schölkopf. Elements of causal inference - foundations and learning algorithms. MIT Press, Cambridge, MA, USA, 2017.

[2] H. Reichenbach. The Direction of Time. University of California Press, Berkeley, CA,, 1956.

[3] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In ICML, 2012.

[4] F. Lv, J. Liang, S. Li, B. Zang. Causality Inspired Representation Learning for Domain Generalization. CVPR, 2022.

[5] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Scholkopf, ¨ and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in international conference on machine learning. PMLR, 2019, pp. 4114–4124.

[6] I. Albuquerque, J. Monteiro, M. Darvishi, T. H. Falk, and I. Mitliagkas, "Adversarial target-invariant representation learning for domain generalization," 2020.

[7] Z. Shen, J. Liu, Y. He, X. Zhang. Towards Out-Of-Distribution Generalization: A Survey. Arxiv, 2021.

[8] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2100–2110.

[9] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Deep domainadversarial image generation for domain generalisation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 13 025–13 032.

[10] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12 556–12 565.

[11] B. Schölkopf, F. Locatello, S. Bauer. Towards Causal Representation Learning. arxiv, 2021.

[12] Z. Shen, P. Cui, K. Kuang, B. Li, and P. Chen, "Causally regularized learning with agnostic data selection bias." in ACM Multimedia, 2018, pp. 411–419.

[13] B. Aubin, A. Słowik, M. Arjovsky, L. Bottou, and D. LopezPaz, "Linear unit-tests for invariance discovery," arXiv preprint arXiv:2102.10867, 2021.

[14] H. Ye, C. Xie, Y. Liu, and Z. Li, "Out-of-distribution generalization analysis via influence function," arXiv preprint arXiv:2101.08521, 2021.

[15] E. Creager, J.-H. Jacobsen, and R. Zemel, "Environment inference for invariant learning," in International Conference on Machine Learning. PMLR, 2021, pp. 2189–2200.

واژه نامه

واژه نامه فارسی به انگلیسی

تغییرات ..... Interventions
عوامل مخدوش کننده ..... Confounders
یادگیری پایدار ..... Stable Learning
سوگیری انتخاب ..... Selection bias
فرا-یادگیری ..... Meta-Learning
یادگیری بازنمایی ثابت ..... Invariant Representation Learning
تقویت دامنه ..... Domain Augmentation
دارای پارامتر ..... parametric
برچسب واقعی ..... ground-truth
جریان ..... stream
تغییر ..... *shift*
تغییر متغیر ..... covariate shift
مدل علّی ساختاری ..... *SCM; structural causal model*
دسته ..... *category*
عوامل علّی ..... *causal factors*
عوامل غیر علّی ..... *non-causal factors*
مکانیسم های علّی ثابت ..... *invariant causal mechanisms*
تغییرات علّی ..... *causal intervention*
اطلاعات مربوط به هر دسته ..... *category-related information*
یادگیری بازنمایی الهام گرفته از علیت . *CIRL; Causality Inspired Representation Learning*
با اطلاعات مرتبط با دامنه‌ی دچار تغییر شده ..... *perturbed domain-related information*
ماژول ماسک متخاصم ..... *adversarial mask module*
پوشش‌دهنده ..... *masker*
به صورت حاشیه ای یا شرطی ثابت-دامنه .... *domain-invariant marginally or conditionally*
اصل علت مشترک ..... *Common Cause Principle*
تابع زیان آنتروپی متقابل ..... *cross-entropy loss*

اصل مکانیزم های علی مستقل ..... *ICM; Independent Causal Mechanisms*

فاکتورسازی علّی ..... *causal factorization*

اندازه دسته ها ..... *batch size*

ابعاد با ارزش پایین تر ..... *inferior dimensions*

ابعاد با ارزش بالاتر ..... *superior dimensions*

ماسک شده ..... *masked*

استنتاج ..... *inference*

سوگیری استقرایی ..... *inductive bias*

یادگیری بازنمایی تفکیک شده ..... Disentangled Representation Learning

جاسازی ..... *embed*

بهینه سازی مقاوم توزیعی ..... *Distributionally Robust Optimization*

بهینه سازی مبتنی بر عدم تغییر ..... *Invariance-Based Optimization*

پراکندگی ..... sparsity

عوامل پنهان ..... latent factor

نیمه نظارتی ..... *semi-supervised*

انتها به انتها ..... *end-to-end*

تعمیم دامنه ..... DG; Domain Generalization

اختلال ..... disturbance

قابل انتقال ..... transferable

مقاوم ..... robust

یادگیری متخاصم دامنه ..... domain adversarial learning

هم ترازی دامنه ..... domain alignment

شبکه عصبی متخاصم دامنه ..... DANN

تطبیق دامنه ..... domain adaptation

فضای چندگانه ..... manifold space

شبکه های متخاصم کلاس خاص ..... class-specific adversarial networks

شبکه متخاصم ثابت شرطی ..... CIAN- conditional invariant adversarial network

حداکثر میانگین اختلاف ..... MMD-maximum mean discrepancy

تجزیه و تحلیل مؤلفه های ثابت دامنه ..... DICA- Domain-Invariant Component Analysis

تنظیم صفت ..... attribute regularization

یادگیری گروهی ..... ensemble learning

فرایادگیری مدل-آگنوستیک ..... MAML; model-agnostic meta-learning

خاص-دامنه ..... domain-specific

یادگیری بدون نظارت نامربوط دامنه..... DIUL; Domain-Irrelevant Unsupervised Learning

الگوریتم حذف تصادفی خود چالش برانگیز ..... self-challenging dropout algorithm

تقویت مبتنی بر تصادفی بودن ..... randomization based augmentation

پیش بینی علّی ثابت ..... ICP- Invariant Causal Prediction

ناهمگنی ..... heterogeneity

زیرجمعیت های آشفته موجود..... available perturbed subpopulations

سطح متغیر خام..... raw variable level

تاثر پیش بینی کننده ..... predictive regret

قاعده ساز کنتراست..... contrastive regularizer

استنتاج محیطی برای یادگیری ثابت .... EIIL- Environment Inference for Invariant Learning

مدل مرجع مغرضانه ..... biased reference model

استراتژی های متعادل کننده متغیر..... variable balancing strategies

رفتار ..... treatment

رفتار های بالقوه..... potential treatments

زیان تعادل سراسری ..... global balancing loss

متغیر رفتار ..... treatment variable

مرحله تعادل سراسری..... the global balancing stage

رمزگذار های خودکار ..... auto-encoders

مشکل رفتار پیوسته..... the continuous treatment problem

مجموعه توزیع عدم قطعیت ..... uncertainty distribution set

شرایط لحظه ای یا پشتیبانی..... moment or support conditions

پیش بینی کننده تغییر ناپذیر حداکثری ..... MIP- Maximal Invariant Predictor

معیار خارج از توزیع..... OOD benchmarks

داده های ساختگی..... synthetic data

داده های تبدیل مصنوعی..... synthetic transformation data

داده های رام نشده ثابت..... fixed wild data

داده های رام نشده قابل کنترل ..... controllable wild data

زمینه های بصری..... visual contexts

Abstract

Machine learning algorithms are usually built on the *iid* assumption that the training and test data are independent and identically distributed. This assumption means that the distribution of training data and test data are the same. In the real world, due to distributional shifts, this assumption is hardly fulfilled, which greatly reduces the accuracy of these classical machine learning algorithms. On the other hand, machine learning algorithms often use statistical models to model the dependence between data and labels, which aims to learn domain-independent representations. However, statistical models are superficial descriptions of reality because they only need to model dependencies rather than intrinsic causal mechanisms. When dependence changes with the target distribution, statistical models may fail to generalize. Causality, by focusing on the representation of structured knowledge about the data generation process that allows for interventions, can help to understand and address some of the limitations of current machine learning methods. Despite the success of statistical learning, these models provide a relatively superficial description of reality that holds only when the experimental conditions are invariant. Instead, the field of causal learning seeks to model the effect of distributional shifts with a combination of data-driven learning and assumptions not previously included in the statistical description of a system.

For the problem caused by the distribution shift, where the distribution of the test data is different from the training data, the out-of-distribution generalization problem arises through which the algorithm can generalize well on the unseen test data. In this seminar, we will review the methods of the out-of-distribution generalization problem. We will also explain one of its methods called causality, which has recently received extensive attention, in more detail.

Keywords: Representation Learning, Domain Generalization, Causal Inference, Out-Of-Distribution Generalization, Invariant Learning

Iran University of Science and Technology
School of Computer Engineering

Causal Representation Learning for Out-Of-Distribution Generalization

A Seminar Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Engineering – Artificial Intelligence

By:
Hossein Rezaei

Supervisor:
Dr. Adel Rahmani

Fall 2022