**اثرات، کاربردها و نقش دیپ لرنینگ در دیتابیس**

هانیه حسنیان - بردیا نیک بخش

زمستان 1401

فهرست

[یادگیری عمیق 3](#_Toc126586648)

[مقدمه 4](#_Toc126586649)

[دیپ لرنینگ در دیتابیس 6](#_Toc126586650)

[1 – رابط کوئری (پرس و جو) 6](#_Toc126586651)

[2 – طرح های کوئری (پرس و جو) 7](#_Toc126586652)

[3 – جمع سپاری و پایگاه های دانش 8](#_Toc126586653)

[4 – داده‌های موقتی و پرت 8](#_Toc126586654)

# یادگیری عمیق

یادگیری عمیق اخیراً به دلیل موفقیت باورنکردنی آن در بسیاری از برنامه‌های پیچیده مبتنی بر داده، از جمله طبقه‌بندی تصویر و تشخیص گفتار، بسیار محبوب شده است. جامعه توسعه دهندگان پایگاه داده سال ها بر روی برنامه های کاربردی مبتنی بر داده کار کرده است و بنابراین باید نقش اصلی را در پشتیبانی از این موج جدید ایفا کند. با این حال، پایگاه‌های داده و یادگیری عمیق هم از نظر تکنیک‌ها و هم از نظر کاربردها متفاوت هستند. در این مقاله به بررسی مشکلات تحقیق در تقاطع دو حوزه می پردازیم. به طور خاص، ما در مورد بهبودهای احتمالی سیستم‌های یادگیری عمیق از منظر پایگاه داده بحث می‌کنیم و برنامه‌های پایگاه داده‌ای را که ممکن است از تکنیک‌های یادگیری عمیق سود ببرند، تجزیه و تحلیل می‌کنیم.

# مقدمه

در سال‌های اخیر، ما شاهد موفقیت برنامه‌های مبتنی بر یادگیری ماشینی **داده‌محور** بوده‌ایم. این امر جامعه پایگاه داده را بر آن داشته است تا فرصت های یکپارچه سازی تکنیک های یادگیری ماشین را در طراحی سیستم ها و برنامه های پایگاه داده بررسی کند. شاخه ای از یادگیری ماشین، به نام یادگیری عمیق، در سال های اخیر به دلیل عملکرد عالی آن در زمینه های مختلف از جمله تشخیص گفتار، طبقه بندی تصویر و پردازش زبان طبیعی (NLP) توجه جهانی را به خود جلب کرده است. اساس یادگیری عمیق حدود بیست سال پیش در قالب شبکه های عصبی ایجاد شد. تجدید حیات اخیر آن عمدتاً توسط سه عامل تحریک می شود: **قدرت محاسباتی بسیار زیاد**، که زمان آموزش و استقرار مدل های جدید را کاهش می دهد، به عنوان مثال. واحد پردازش گرافیکی (GPU) سیستم های آموزشی را قادر می سازد بسیار سریعتر از سیستم های دهه 1990 اجرا شوند. **مجموعه داده های آموزشی عظیم** (برچسب گذاری شده) (به عنوان مثال ImageNet) دانش جامع تری را در مورد دامنه به دست می آورد. **مدل های جدید** یادگیری عمیق (به عنوان مثال AlexNet) توانایی ضبط منظم داده ها را بهبود می بخشد. محققان پایگاه داده از دهه 1970 بر روی بهینه سازی سیستم و برنامه های کاربردی مبتنی بر داده در مقیاس بزرگ کار کرده اند که ارتباط نزدیکی با دو عامل اول دارند. طبیعی است که به روابط بین پایگاه های داده و یادگیری عمیق فکر کنیم. نشان داده شده است که مجموعه داده های آموزشی بزرگتر و ساختار مدل عمیق تر، دقت مدل های یادگیری عمیق را بهبود می بخشد. با این حال، نتیجه جانبی این است که آموزش پرهزینه تر می شود. رویکردهایی برای تسریع سرعت آموزش از دیدگاه سیستم و تئوری پیشنهاد شده است. از آنجایی که جامعه پایگاه داده تجربه غنی با بهینه سازی سیستم دارد، مناسب است که در مورد کاربرد تکنیک های پایگاه داده برای بهینه سازی سیستم های یادگیری عمیق بحث کنیم. به عنوان مثال، **محاسبات توزیع شده** و **مدیریت حافظه** از فناوری های کلیدی پایگاه داده هستند. آنها همچنین در یادگیری عمیق نقش اساسی دارند.

# دیپ لرنینگ در دیتابیس

برنامه های یادگیری عمیق، مانند بینایی کامپیوتر و NLP، ممکن است بسیار متفاوت از برنامه های کاربردی پایگاه داده به نظر برسند. با این حال، ایده اصلی یادگیری عمیق، که به عنوان یادگیری ویژگی (یا بازنمایی) شناخته می شود، برای طیف گسترده ای از برنامه ها قابل اجرا است. به طور شهودی، هنگامی که نمایش‌های مؤثری برای موجودیت‌ها، به عنوان مثال، تصاویر، کلمات، ردیف‌ها یا ستون‌های جدول داشته باشیم، می‌توانیم شباهت موجودیت را محاسبه کنیم، خوشه‌بندی کنیم، مدل‌های پیش‌بینی را آموزش دهیم، و داده‌ها را با روش‌های مختلف و غیره بازیابی کنیم. ما چند مدل یادگیری عمیق را برجسته خواهیم کرد. که می تواند برای برنامه های پایگاه داده زیر سازگار شود.

## 1 – رابط کوئری (پرس و جو)

رابط‌های پرس و جوی زبان طبیعی برای چندین دهه به دلیل مطلوبیت زیادشان، به ویژه برای کاربران غیر متخصص پایگاه داده، وایرال شده‌اند. با این حال، برای سیستم های پایگاه داده تفسیر (یا درک) معنای پرس و جوهای زبان طبیعی چالش برانگیز است. اخیراً، مدل‌های یادگیری عمیق به عملکرد پیشرفته‌ای برای وظایف **NLP** دست یافته‌اند. علاوه بر این، نشان داده شده است که RNN قادر به یادگیری خروجی ساختاریافته است. به عنوان یکی از این راه‌حل‌ها، می‌توانیم مدل‌های RNN را برای تجزیه پرس‌و‌جوهای زبان طبیعی برای تولید پرسش‌های SQL و اصلاح آن با استفاده از رویکردهای پایگاه داده موجود اعمال کنیم. به عنوان مثال، **قوانین اکتشافی** را می توان برای تصحیح خطاهای دستور زبان در جستارهای SQL تولید شده به کار برد. چالش این است که مقدار زیادی از نمونه های آموزشی (برچسب‌دار) برای آموزش مدل مورد نیاز است. یک راه حل ممکن، آموزش یک مدل پایه با مجموعه داده کوچک و اصلاح تدریجی آن با بازخورد کاربران است. به عنوان مثال، کاربران می توانند به تصحیح پرس و جوی SQL تولید شده کمک کنند، و این بازخوردها اساساً به عنوان داده های برچسب گذاری شده برای آموزش های بعدی خدمت می کنند.

## 2 – طرح های کوئری (پرس و جو)

بهینه سازی طرح پرس و جو یک مشکل پایگاه داده سنتی است. اکثر سیستم های پایگاه داده فعلی از مدل های اکتشافی و هزینه پیچیده برای تولید طرح پرس و جو استفاده می کنند. با توجه به، هر طرح پرس و جو از یک الگوی پرس و جو پارامتری SQL دارای یک منطقه بهینه است. تا زمانی که پارامترهای پرس و جوی SQL در این منطقه هستند، طرح پرس و جو بهینه تغییر نمی کند. به عبارت دیگر، طرح های پرس و جو نسبت به تغییرات کوچک پارامترهای ورودی حساس نیستند. بنابراین، ما می توانیم برنامه ریز پرس و جو را آموزش دهیم که از مجموعه ای از جفت پرس و جوهای SQL و طرح های بهینه یاد می گیرد تا طرح های مشابه برای پرس و جوهای جدید (مشابه) تولید کند. برای توضیح بیشتر، می‌توانیم یک مدل RNN آموزش دهیم که عناصر پرس و جوی SQL و متا داده‌ها (مانند اندازه بافر و کلید اصلی) را به عنوان ورودی می‌پذیرد و یک ساختار درختی تولید می‌کند که طرح پرس و جو را نشان می‌دهد. توجه داشته باشید که رویکردهای صرفاً مبتنی بر مدل‌های یادگیری عمیق ممکن است چندان مؤثر نباشند. به طور خاص، مجموعه داده آموزشی ممکن است جامع نباشد که شامل تمام الگوهای پرس و جو باشد، به عنوان مثال نتایج ممکن است در مجموعه داده های آموزشی گم شوند. برای حل این مشکلات، یک رویکرد بهتر ترکیب راه حل های پایگاه داده و یادگیری عمیق است.

## 3 – جمع سپاری و پایگاه های دانش

بسیاری از کاربردهای جمع‌سپاری و پایگاه دانش شامل استخراج موجودیت، ابهام‌زدایی و مشکلات ترکیبی هستند، جایی که موجودیت می‌تواند یک ردیف از پایگاه داده، یک گره در یک نمودار و غیره باشد. با پیشرفت‌های مدل‌های یادگیری عمیق در NLP، فرصتی است که یادگیری عمیق را برای این مشکلات به طور خاص در نظر بگیریم. می‌توانیم نمایش‌هایی را برای موجودیت‌ها یاد بگیریم و سپس با استفاده از نمایش‌های آموخته‌شده، استدلال رابطه موجودیت و محاسبه شباهت را انجام دهیم.

## 4 – داده‌های موقتی و پرت

داده‌های مکانی و زمانی انواع داده‌های رایج در سیستم‌های پایگاه داده هستند و معمولاً برای تحلیل روند، مدل‌سازی پیشرفت و تحلیل پیش‌بینی استفاده می‌شوند. داده های مکانی معمولاً با نگاشت اجسام متحرک در بلوک های مستطیلی پردازش می شوند. اگر هر بلوک را پیکسلی از یک تصویر در نظر بگیریم، مدل‌های یادگیری عمیق، به عنوان مثال، CNN، می‌توانند برای استخراج موقعیت مکانی بین بلوک‌های مجاور مورد سوء استفاده قرار گیرند. به عنوان مثال، اگر داده‌های موقعیت مکانی بی‌درنگ (مثلاً داده‌های GPS) اجسام متحرک را داشته باشیم، می‌توانیم یک مدل CNN برای ثبت روابط چگالی مناطق مجاور برای پیش‌بینی ترازویی یاد بگیریم.