ادامه ترجمه مقاله زبان تخصصی از ابتدای صفحه 21 تا انتهای صفحه 30

احسان بیرنگ

10) ورزش و سرگرمی

تجزیه و تحلیل های ورزشی طی سالهای اخیر به سرعت در حال تحول بوده و نقش مهمی در ایجاد مزیت رقابتی برای تیم یا بازیکن ایفا می کند. تیم های ورزشی حرفه ای امروزه برای تجزیه و تحلیل خود دارای بخش ها یا کارمندان ویژه هستند. از تجزیه و تحلیل ها و پیش بینی های موجود در این زمینه می توان برای ردیابی رفتار بازیکنان ، عملکرد ، گرفتن امتیاز و... استفاده کرد. DL برای این حوزه جدید است و فقط تعداد کمی از آثار DNN را در ورزش های مختلف استفاده کرده اند. یک روش DL برای ساخت یک زمین بسکتبال هوشمند پیشنهاد شده است. این سیستم از SVM برای انتخاب بهترین دوربین برای پخش در زمان واقعی از بین دوربینهای موجود در اطراف جایگاه استفاده می کند. آنها همچنین تصاویر هیجانی بسکتبال[1] را به CNN تغذیه کردند تا از ضربات عکس بگیرند و کلیپ هایی را که از آنها گلزنی نشده است ثبت کنند، از این رو گزارش دقیق امتیاز آنلاین و کلیپ های برجسته جالب را ارائه می دهند. این سیستم برای دستیابی به دقت 94.59 درصد در ضبط کلیپ های امتیاز با دقت 45 میلی ثانیه زمان پردازش برای هر فریم نشان داده شد.

در کار دیگر وانگ و همکاران ، از RNN برای طبقه بندی بازی های تهاجمی بسکتبال در مسابقات NBA استفاده شده است. نویسندگان در کلیپ های ویدیویی بازی ها از مجموعه داده SportVU[2] استفاده کردند. این مجموعه داده ویدئویی با نرخ 25 فریم در ثانیه با شناسایی شناسه منحصر به فرد بازیکنان ، مکان آنها در زمین و موقعیت توپ را در اختیار شما قرار می دهد. نشان داده شده است که مدل آنها در66 و 80 درصد موارد رده یکم و سوم را بدست آورده است. به طور مشابه ، از RNN با واحدهای LSTM در بیش از همان مجموعه داده برای پیش بینی میزان موفقیت عکس های سه نقطه استفاده کرده و دقت طبقه بندی بهتری را نسبت به دستگاه تقویت شده شیب (GBM) و مدل خطی تعمیم یافته (GLM) گزارش کرده است.

کاوتز و همکاران به شناسایي فعالیت بازیکنان در بازیهای والیبال پرداختند. برای دستیابی به این کار از داده های حسگر پوشیدنی و CNN استفاده شده است و دقت طبقه بندی 83.2 درصد برای شناسایی فعالیت های بازیکنان مشاهده شده است.

شناخت فعالیت گروهی مبحث جالب دیگری در تیم های ورزشی است. ابراهیم و همکاران با استفاده از مدل LSTM سلسله مراتبی ، این گزینه را در تیم والیبال بررسی کردند. در این کار ، یک مدل LSTM واحد برای به دست آوردن فعالیت های هر بازیکن ساخته شده است ، و یک مدل LSTM سطح بالا برای جمع آوری مدل های فردی برای شناسایی رفتار کلی تیم طراحی گردیده. یک مدل CNN برای استخراج ویژگی ها از قاب های ویدیویی و تغذیه آنها به مدل های

[1] تصویر هیجانی بسکتبال آهنگ های مکانی و زمانی بسکتبال در منطقه کانون است. منطقه کانون شامل تخته پشتی بسکتبال ، حلقه باز و سبد است.

LSTM فرد استفاده شد. در مقایسه با چندین مدل پایه ، مدل سلسله مراتبی پیشنهادی نتایج طبقه بندی بهتری به دست می آورد.

11) خرده فروشی:

با توجه به گسترش دستگاه های تلفن همراه ، خرید آنلاین بسیار افزایش یافته است. تغییر اخیر به سمت بازیابی تصویر محصول از طریق تکنیک های جستجوی بصری مشاهده شد. از CNN ها برای جستجوی بصری بازار لباس و مد استفاده شده است ، تا در فروشگاه های آنلاین اقلام مشابه و مشابه آنچه را که در یک فیلم یا خیابان دیده اید ، پیدا کنید.

علاوه بر این ، خرید برای افراد کم بینا باید راحت انجام شود. ترکیبی از فناوری های IoT ، از جمله چرخ دستی های هوشمند ، با روش های DL یکپارچه می توانند راه حلی برای این مشکل باشند. یک سیستم بصری که شامل عینک هوشمند ، دستکش و سبد خرید است برای کمک به افراد کم بینا در خرید طراحی شده است. این سیستم همچنین از CNN برای شناسایی موانع و اشیا موجود در راهروها استفاده می کند.

علاوه بر این ، پیشخوان های بازرسی در فروشگاه های خرده فروشی معمولاً تنگناهایی هستند که افراد برای پرداخت خریدهای خود صف می کشند. توسعه چرخ دستی های هوشمند می تواند خودآزمایی را در زمان واقعی امکان پذیر کند و افزایش چنین سیستم هایی با قابلیت های پیش بینی می تواند کالایی را ارائه دهد که ممکن است مشتری براساس خرید قبلی خود به آن نیاز داشته باشد.

علاوه بر این ، توصیه کردن موارد به خریداران یک برنامه محبوب IoT برای خرده فروشی هایی است که از فناوری های مختلفی مانند سیگنال های BLE یا دوربین های تصویری استفاده می کند. رویکرد دوم می تواند از طریق شناسایی وسایل فروشگاه یا اقدامات خریداران (به عنوان مثال ، رسیدن به قفسه ، عقب کشیدن از قفسه و...) و تهیه لیست موارد مرتبط برای اقدامات کشف شده انجام شود.

برای تجزیه و تحلیل علاقه مشتری به کالاها ، لیو و همکاران یک سیستم تخمین و جهت گیری مشتری را بر اساس DNN متشکل از CNN و RNN پیشنهاد دادند. اطلاعات ورودی از دوربین های مداربسته تهیه می شود. شبکه CNN برای استخراج ویژگی های تصویر استفاده می شود. ویژگی های تصویر و آخرین ویژگی جهت گیری پیش بینی شده پس از آن به RNN تغذیه می شوند تا نمایش و جهت گیری خروجی را بدست آورند.

12) زیرساخت هوشمند IoT:

محیط IoT از تعداد زیادی سنسور ، محرک ، رسانه و بسیاری از دستگاههای دیگر تشکیل شده است که M2M بزرگ و داده های ترافیک شبکه را تولید می کنند. بنابراین ، مدیریت ، نظارت و هماهنگی این دستگاه ها منوط به پردازش چنین داده های بزرگی با تکنیک های پیشرفته یادگیری ماشین برای شناسایی تنگناها ، بهبود عملکرد و همچنین تضمین کیفیت خدمات می باشد.

یکی از وظایف محبوب برای مدیریت زیرساخت ها تشخیص ناهنجاری است. به عنوان مثال ، تشخیص طیف ناهنجاری در ارتباطات بی سیم با استفاده از AE توسط فنگ و همکاران پیشنهاد شده است. در این کار ، یک مدل AE برای

تشخیص ناهنجاری که ممکن است به دلیل تغییر ناگهانی نسبت سیگنال صوتی کانال ارتباطی رخ دهد ، ایجاد گردیده. این مدل بر اساس ویژگی های مبتنی بر نمودار فرکانس زمانی سیگنال های ورودی آموزش داده شده است. نتیجه آنها نشان داد که یک AE عمیق تر عملکرد بهتری نسبت به شبکه های کم عمق معمولی دارد. لوپز-مارتین و همکاران و شون و همکاران به ترتیب از VAE شرطی و AE عمیق برای شناسایی نفوذ به شبکه استفاده کرده اند. در شرط VAE شرطی ، از برچسب های نمونه علاوه بر متغیرهای نهفته به عنوان ورودی اضافی به شبکه رمزگشایی استفاده می شود.

آثار ناچیز از ترافیک IoT ممکن است منجر به تراکم در ستون فقرات نشود. با این حال ، نیاز به دسترسی همزمان به کانال توسط تعداد زیادی دستگاه IoT می تواند در مرحله دسترسی به کانال منجر به مشاجره شود. بحث در مورد دسترسی کانال با افزایش تأخیرهای دسترسی به یک مشکل جدی تبدیل می شود. بنابراین ، توازن بار یک راه حل مناسب است که می تواند توسط مدل های DL برای پیش بینی معیارهای راهنمایی و رانندگی و پیشنهاد مسیرهای جایگزین انجام شود. کیم و همکاران از DBN ها برای انجام تعادل بار در IoT استفاده کردند. مدل DL آنها به تعداد زیادی از داده های کاربر و بارهای شبکه آموزش داده می شود. شناسایی تداخل همچنین می تواند توسط DNN ها انجام شود همانطور که اشمیت و همکاران نشان دادند ، جایی که یك سیستم شناسایی تداخل بی سیم مبتنی بر CNN ارائه شده است. احد و همکاران نظرسنجی را در مورد کاربرد شبکه های عصبی در شبکه های بی سیم ارائه دادند. آنها ادبیات مرتبط با کیفیت خدمات و کیفیت تجربه ، تعادل بار ، بهبود امنیت و... را مرور کردند.

از آنجا که شبکه های سلولی نسل 5 (G5) در حال ظهور یکی از ستون های اصلی زیرساخت IoT است ، لازم است از فناوری های برش برای پیشرفت جنبه های مختلف شبکه های سلولی از جمله مدیریت منابع رادیویی ، مدیریت تحرک و مدیریت ارائه خدمات استفاده کرد ؛ هم چنان که خود سازماندهی ، یک راه حل کارآمد و دقیق برای مشکلات پیچیده پیکربندی می باشد. به عنوان بخشی از این تلاش ها ، استفاده از داده های شبکه های سلولی با منبع جمعیت (مثلا قدرت سیگنال) می تواند به دستیابی به راه حل های قابل اعتماد کمک کند. به عنوان مثال ، از چنین داده های بزرگی می توان برای ایجاد نقشه های پوشش دقیق تر برای شبکه های سلولی برای بهبود عملکرد شبکه استفاده کرد ، همانطور که توسط برنامه موبایل OpenSignal انجام شده است.

ج) دروس آموخته شده

در این بخش ، پنج کلاس از خدمات IoT را به عنوان خدمات اساسی معرفی کردیم که می تواند در طیف گسترده ای از برنامه های IoT مورد استفاده قرار بگیرد. ما در مورد چگونگی استفاده از DL برای دستیابی به این خدمات بحث کرده ایم. علاوه بر این ، ما طیف گسترده ای از دامنه های IoT را طی کردیم تا بدانیم که چگونه آنها از DL برای ارائه یک سرویس هوشمند بهره برداری می کنند. جدول 4 آثاری را که از خدمات بنیادی در حوزه های IoT استفاده کرده اند ، نشان می دهد.

بسیاری از حوزه ها و برنامه های IoT از شناخت تصویر سود زیادی کسب کرده اند. انتظار می رود این علاقه به سرعت افزایش یابد زیرا دوربین های با وضوح بالای تعبیه شده در تلفن های هوشمند منجر به تولید آسان تر داده های تصویری و ویدئویی می شوند. استفاده از سایر کاربردهای اساسی ، به ویژه تشخیصهای فیزیولوژیکی و روانی و

همچنین بومی سازی ، در زمینه های مختلف قابل مشاهده است. با این حال ، استفاده از خدمات امنیتی و حریم خصوصی محدود به نظر می رسد. این شکاف در توسعه برنامه های کاربردی هوشمند IoT است ، جایی که فعالیت های احتمالی هکرها و مهاجمان نادیده گرفته می شود. همچنین ، تشخیص صدا با DL به طور گسترده ای در برنامه های IoT متعلق به چندین حوزه ، مانند خانه های هوشمند ، آموزش ، ITS و صنعت مورد استفاده قرار نمی گیرد. آثاری وجود دارد که از تشخیص صدا با رویکردهای یادگیری سنتی ماشین استفاده می کند. تشخیص صدا پیشرفت قابل توجهی با DL نشان داده است. یکی از دلایل کمی ظاهر این تکنیک در برنامه های IoT عدم وجود مجموعه داده های آموزشی جامع برای هر دامنه است ، زیرا برای آموزش DNN های تشخیص صدا نیاز به مجموعه داده های آموزشی بزرگ است.

خدمات بنیادی نیاز به تجزیه و تحلیل داده های سریع دارند تا در زمینه آنها کارآمد باشند. با وجود چندین کار در این راستا ، تجزیه و تحلیل سریع داده IoT مبتنی بر DL ، فضاهای زیادی را برای توسعه الگوریتم ها و معماری ها در اختیار دارد.

جدول 5 تحقیق در هر دامنه و مدل DL آنها را خلاصه می کند. شکل 18 همچنین فرکانس مدلهای مختلفی را که در کارهای تحقیقاتی مختلف مورد استفاده قرار گرفته است ، نشان می دهد. حدود 43 درصد از مقالات ، از CNN در ساختن سیستم هاي پیشنهادي خود استفاده کرده اند در حالي که از DBN کمتر از سایر مدل ها (حدود 7 درصد) استفاده مي شود. RNN و LSTM با هم ، به عنوان مدل های سری زمانی ، در 30٪ از آثار استفاده شده اند. این جدول همچنین بر تأثیر زیاد کارهای مرتبط با تشخیص تصویر در برنامه های IoT تأکید دارد. علاوه بر این ، یک سوم برنامه های IoT مربوط به داده های سری زمانی یا سریالی است که در آن استفاده از RNN ها یک روش مفید است.

1) پیچیدگی در مقابل عملکرد:

کانزیانی و همکاران ، چندین مدل DNN مدرن را مورد بررسی قرار دادند تا رابطه بین دقت ، میزان استفاده از حافظه ، شمارش عملیات ، زمان استنتاج و مصرف برق را بررسی کنند. آنها دریافتند که صحت و زمان استنتاج یک رابطه هذلولی را نشان می دهد به طوری که افزایش جزئی در دقت منجر به زمان محاسباتی طولانی می شود. آنها همچنین نشان دادند که تعداد عملیات در یک مدل شبکه ، رابطه خطی با زمان استنتاج دارد. نتایج آنها همچنین نشان داد که اعمال محدودیت انرژی ، حداکثر دقت قابل دستیابی را محدود می کند. با توجه به ردپای حافظه و اندازه دسته ای ، نتایج نشان داد که حداکثر استفاده از حافظه در طول تخصیص حافظه اولیه مدل ثابت است و سپس با اندازه دسته ای افزایش می یابد. با توجه به اینکه نورونها اصلی ترین سازه یک مدل هستند که عملیات را انجام می دهند ، تعداد عملیات متناسب با تعداد سلولهای عصبی است. بنابراین ، می توان پیچیدگی را به عنوان تعداد نورون های موجود در شبکه بیان کرد ، به گونه ای که افزایش تعداد نورون ها به طور مستقیم بر زمان اجرا تأثیر می گذارد.

با این حال ، بین صحت و تعداد لایه ها (یعنی عمق) یا تعداد نورون ها رابطه روشن وجود ندارد. گزارش هایی وجود دارد که نشان می دهد تخریب صحت بعد از افزایش تعداد لایه ها بیش از برخی از نقاط است. به عنوان مثال ، ژانگ و همکاران ادعا می کنند که تعداد لایه ها و نورون های پنهان تأثیر مستقیمی در صحت سیستم بومی سازی دارند. افزایش لایه ها در ابتدا به نتایج بهتری می انجامد ، اما در بعضی از مواقع وقتی شبکه عمیق تر شود ، نتایج تخریب می گردند. بهترین نتیجه آنها با شبکه ای از سه لایه پنهان بدست آمد. از طرف دیگر ، نشان داده شده است که عمق

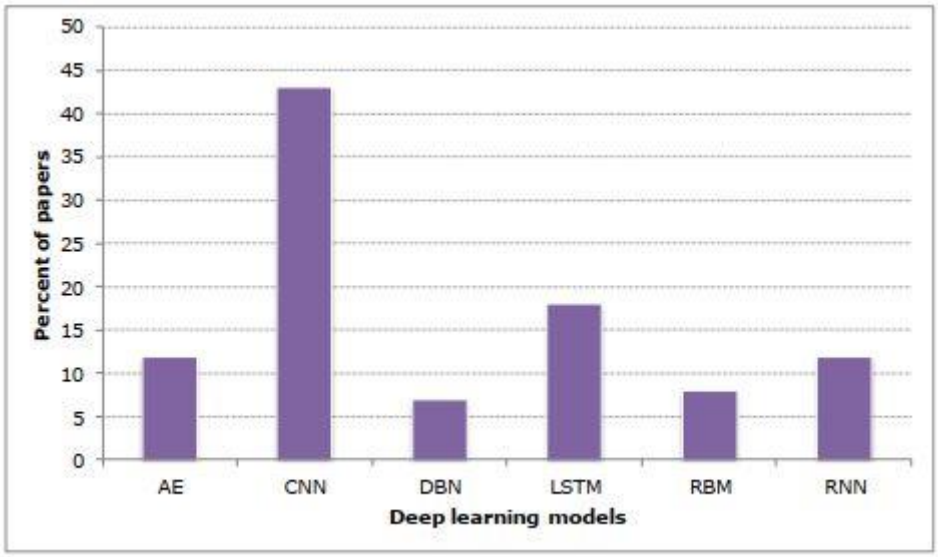
بازنمایی برای بسیاری از کارهای تشخیص تصویر ، بیشترین سود را دارد. دقت بالای کارهای مبتنی بر بینایی ، تا حدودی به دلیل معرفی شبکه های عمیق تر با تعداد بیشتری پارامتر است. بسیاری از پارامترهای بیش از حد برای بهینه سازی (به عنوان مثال ، شمارش دوره ، عملکرد از دست دادن ، عملکرد فعال سازی ، عملکرد بهینه سازی ، میزان یادگیری و...) وجود دارند که روند توسعه مدل های DL خوب و دقیق را پیچیده می کنند. جدول 6 ویژگی های مدل های DNN را در چندین برنامه ای خلاصه می کند. در جدول زمان آزمایش برای یک نمونه است ، مگر اینکه خلاف آن مشخص شده باشد.

2) مشکلات و انتقادها:

مدلهای DL نشان دادند که گامی بزرگ در جهت ایجاد سیستم های قدرتمند هوش مصنوعی برداشته اند ، اما آنها برای يك مشكل يك راه حل واحد نیستند. تکنیک های DL به عنوان جعبه های سیاه شناخته می شوند که قابلیت پیش بینی بالا اما تفسیر کم را نشان می دهد. در حالی که قابلیت پیش بینی قدرتمند از دیدگاه علمی مطلوب است ، تفسیرپذیری و قابلیت توضیح مدل ها از دیدگاه تجارت ترجیح داده می شود.

		IoT Foundational Services				
		Image Recognition	**Voice Recognition**	**Physiological & Psychological Detection**	**Localization**	**Security & Privacy**
IoT Domains	Smart Home					
	Smart City	[81], [119], [120]			[116], [117]	
	Energy					[107]
	ITS	[125], [126]				[108]
	Healthcare	[82], [128], [129], [131]	[130]	[132]		
	Agriculture	[83], [135]–[139]				
	Education	[145]		[77]		
	Industry	[78], [147]			[54]	
	Government	[84], [150], [152]				
	Sport	[154]–[156]		[157], [158]	[97]	
	Retail	[159]–[162]		[99]	[163]	

جدول 4. استفاده از خدمات مؤثر در دامنه های IOT.



شکل 18. درصد مقالات مورد بررسی که از مدلهای DL استفاده کرده اند.

چولت اظهار داشت که مشکلات مستدل ، برنامه ریزی طولانی مدت و دستکاری داده های الگوریتمی ، با مدل های یادگیری عمیق قابل حل نیست. این به دلیل ماهیت تکنیک های DL است که صرفاً چقدر داده را به آنها تغذیه می کنید ، زیرا فقط یک فضای بردار را به دیگری تبدیل می کند.

علاوه بر این ، انتقاداتی که در مورد عملکرد مدلهای DNN وجود دارد ، نشان می دهد که مدلهای سنتی ممکن است به نتایج قابل مقایسه یا حتی بهتر از مدلهای عمیق دست یابند. طبق گفته های چاتفیلد و همکاران ، ابعاد لایه های محرمانه در CNN ها می تواند بدون تأثیر منفی بر عملکرد کاهش یابد. آنها همچنین گفتند که اگر در مدلهای پیشین از روشهای تقویت داده بهره بگیریم که معمولاً برای روشهای مبتنی بر CNN استفاده می شود ، روشهای کم عمق می تواند به کارایی مشابه مدلهای عمیق CNN برسد. با و همکارانش چندین آزمایش تجربی انجام دادند که ادعا می کنند FNN های کم عمق می توانند توابع پیچیده را بیاموزند و به صحت هایی که قبلاً فقط توسط مدل های عمیق امکان پذیر بودند ، دست یابند. اریکسون و همکاران در کار خود روی سیستم های ارتباطی نوری نشان دادند که با استفاده از توالی شبه بیت های تصادفی یاتوالی های مکرر کوتاه می توانند به سیگنال-نویز بیش از حد نسبت منجر شوند.

به طور کلی ، مدل های DL به ساختار و اندازه داده ها حساس هستند. در مقایسه با مدل های کم عمق ، وقتی تعداد زیادی از داده های آموزش با طیف گسترده ای از ویژگی ها وجود داشته باشد ، آنها بهتر کار می کنند. در غیر این صورت ، معمولاً مدل های کم عمق منجر به نتایج بهتری می شوند.

5. DL در دستگاه های IOT

پیش از دوران IoT ، بیشتر تحقیقات در مورد DL برای هدف قرار دادن مدل ها و الگوریتم های خود ، به جهت کارآمد سازی هنگام کارکرد مقیاس مشکل به داده های بزرگ ، با تلاش برای استقرار مدل های کارآمد بر روی سیستم عامل های ابری بوده است. ظهور IoT هنگامی که مقیاس مشکلات به دستگاههای دارای محدودیت منابع رسیده و نیاز به تحلیل در زمان واقعی کاهش یافته است ، جهت کاملاً متفاوتی را باز کرده است.

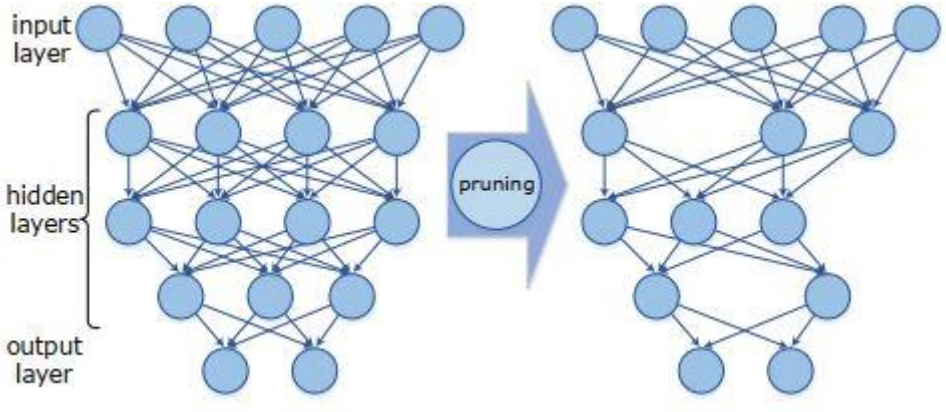
اشیاء هوشمند باید به نوعی از هوش سبک وزن پشتیبانی کنند. با توجه به نتایج موفق DL در برنامه های گفتاری و ویدیویی که از جمله خدمات اساسی و کاربردهای رایج IoT است ، تطبیق مدلها و رویکردهای آن برای استقرار در دستگاههای دارای محدودیت منابع ، به یک مطالعه بسیار مهم تبدیل شد. تاکنون روش های DL نمی تواند در مقاصد آموزشی IoT و محدود کننده منابع برای اهداف آموزشی استفاده شود زیرا مدل های DL به بخش بزرگی از منابع مانند پردازنده ها ، انرژی باتری و حافظه احتیاج دارند. در بعضی موارد ، منابع موجود حتی برای اجرای یک الگوریتم DL از پیش آموزش دیده برای کارهای استنتاجی کافی نیستند ؛ اما خوشبختانه اخیراً نشان داده شده است که بسیاری از پارامترهای ذخیره شده در DNN ها ممکن است زائد باشند. همچنین گاهی اوقات استفاده از تعداد زیادی لایه پنهان برای به دست آوردن دقت بالا ، غیر ضروری است. در نتیجه ، از بین بردن کارآمد این پارامترها و لایه ها ، پیچیدگی این DNN ها را بدون تخریب زیاد در خروجی به طور قابل توجهی کاهش می دهد و آنها را به سیستم IoT پسند تبدیل می کند. در ادامه این بخش ، ما در مورد روش ها و فن آوری ها برای دستیابی به این نتایج بحث خواهیم کرد و کاربردهای آنها را در حوزه های مختلف نشان خواهیم داد.

الف) روش ها و فن آوری ها:

مدل های DL ممکن است شامل میلیون ها یا حتی میلیارد ها پارامتر باشد که نیاز به محاسبات پیشرفته و منابع بزرگ ذخیره سازی دارند. در این بخش ، چندین رویکرد برتر از هنر را مورد بحث قرار می دهیم که مدل های DL را به دستگاه های جاسازی شده و منبع محدود IoT منتقل می کنند.

Domain	Usage of DNNs					
	AE	CNN	DBN	LSTM	RBM	RNN
Image Recognition	[150]	[81]–[83], [119], [120] [125], [126], [129], [131] [135], [138], [145] [154], [158], [160]	[147]	[156]		[155], [156]
Physiological & Psychological Detection	[104], [105]	[98], [100], [102] [103], [105], [106]	[105]	[100], [101]		[103], [104], [106]
Localization	[93], [94]	[95], [96]		[97]	[92]	
Privacy and Security		[110]	[107], [109]			
Smart home		[113]		[113]		
Smart city		[81], [119], [120]		[116]		[117]
Energy	[123]		[123]	[123] [86]	[121] [122]	[122]
ITS		[125], [126]	[108]	[124]	[79]	[79]
Healthcare		[82], [128], [129], [131]	[132]	[133]	[132]	
Agriculture		[83], [135]–[139]				
Education		[145]		[144]		[143], [144]
Industry	[146], [148], [149]	[78]	[147]			
Government	[150]	[84], [152]		[151]		
Sport		[154], [157], [158]		[156], [158]		[155]
Retail		[159]–[162]				[163]
IoT Infrastructure	[164]–[166]	[169]	[168]			

جدول V استفاده از مدل های مختلف DNN در دامنه های IOT.



شکل 19. مفهوم کلی هَرَس DNN.

1) فشرده سازی شبکه:

یکی از راه های پذیرشی DNN ها برای دستگاه های دارای منابع محدود ، استفاده از فشرده سازی شبکه است که در آن یک شبکه متراکم به یک شبکه پراکنده تبدیل می شود. این روش هنگامی که از آنها برای طبقه بندی یا انواع دیگر استنباط در دستگاه های IoT استفاده می شود ، در کاهش ذخیره سازی و نیازهای محاسباتی DNN ها کمک می کند. محدودیت اصلی این رویکرد این است که آنها به اندازه کافی برای پشتیبانی از انواع شبکه ، کلی نیستند. این فقط برای مدلهای خاص شبکه قابل استفاده است که می توانند چنین کمبودهایی را نشان دهند.

مطالعه جالب دیگر برای اتخاذ مدل های DL فشرده شده بر روی دستگاه های IoT ، موردی است که توسط لین و همکاران انجام شده است. در این مطالعه ، نویسندگان فاکتورهای مختلفی را که دستگاه های جاسازی شده ، موبایل و پوشیدنی می توانند برای اجرای الگوریتم های DL تحمل کنند ، اندازه گیری می کنند. این عوامل شامل اندازه گیری زمان اجرا ، مصرف انرژی و ردپای حافظه بود. در این تحقیق به بررسی رفتار CNN و DNN در سه سیستم عامل سخت افزاری که در IoT ، موبایل و برنامه های پوشیدنی استفاده می شود ، یعنی Snapdragon 800 که در برخی از مدل های تلفن های هوشمند و تبلت ها مورد استفاده قرار می گیرد و Intel Edison که در پوشیدنی و حساس به فرم استفاده می شود ؛ IoT و NvidiaTegra K1 نیز در تلفن های هوشمند و همچنین وسایل نقلیه دارای قابلیت فعال سازی IoT کار می کنند. Torch برای توسعه و آموزش DNN ها استفاده شده است و AlexNet مدل غالب مورد استفاده در این سیستم عامل ها بوده است. اندازه گیری آنها از مصرف انرژی نشان می دهد که تمام سیستم عامل ها ، از جمله Intel Edison (که ضعیف ترین آن است) ، قادر به اجرای مدل های فشرده شده بودند. از نظر زمان اجرای CNN ها ، نشان داده شده است که لایه های پیچشی بعد دار با کاهش ابعاد ، زمان کمتری صرف می کنند. علاوه بر این ، شناخته شده است که لایه های تغذیه پیشرو بسیار سریعتر از لایه های حلقوی در CNN ها هستند. در نتیجه ، یک رویکرد مناسب برای بهبود مدلهای CNN در دستگاههای دارای منابع محدود ، جایگزینی لایه های محرمانه با لایه های تغذیه پیشرو در هر زمان ممکن است. علاوه بر این ، انتخاب عملکرد فعال سازی به کار رفته در DNN ها می تواند تأثیر زیادی در بازده زمانی داشته باشد. به عنوان مثال ، چندین آزمایش نشان داده است که عملکردهای ReLU با زمان بیشتری کار می کنند و به دنبال آن Tanh ، و سپس Sigmoid هستند. با این حال ، کاهش کلی زمان اجرای چنین

انتخابی (حداقل 25 درصد) در مقایسه با زمان اجرای این لایه ها معنی دار نیست (کمتر از 7 درصد). از نظر استفاده از حافظه ، CNN ها به دلیل کمترین پارامترهای ذخیره شده در لایه های محرمانه نسبت به همتای خود در DNN ها ، از فضای کمتری نسبت به DNN استفاده می کنند.

همانطور که قبلاً گفته شد ، کاهش تعداد پارامترهای به کار رفته در DNN ها ، با هَرَس موارد اضافی و بی اهمیت تر ، یک رویکرد مهم دیگر برای اجرای DNN ها در دستگاه های دارای منابع محدود است. یکی از اولین کارهای این رویکرد ، Optimal Brain Damage در سال 1989 است. در هسته اصلی این روش از مشتقات مرتبه دوم پارامترها برای محاسبه اهمیت پارامترها و هَرَس پارامترهای بی اهمیت برای شبکه استفاده می شود.

Work	Application	Type of DNN	Depth	Layers Sizes	Training Time	Test Time
[79]	Transportation analysis	RNN+RBM	2	R(100)-RBM(150)	NA	354 (s), whole test set
[92]	Localization	RBM DBN	4 4	500-300-150-50 300-150-100-50	NA	NA
[93]	Localization	SdA DBN ML-ELM SDELM	4 3 5 5	26-200-200-71 26-300-71 26-300-300-1500-71 26-300-300-1500-71	451 (s) 110 (s) 14 (s) 24 (s)	NA
[94]	Localization	SdA	5	163-200-200-200-91	NA	0.25 (s)
[98]	Pose detection	CNN	12	C(55×55×96)-LRN-P- C(27×27×256)-LRN-P- C(13×13×384)- C(13×13×384)- C(13×13×256)-P- F(4096)-F(4096)-SM	NA	0.1 (s)
[100]	Human activity detection	CNN+LSTM	7	C(384)-C(20544)-C(20544)- C(20544)-L(942592)- L(33280)-SM	340 (min)	7 (s), whole test set
[101]	Human activity detection	LSTM	5	L(4)-FF(6)-L(10)-SG-SM	NA	2.7 (ms)
[107]	FDI detection	DBN	4	50-50-50-50	NA	1.01 (ms)
[109]	Malware detection	DBN	3	150-150-150	NA	NA
[120]	Parking space	CNN	8	C(64×11×11)-C(256×5×5)- C(256×3×3)-C(256×3×3)- C(256×3×3)-F(4096)-F(2)-SM	NA	0.22 (s)
[126]	Traffic sign detection	CNN	6	C(36×36×20)-P- C(14×14×50)-P- FC(250)-SM	NA	29.6 (ms) on GPU 4264 (ms) on CPU
[128]	food recognition	CNN	22	Used GoogLeNet [75]	NA	50 (s)
[135]	Crop recognition	CNN	6	C(96×7×7)-P- C(96×4×4)-P-F(96)-F(96)	12 (h)	NA
[145]	Classroom Occupancy	CNN	5	C(6×28×28)-P- C(16×10×10)-P-F(120)	2.5 (h)	2 (s) (4 thread)
[146]	Fault diagnosis	AE	4	300-300-300-300	NA	91 (s)
[152]	Road damage detection	CNN	8	Used AlexNet [37]	NA	1.08 (s)
[155]	Classifying offensive plays	RNN	3	10-10-11	NA	10 (ms)

جدول 6. اندازه ها و ویژگی های DNN در برنامه های مختلف.

روش ارائه شده همچنین مبتنی بر هرس و اتصالات اضافی و غیر ضروری نورونها و همچنین استفاده از مکانیسم های تقسیم وزن است. تقسیم وزن هر وزن را با یک فهرست بیت n از یک جدول مشترک که دارای n2 مقدار ممکن است جایگزین می کند. مراحل هرس شبکه همانطور که توسط هان و همکاران شرح داده شده است شامل موارد زیر است:

- برای پیدا کردن اتصالات با وزن های زیاد ، به شبکه آموزش دهید.
- اتصالات بی اهمیت را که وزن کمتری از آستانه دارند ، هرس کنید.

- پس از هرس ممکن است برخی از سلولهای عصبی بدون اتصال ورودی و خروجی باقی بمانند. فرآیند هرس این سلولهای عصبی را مشخص می کند و آنها و همچنین تمام اتصالات باقیمانده آنها را از بین می برد.
- شبکه را مجدداً راه اندازی کنید تا وزن مدل به روز شده را تنظیم کنید. وزنها باید به جای ابتدای کار ، از مراحل آموزش قبلی منتقل شوند ، در غیر این صورت عملکرد تا حدی کاهش می یابد.

نویسندگان این رویکرد را در چهار مدل مربوط به بینایی ، یعنی AlexNet، VGG-16، LeNet-300-100 و LeNet-5 ارزیابی کردند. مدلها حداقل 9 بار برای AlexNet و 13 بار در VGG-16 فشرده شدند ، در حالی که دقت مدلها تقریباً حفظ شده بود. یک محدودیت این رویکرد این است که نمی توان از آن برای سایر مدل های DNN استفاده کرد. علاوه بر این ، شبکه های فشرده شده به دست آمده در تمام سیستم عامل های سخت افزاری و معماری CPU به اندازه کافی کارآمد نیستند ، بنابراین به انواع جدیدی از شتاب دهنده نیاز دارند که بتوانند پراکندگی فعال سازی پویا و تقسیم وزن را کنترل کنند. شکل 19 مفهوم هرس DNN را نشان می دهد.

یک موتور استنتاج به نام EIE با معماری سخت افزاری ویژه و SRAM به جای DRAM طراحی شده است و به نظر می رسد که با مدل های شبکه فشرده خوب کار می کند. در این معماری ، ضرب بردار ماتریس پراکنده سفارشی و تقسیم وزن بدون از دست دادن کارایی شبکه به کار گرفته می شود. موتور از یک عنصر مقیاس پذیر از عناصر پردازش (PE) تشکیل شده است ، که هر یک بخشی از شبکه را در یک SRAM نگه می دارند و محاسبات مربوطه را انجام می دهند. از آنجا که بیشتر انرژی مورد استفاده در شبکه های عصبی برای دسترسی به حافظه مصرف می شود ، مصرف انرژی با این شتاب دهنده طراحی شده 120 برابر کمتر از مصرف انرژی شبکه اصلی مربوطه است.

در HashedNets ، وزن اتصال به شبکه عصبی بطور تصادفی با استفاده از یک عملکرد هش در سطل های هش دسته بندی می شود. تمام اتصالات که در یک سطل یکسان قرار می گیرند توسط یک پارامتر واحد نمایش داده می شوند. انتشار بازگشتی برای تنظیم دقیق پارامترها در طول آموزش استفاده می شود. نتایج آزمایش نشان می دهد که صحت این مدل فشرده سازی مبتنی بر هش نسبت به سایر روش های پایه فشرده سازی بهتر است.

این کار توسط کورباریاکس و همکاران پیشنهاد شده است که وزن و شبکه های عصبی را در هر دو مرحله استنتاج و در کل مراحل آموزش باینریزه کند تا بتواند ردپای حافظه و دسترسی ها را کاهش دهد. این شبکه همچنین می تواند بیشتر عملیات حسابی را از طریق عملیات بیت خرد انجام دهد و منجر به کاهش مصرف برق شود. مجموعه داده های MNIST، CIFAR-10 و Street View House Number (SVHN) در چارچوب های Torch7 و Theano با استفاده از این رویکرد مورد آزمایش قرار گرفت و نتایج به دست آمده امیدوار کننده بود.

2) محاسبات تقریبی:

محاسبات تقریبی روش دیگری برای اجرای ابزارهای یادگیری ماشین در دستگاه های IoT و کمک به صرفه جویی در مصرف انرژی در میزبان آنها است. اعتبار این رویکرد از این واقعیت ناشی می شود که در بسیاری از برنامه های IoT ، خروجی های یادگیری ماشین (به عنوان مثال پیش بینی ها) لازم نیست که دقیق باشند ، بلکه در یک محدوده قابل قبول قرار می گیرند که کیفیت مطلوب را ارائه می دهند. در واقع ، این رویکردها باید آستانه های کیفیتی را تعریف

کنند که خروجی نباید از آن عبور کند. تلفیق مدل های DL با محاسبات تقریبی می تواند به مدلهای DL کارآمدتر برای دستگاههای دارای منبع محدود منجر شود. ونکاتارامانی و همکارانش گسترش محاسبات تقریبی به شبکه های عصبی را پیشنهاد داده و یک شبکه عصبی را به یک شبکه عصبی تقریبی تبدیل کردند. در رویکردشان ، نویسندگان با هدف شناسایی نورونهایی که کمترین تأثیر را بر صحت خروجی دارند ، از پردازش مجدد استفاده می کنند. سپس ، NN تقریبی با جایگزینی نورونهای کم اهمیت در شبکه اصلی با همتایان تقریبی آنها شکل می گیرد. ساخت نورون تقریبی با یک روش طراحی تقریبی به نام مقیاس گذاری دقیق انجام می شود. بجای استفاده از یک عدد ثابت بیت معمولی (فرمت 16 بیتی یا 32 بیتی) برای ارائه محاسبات ، تعداد مختلفی بیت (4 - 10 بیت) در این تکنیک استفاده می شود. پس از تشکیل شبکه تقریبی ، دقت در ورودی ها و وزن نورون ها تنظیم می شود تا در یک رابطه بهینه بین دقت و انرژی به وجود آید. همچنین تلاشهای دیگری نیز وجود دارد که از کاربرد محاسبات تقریبی با مقیاس گذاری دقیق بر روی CNN ها و DBN ها گزارش کرده است. با این حال ، تمرین فعلی مستلزم این است که فرایند آموزش مدل و تبدیل آن به تقریب DL در یک سکوی غنی از منابع صورت گیرد و سپس مدل تبدیل شده در یک وسیله محدودکننده منبع مستقر شود.

3) شتاب دهنده ها:

طراحی سخت افزارها و مدارهای خاص ، یکی دیگر از جهات تحقیقاتی فعال با هدف بهینه سازی بهره وری انرژی و ردپای حافظه مدل های DL در دستگاه های IoT است. تمرکز چنین کارهای پژوهشی بر زمان استنباط مدلهای DL است ، زیرا روند آموزش مدلهای پیچیده زمان و انرژی زیادی می برد. چندین روش برای بهبود هوش دستگاه های IoT از جمله طراحی شتاب دهنده برای DNN و استفاده از فناوری های Post-CMOS مانند اسپینترونیک که از مکانیسم چرخش الکترونی استفاده می کند ، شناسایی شده اند. این فناوری اخیر یک گام به سمت توسعه دستگاههای ترکیبی را نشان می دهد که می توانند داده ها را ذخیره کنند ، محاسبات و ارتباطات را در همان فناوری مواد انجام دهند.

تحقیقات انجام شده و تحقیقات مربوط به توسعه ، شتابدهنده های DNNs و CNN را به ترتیب گزارش داده است. فراتر از شتاب دهنده های سخت افزاری ، کار در پیشنهاد استفاده از شتاب دهنده نرم افزار برای مرحله استنتاج مدل های DL در دستگاه های تلفن همراه است. این دو الگوریتم کنترل منابع را در زمان اجرا به کار می برند ، یکی لایه ها را فشرده می کند و دیگری مدل های معماری عمیق را در پردازنده های موجود تجزیه می کند. این شتاب دهنده نرم افزار می تواند یک راه حل مکمل برای طراحی های شتاب دهنده سخت افزاری باشد.

4) Tinymotes:

علاوه بر تمام راه حل های قبلی ، پیشرفت پردازنده های کوچک (میکروموت) با قابلیت های قوی DL رو به افزایش است. طیف وسیعی به اندازه یک میلی متر مکعب طراحی شده است ، چنین پردازنده هایی را می توان با باتری ها اداره کرد ، در حالی که در هنگام انجام آنالیز و پیش بینی پردازنده از طریق شتاب دهنده های شبکه عمیق ، تنها حدود 300 میکروووات مصرف می کند. با استفاده از این فناوری ، بسیاری از برنامه های مهم IoT می توانند به جای ارسال داده به رایانه های با کارایی بالا و در انتظار پاسخ آنها ، تصمیم گیری را روی دستگاه انجام دهند. برای برنامه های

کاربردی که امنیت داده ها و حفظ حریم خصوصی اصلی ترین نگرانی ها هستند ، این ادغام سخت افزار و DL این نگرانی ها را تا حدودی کاهش می دهد ، زیرا برای تجزیه و تحلیل هیچ یا فقط مقدار محدودی نیاز به داده ابری است. مونز و همکاران همچنین یک پردازنده کوچک برای CNN ها (کل مساحت فعال 1.2 * 2 میلی متر مربع) تولید کرده اند که از نظر انرژی کارآمد است (مصرف برق 25 تا 288 مگاوات است).

ب) برنامه ها:

برنامه های موبایلی وجود دارد که DNN های از قبل آموزش دیده را برای انجام کارهای تحلیلی و پیش بینی خود استفاده می کند ، مانند استفاده از CNN برای شناسایی زباله در تصاویر. با این حال ، مصرف منابع در این برنامه ها هنوز هم بسیار زیاد است. در واقع ، زمانی در حدود 5.6 ثانیه برای بازگشت پاسخ پیش بینی ، ضمن مصرف 83 درصد از CPU و 67 MB از حافظه گزارش می کند. هوارد و همکاران معماری MobileNets را برای استفاده در برنامه های دیداری موبایل و جاسازی شده پیشنهاد کردند. با بازسازی یک مدل پیچیده از طریق فاکتور سازی یک لایه پیچش استاندارد به یک محور عمقی و یک نتیجه گیری 1 × 1 ، آنها قادر به دستیابی به مدل های کوچکتر و محاسباتی کارآمد برای GoogleNet و VGG16 شدند. آنها همچنین چندین مورد استفاده از مدل خود را از جمله تشخیص شیء ، طبقه بندی تصویر و شناسایی ویژگی های چهره نشان دادند.

آماتو و همکارانش یک CNN را روی تابلوهای Raspberry Pi که در دوربین های هوشمند گنجانده شده اند برای یافتن شکاف های خالی پارکینگ اجرا کردند. راوی و همکاران از توسعه یک برنامه تناسب اندام برای دستگاه های تلفن همراه استفاده کرده اند که از DL برای طبقه بندی فعالیت های انسانی استفاده می کند.

Method / Technology	Reference	Pros	Cons
Network Compression	[184] [186] [187]	• Reduce storage and computation	• Not general for all DL models • Need specific hardware • The pruning process bring overload to training
Approximate Computing	[188], [189]	• Makes fast DL models • Save energy	• Not suitable for precise systems
Accelerators	[91] [185] [190] [191] [192] [193]	• Integrates DL model with the hardware • Efficient computations	• Does not work with the traditional hardware platforms
Tinymote with DL	[194]	• Good for time-critical IoT apps • Energy-efficient • Provides more security and privacy for data	• Special-purpose networks

جدول 7. روش ها و فناوری هایی برای آوردن DL روی دستگاه های IOT.

مدل DL بر روی یک ماشین استاندارد آموزش داده شده و سپس جهت تشخیص فعالیت به سکوی موبایل منتقل می شود. با این حال ، ورودی مدل DL برای بهبود دقت ، با چندین ویژگی مهندسی مخلوط شده است. همانطور که نویسندگان توصیف می کنند ، تعداد کمی از لایه ها در مدل های DNN اختصاص داده شده برای یک وسیله محدود منابع ، یک دلیل بالقوه برای دستیابی به عملکرد ضعیف است. علاوه بر این ، اگر داده های آموزشی به خوبی نمایانگر کل اکوسیستم نباشد ، عملکرد رضایت بخش نخواهد بود.

نگوین و همکاران برای پشتیبانی از برنامه های هوشمند IoT یک چارچوب نرم افزاری سخت افزاری مفهومی را پیشنهاد کردند. چارچوب آنها از یک موتور شناختی و یک جزء اتصال هوشمند تشکیل شده است. موتور شناختی که

قابلیت شناختی را برای اشیاء هوشمند فراهم می کند ، از هر دو الگوریتم DL و تجزیه و تحلیل تصمیم گیری نظری بازی استفاده می کند. این الگوریتم ها باید در پردازنده های مخصوص برنامه قدرت پایین مستقر شوند تا برای IoT مناسب باشد. مؤلفه اتصال هوشمند با فرستنده های رادیویی شناختی و پردازنده های باند پایه ادغام می شود تا از اتصالات قابل انعطاف و قابل اعتماد با اشیاء هوشمند IoT استفاده کند.

ج) درس های آموخته شده:

در این بخش ، لزوم حرکت به سمت پشتیبانی از DL در دستگاه های جاسازی شده و منبع محدود IoT مورد بحث قرار گرفته است. ویژگی های مختلف دستگاه های IoT و تکنیک های DL این مسیر را چالشی تر می کند زیرا دستگاه های IoT به ندرت می توانند مدلهای DL را میزبانی کنند حتی به دلیل محدودیت منابع خودشان فقط پیش بینی ها را انجام دهند. برای مقابله با این چالش ها ، روش های مختلفی در ادبیات اخیر معرفی شده است از جمله:

- فشرده سازی DNN
- محاسبه تقریبی برای DL
- شتاب دهنده ها
- Tinymotes با DL.

این رویکردها بر عملکرد استنباط مدلهای DL موجود یا از قبل آموزش دیده تمرکز دارند. بنابراین ، آموزش مدلهای DL در دستگاههای محدود و منابع جاسازی شده هنوز یک چالش جدی است. تغییر روند آموزش به دستگاه های IoT برای استقرار مقیاس پذیر و توزیع شده دستگاه های IoT مطلوب است. به عنوان مثال ، با داشتن صدها دوربین امنیتی هوشمند مستقر در یک جامعه برای تأیید هویت چهره ، روند آموزش برای هر دوربین می تواند در سایت انجام شود. فشرده سازی شبکه شامل شناسایی اتصالات و نورون های بی اهمیت در یک DNN از طریق چندین دوره آموزش است. در حالی که این یک روش امیدوار کننده برای نزدیک شدن به تجزیه و تحلیل در زمان واقعی در دستگاه های IoT است ، برای مقابله با چندین چالشی باید تحقیقات بیشتری انجام شود:

- مشخص نیست که آیا روشهای فشرده سازی شبکه برای جریان داده مناسب هستند یا نه ، به خصوص هنگامی که مدل DL پویا است و ممکن است با گذشت زمان تکامل یابد.
- روش های فشرده سازی برای معماری های سری زمانی ، مانند RNN و LSTM ، به خوبی مورد بررسی قرار نگرفته است ، و این شکاف وجود دارد که آیا روش های فشرده سازی موجود برای این معماری DL کاربرد دارد یا خیر.
- نیاز به مشخص کردن معامله بین میزان فشرده سازی و دقت DNN وجود دارد ، زیرا فشرده سازی بیشتر منجر به تخریب دقت می شود.

اخیراً روشهای تقریبی محاسباتی نیز در ساخت مدلهای DL ساده تر و با صرفه تر انرژی مورد استفاده قرار گرفته است تا بتوان آنها را در دستگاههای دارای محدودیت منابع کار گذاشت. مشابه روشهای فشرده سازی شبکه ، این روش ها از نورون های ناچیز نیز بهره می گیرند. با این حال ، به جای دستکاری ساختار شبکه ، آنها ساختار را حفظ می کنند اما بازنماهای محاسبات را از طریق کاهش طول بیتها تغییر می دهند. به همین دلیل ، به نظر می رسد آنها برای انواع معماری DL قابل اجرا هستند و حتی می توانند تحول پویای مدل های شبکه را در زمان اجرا پوشش دهند. حفظ تعادل بین دقت و مصرف انرژی ، هدف مشترک آنها است. با این وجود ، برای یافتن برتری یکی از این رویکردها برای تعبیه مدلهای DL در دستگاههای IoT ، کارهای بیشتری لازم است.

علاوه بر این ، ما در مورد ظهور سخت افزار شکل خاص و کوچک که به منظور اجرای کارآمد مدل های DL در دستگاه های محدود شده و منبع استفاده شده است ، بحث کردیم. این معماری ها به دلیل کاهش تقاضای منابع و کاربرد آنها در برنامه های IoT حساس به زمان ، در دستگاه های پوشیدنی ، سیار و IoT قابل استفاده هستند. با این حال ، کلی بودن آنها برای پشتیبانی از هر نوع DNN و همچنین قابلیت همکاری و سازگاری آنها با سیستم عامل های سخت افزاری موجود همچنان به عنوان چالش های واضح باقی مانده است.

جدول 7 روش ها و فن آوری های استفاده شده در ادبیات اخیر را برای میزبانی از تجزیه و تحلیل DL در دستگاه های IoT به همراه جوانب مثبت و منفی آنها خلاصه می کند.

ما همچنین برخی از برنامه های کاربردی که DL را در دستگاه های محدودکننده منابع پیاده سازی کرده اند ، مرور کردیم. با توجه به چالش های ذکر شده ، برنامه های کاربردی خوبی در این دسته وجود ندارد. هرچند که با برطرف کردن این چالش ها و موانع ، شاهد ظهور بسیاری از برنامه های IoT خواهیم بود که مدل اصلی DL آنها در حسگرها ، محرک ها و اشیاء هوشمند IoT تعبیه شده است.

6. DL مه و ابر محور برای IOT

محاسبات ابری یک راه حل امیدوار کننده برای تجزیه و تحلیل داده های بزرگ IoT به حساب می آید. با این وجود ، ممکن است برای داده های IoT با محدودیت های امنیتی ، قانونی / سیاسی ایده آل روبرو باشد به عنوان مثال ، داده ها نباید به مراکز ابری که در خارج از قلمرو ملی میزبانی می شوند منتقل شوند یا مثلا محدودیتهای زمانی. از سوی دیگر ، انتزاع سطح بالا از داده ها برای برخی از اهداف تحلیلی باید با جمع آوری چندین منبع داده IoT بدست آید. از این رو ، در این موارد نمی توان از راه حل های تحلیلی بر روی گره های IoT جداگانه استفاده کرد.

به جای قرار گرفتن صرفا در فضای ابری ، ایده نزدیک کردن محاسبات و آنالیزها به کاربران یا دستگاههای نهایی اخیراً تحت نام محاسبات مه ارائه شده است. با تکیه بر تجزیه و تحلیل مبتنی بر مه ، می توان از مزایای محاسبات ابری ضمن کاهش یا جلوگیری از اشکالات آن ، مانند تأخیر شبکه و خطرات امنیتی ، بهره برد. نشان داده شده است که با میزبانی از تجزیه و تحلیل داده ها بر روی گره های محاسبه مه ، می توان عملکرد کلی را به دلیل جلوگیری از انتقال مقادیر زیادی از داده های خام به گره های ابری دوردست بهبود داد. همچنین ممکن است انجام تجزیه و تحلیل در زمان واقعی تا حدی امکان پذیر باشد زیرا مه در محلی نزدیک به منبع داده میزبان است. دروازه های کاربردی

هوشمند عناصر اصلی این فناوری جدید مه است که برخی از کارهایی را که اکنون توسط محاسبات ابری انجام می شود مانند جمع آوری داده ها ، طبقه بندی ، ادغام و تفسیر انجام می دهد ، بنابراین استفاده از منابع محاسباتی محلی IoT را تسهیل می کند.

این کار یک دروازه هوشمند IoT را توضیح می دهد که از مکانیزم هایی پشتیبانی می کند که با استفاده از آن ، کاربران نهایی می توانند پروتکل های برنامه را کنترل کنند تا عملکرد بهینه شود. دروازه هوشمند در درجه اول از همکاری انواع مختلفی از IoT و دستگاههای غنی از منابع پشتیبانی می کند و باعث می شود با آنها به طور مشابه رفتار شود. در دروازه هوشمند پیشنهادی ، یک ابزار تحلیلی سبک وزن برای افزایش عملکرد در سطح برنامه تعبیه شده است. مجهز بودن به دروازه های IoT و گره های لبه با الگوریتم های کارآمد DL می تواند بسیاری از کارهای پیچیده تحلیلی را که اکنون در ابر انجام می شود بومی سازی کند. جدول 8 چندین محصول را که شامل DL در هسته هوشمندشان هستند خلاصه می کند و می تواند دامنه های IoT را در مه یا ابر ارائه دهد.

در زیربخش زیر ، چندین فناوری پیشرفته را که امکان یادگیری عمیق بر روی سکوهای مه و ابر را تسهیل می کنند ، مرور می کنیم.

الف) فعال کردن فن آوری ها و بسترهای نرم افزاری

علیرغم معرفی DL در زیرساخت مه ، محاسبات ابری تنها راه حل مناسب برای تجزیه و تحلیل در بسیاری از برنامه های IoT است که با محاسبات مه قابل دستیابی نیست. به عنوان مثال ، کارهای پیچیده مانند تجزیه و تحلیل فیلم نیاز به مدل های بزرگ و پیچیده با منابع محاسباتی زیادی دارد. بنابراین ، طراحی مدلها و الگوریتم های DNN محور ابعاد مقیاس پذیر و با کارایی بالا ، که می تواند تجزیه و تحلیل را بر روی داده های گسترده IoT انجام دهد ، هنوز یک زمینه مهم برای تحقیق است. کوتس و همکاران سیستمی را مبتنی بر خوشه ای از سرورهای GPU ، در مقیاس بزرگ پیشنهاد کردند که می تواند طی چند روز آموزش شبکه های عصبی را با 1 میلیارد پارامتر در 3 دستگاه انجام دهد. این سیستم همچنین می تواند برای آموزش شبکه ها با 11 میلیارد پارامتر در 16 دستگاه مقیاس پذیر باشد.

پروژه آدام تلاش دیگری برای تهیه یک مدل DL مقیاس پذیر و کارآمد است. این سیستم مبتنی بر DL توزیع شده است که در آن محاسبات و ارتباطات کل سیستم برای مقیاس پذیری و کارآیی بالا بهینه شده است. ارزیابی این سیستم با استفاده از خوشه 120 دستگاه نشان می دهد که آموزش یک DNN بزرگ با 2 میلیارد اتصال ، در مقایسه با یک سیستم پایه ، دو برابر دقت بیشتری کسب می کند ، در حالی که از تعداد دستگاه های 30 برابر کمتری استفاده می شود.

واحد پردازش تانسور (TPU) یک پردازشگر ویژه برای DNN در مراکز داده Google است. این طرح در سال 2015 با هدف تسریع در مرحله استنباط DNN هایی که توسط چارچوب TensorFlow نوشته شده اند ، طراحی شده است. از 95 درصد نمایندگی های DNN در مراکز داده خود ، CNN ها فقط 5 درصد از حجم کار را تشکیل می دهند ، در حالی که MLP ها و LSTM ها 90 درصد دیگر را پوشش می دهند. ارزیابی عملکرد نشان داد که TPU با دستیابی به سرعت 15 تا 30 برابر سریعتر اجرای عملیات ، (در حالی که 30 تا 80 برابر انرژی کمتری در هر TeraOps / ثانیه مصرف می کند) ، به طور متوسط از GPU یا CPU های معاصر خود بهتر عمل می کند.

فراتر از پیشرفت های زیرساختی برای میزبانی مدل های DL مقیاس پذیر بر روی سیستم عامل های ابری ، به مکانیسم ها و روش های لازم برای دسترسی مدل های DL از طریق API ها نیاز است تا بتواند به راحتی در برنامه های IoT ادغام شود. این جنبه زیاد مورد بررسی قرار نگرفته است و فقط چند محصول موجود است ، مانند AWS DL AMI آمازون ، Google cloud ML و IBM Watson. این فرصت را برای ارائه دهندگان ابر ایجاد می کند تا "مدل های DL به عنوان یک سرویس" را به عنوان زیر مجموعه جدیدی از نرم افزارها (SaaS) ارائه دهند. با این حال ، این چندین چالش را برای ارائه دهندگان ابر تحمیل می کند ، زیرا وظایف DL از نظر محاسباتی فشرده هستند و ممکن است سایر سرویس های ابری را گرسنه کنند. علاوه بر این ، با توجه به تشنگی داده های مدل های DL ، انتقال داده ها به ابر ممکن است وارد تنگنا شود. به منظور ارائه تجزیه و تحلیل DL بر روی ابر ، شکل 20 یک پشتوانه کلی برای مدل های DL را به عنوان خدمات ارائه می دهد. ارائه دهندگان مختلف ممکن است از اطلاعات سفارشی شده خود استفاده کنند.

ب) چالش ها

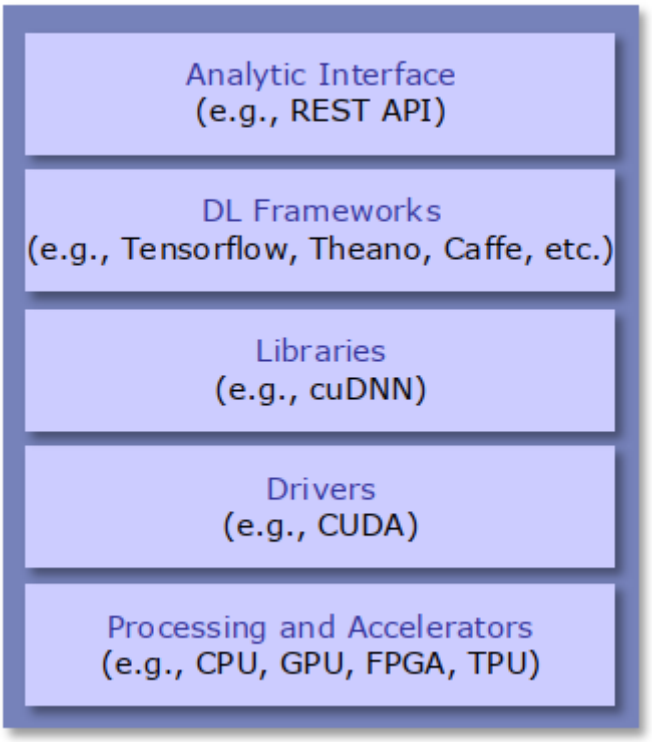
هنگامی که تجزیه و تحلیل DL به گره های مه آلود می آید ، باید چندین چالش از جمله موارد زیر مورد بررسی قرار گیرد:

• کشف سرویس DL: گره های مه در مناطق جغرافیایی متراکم توزیع می شوند و ممکن است هر گره از قابلیت های تحلیلی خاصی برخوردار باشد (به عنوان مثال ، یک گره مدل های CNN را برای تشخیص تصویر اجرا می کند ، یک گره دیگر RNN ها را برای پیش بینی داده های شایع و غیره اجرا می کند). بنابراین ، دستگاه ها باید از طریق نوعی پروتکل کشف خدمات گسترده برای تجزیه و تحلیل DL منابع ارائه دهنده های تحلیلی مناسب را شناسایی کنند.

• مدل DL و توزیع کار: پارتیشن بندی اجرای مدل های DL و وظایف در میان گره های مه ، و توزیع بهینه جریان داده در میان گره های موجود برای برنامه های حساس به زمان بسیار مهم است. جمع آوری نتایج نهایی از گره های محاسباتی و بازگشت عمل با کمترین تاخیر ، روی دیگر سکه است.

• فاکتورهای طراحی: از آنجا که محیط های محاسبات مه در مراحل ابتدایی خود قرار دارند و انتظار می رود که تکامل یابند ، شایسته است بررسی شود که چگونه عوامل طراحی محیط مه (مانند معماری ها ، مدیریت منابع و ...) و استقرار مدل های DL در این محیط می تواند بر کیفیت خدمات تحلیلی تأثیر بگذارد. از طرف دیگر ، جالب است بدانید که تا چه حد می توان این فاکتورهای طراحی را برای بهبود کارآیی و کیفیت تجزیه و تحلیل DL تنظیم کرد.

• لبه موبایل: از طریق همه گیری محیط های محاسبات لبه موبایل و سهم آنها در تجزیه و تحلیل IoT ، مهم است که پویایی چنین محیط هایی را برای طراحی آنالیز DL با کمک لبه در نظر بگیرید زیرا دستگاه های تلفن همراه ممکن است بپیوندند و سیستم را ترک کنند. همچنین ، در هنگام واگذاری وظایف تحلیلی به آنها ، مدیریت انرژی دستگاه های لبه تلفن همراه باید دقیق باشد.



شکل 20. یک پشته کلی از مدل های DL به عنوان سرویس در سیستم عامل های ابری را ارائه می دهد.

Product	Description	Application	Platform
Amazon Alexa	Intelligent personal assistant (IPA)	Smart home	Fog
Microsoft Cortana	IPA	Smart Car, XBox	Fog
Google Assistant	IPA	Smart Car, Smart home	Fog
IBM Watson	Cognitive framework	IoT domains	Cloud

جدول 8. برخی از محصولاتی که با استفاده از یادگیری عمیق و سرویس دهی از بسیاری از سایت های مه روی مه و یا ابری استفاده کرده اند.

چند تلاش از ادغام DL در گره های مه در اکوسیستم IoT خبر داد. به عنوان مثال ، اثبات مفهوم استقرار مدل های CNN بر روی گره های مه برای پیش بینی سلامت دستگاه توسط قیصر و همکاران ارائه شده است. در کار آنها ، جستجوی کاملی در بین گره های مه ، برای یافتن گره های رایگان برای واگذاری وظایف تحلیلی انجام می شود. همچنین ، لی و همکاران سیستمی را پیشنهاد کردند که با استفاده از دستگاههای تلفن همراه و لبه در حال اجرا ، مدلهای CNN برای شناسایی شیء ، ارائه می دهد.

ج) دروس آموخته شده

در این بخش ما به نقش محاسبات ابری و مه و فناوری های فعال ، سیستم عامل ها و چالش های آنها برای ارائه تحلیل DL به برنامه های IoT اشاره کرده ایم. موفقیت بزرگ محاسبات ابری در پشتیبانی از DL با پیشرفت و به کارگیری پردازنده های بهینه شده برای شبکه های عصبی و همچنین الگوریتم های توزیع پذیر و مقیاس پذیر پشتیبانی می شود. استقرار مدل های DL در سکوهای مه برای برنامه های IoT ، مانند خانه ها و شبکه های هوشمند ، به دلیل سهولت دسترسی و زمان پاسخ سریع ، توجه کاربران نهایی را به خود جلب می کند. با این وجود ، تجزیه و تحلیل DL مبتنی بر ابر می تواند برای تجزیه و تحلیل داده های طولانی مدت و پیچیده که از قابلیت های محاسبات مه جلوگیری می کند ، از اهمیت بالایی برخوردار باشد. برخی از برنامه های شهر هوشمند ، بخش دولتی و استقرار IoT در سطح کشور باید از زیرساخت های DL مبتنی بر ابر استفاده کنند.

در حال حاضر ، ادغام تجزیه و تحلیل DL در برنامه های IoT محدود به API های RESTful ، بر اساس پروتکل HTTP است. در حالی که چندین پروتکل برنامه کاربردی دیگر وجود دارد که بطور گسترده در برنامه های IoT مورد استفاده قرار می گیرد ، از جمله: Telemetry Transport (MQTT) ، پروتکل کاربرد محدود (CoAP) ، پروتکل پیام رسانی گسترده و حضور (XMPP) ، و پروتکل پیشرفته پیام رسانی (AMQP) ؛ ادغام این پروتکل ها با رابط های تحلیلی DL ، تقویت سازگاری آنها با پروتکل های فوق را برای رفع نیاز به پراکسی های تبدیل پیام ، که سربار اضافی را در زمان پاسخ تحلیلی تحمیل می کند ، می طلبد.

ما چندین چالش مربوط به استقرار و استفاده از مدل های DL را در پشتیبانی از تحلیلی بر روی گره های مه شناسایی کردیم. کشف سرویس DL به دلیل استقرار متراکم گره های مه ، یک نیاز ضروری است که باعث می شود جستجوی بی رحمانه برای خدمات موجود یک رویکرد ناکارآمد باشد. پروتکل های کشف سرویس که در حال حاضر در برنامه های IoT استفاده می کنند ، مانند multicast DNS (mDNS) یا DNS Service Discovery (DNS-SD) ، برای پشتیبانی از کشف سرویس DL (برای مثال ، اعلام نوع تحلیلی ، مدل DL ، شکل ورودی ، نیاز به گسترش دارند). توزیع کارآمد مدل ها و وظایف DL و توزیع جریان داده ها بر روی گره های مه و جمع آوری نتایج از دیگر الزاماتی است که باید مورد توجه قرار گیرد.

7. چالش های IOT برای یادگیری عمیق ، و دستورالعمل های آینده

در این بخش ابتدا چندین چالش مهم را از منظر یادگیری ماشین برای پیاده سازی و توسعه تجزیه و تحلیل IoT بررسی می کنیم. سپس ما به مسیرهای تحقیق اشاره می کنیم که می تواند شکافهای موجود برای تجزیه و تحلیل IoT را بر اساس رویکردهای DL پر کند.

الف) چالش ها

1) عدم وجود مجموعه داده بزرگ IoT:

عدم دسترسی به مجموعه داده های بزرگ دنیای واقعی برای برنامه های IoT یک مشکل اصلی برای ترکیب مدل های DL در IoT است ، زیرا داده های بیشتری برای دستیابی به دقت بیشتر برای DL مورد نیاز است.

Dataset Name	Domain	Provider	Notes	Address/Link
CGIAR dataset	Agriculture, Climate	CCAFS	High-resolution climate datasets for a variety of fields including agricultural	http://www.ccafs-climate.org/
Educational Process Mining	Education	University of Genova	Recordings of 115 subjects' activities through a logging application while learning with an educational simulator	http://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Set
Commercial Building Energy Dataset	Energy, Smart Building	IIITD	Energy related data set from a commercial building where data is sampled more than once a minute.	http://combed.github.io/
Individual household electric power consumption	Energy, Smart home	EDF R&D, Clamart, France	One-minute sampling rate over a period of almost 4 years	http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption
AMPds dataset	Energy, Smart home	S. Makonin	AMPds contains electricity, water, and natural gas measurements at one minute intervals for 2 years of monitoring	http://ampds.org/
UK Domestic Appliance-Level Electricity	Energy, Smart Home	Kelly and Knottenbelt	Power demand from five houses. In each house both the whole-house mains power demand as well as power demand from individual appliances are recorded.	http://www.doc.ic.ac.uk/~dk3810/data/
PhysioBank databases	Healthcare	PhysioNet	Archive of over 80 physiological datasets.	https://physionet.org/physiobank/database/
Saarbruecken Voice Database	Healthcare	Universität des Saarlandes	A collection of voice recordings from more than 2000 persons for pathological voice detection.	http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4
T-LESS	Industry	CMP at Czech Technical University	An RGB-D dataset and evaluation methodology for detection and 6D pose estimation of texture-less objects	http://cmp.felk.cvut.cz/t-less/
CityPulse Dataset Collection	Smart City	CityPulse EU FP7 project	Road Traffic Data, Pollution Data, Weather, Parking	http://iot.ee.surrey.ac.uk:8080/datasets.html
Open Data Institute - node Trento	Smart City	Telecom Italia	Weather, Air quality, Electricity, Telecommunication	http://theodi.fbk.eu/openbigdata/
Málaga datasets	Smart City	City of Malaga	A broad range of categories such as energy, ITS, weather, Industry, Sport, etc.	http://datosabiertos.malaga.eu/dataset
Gas sensors for home activity monitoring	Smart home	Univ. of California San Diego	Recordings of 8 gas sensors under three conditions including background, wine and banana presentations.	http://archive.ics.uci.edu/ml/datasets/Gas+sensors+for+home+activity+monitoring
CASAS datasets for activities of daily living	Smart home	Washington State University	Several public datasets related to Activities of Daily Living (ADL) performance in a two-story home, an apartment, and an office settings.	http://ailab.wsu.edu/casas/datasets.html
ARAS Human Activity Dataset	Smart home	Bogazici University	Human activity recognition datasets collected from two real houses with multiple residents during two months.	https://www.cmpe.boun.edu.tr/aras/
MERLSense Data	Smart home, building	Mitsubishi Electric Research Labs	Motion sensor data of residual traces from a network of over 200 sensors for two years, containing over 50 million records.	http://www.merl.com/wmd
SportVU	Sport	Stats LLC	Video of basketball and soccer games captured from 6 cameras.	http://go.stats.com/sportvu
RealDisp	Sport	O. Banos	Includes a wide range of physical activities (warm up, cool down and fitness exercises).	http://orestibanos.com/datasets.htm

جدول 9. مجموعه داده های مشترک برای یادگیری عمیق در IOT.