چکیده

به طور کلی، اغلب سامانه‌های دسته‌بندی ترافیک، مبتنی بر ویژگی‌های از قبل استخراج‌شده توسط یک فرد خبره در زمینه شبکه‌های کامپیوتری هستند. این ویژگی‌ها مواردی همچون عبارات منظم مخصوص هر پروتکل، شماره درگاه، اطلاعات موجود در سرآیند لایه‌های مختلف و ویژگی‌های آماری استخراجی از جریان را شامل می‌شوند. مشکل اصلی تحلیل ترافیک به منظور دسته‌بندی و یا کشف ناهنجاری، یافتن ویژگی‌های مناسب در ترافیک است. فرآیند یافتن ویژگی‌های مناسب، عموماً امری زمان‌بر است و نیاز به فردی خبره دارد که این ویژگی‌ها را معین کرده و اقدام به استخراج آن‌ها نماید. همچنین دسته‌بندی برخی از انواع مختلف ترافیک مانند ترافیک رمزشده با استفاده ویژگی‌های مذکور، غیر ممکن است. مجتمع نبودن سامانه استخراج ویژگی و دسته‌بندی هم از دیگر مشکلات سامانه‌های تشخیص ترافیک است که منجر به زمان‌بر و هزینه‌بر شدن این امر می‌گردد. برای حل مشکلات مذکور ما روشی مبتنی بر یادگیری ژرف و شبکه‌های عصبی مصنوعی را برای یادگیری ویژگی و دسته‌بندی ترافیک ارائه می‌کنیم. ما با استفاده از شبکه‌های خودرمزنگار و شبکه‌های پیچشی اقدام به دسته‌بندی و مشخصه‌سازی ترافیک شبکه می‌کنیم. نتایج شبیه‌سازی‌های این روش نشان می‌دهند که روش پیشنهادی موفق به شناسایی ترافیک رمزشده می‌شود . دسته‌بند پیشنهادی با وجود افزایش تعداد دسته‌ها، دقت روش‌های موجود دسته‌بندی کاربرد را در حدود ۲ درصد و دقت مشخصه‌سازی را نیز ۶ درصد افزایش می‌دهد.

کلمات کلیدی: یادگیری ژرف، یادگیری ویژگی، تشخیص ترافیک، شبکه‌های عصبی مصنوعی.

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱

مقدمه

امروزه با پیشرفت فنّاوری، حجم و تراکم ترافیک شبکه‌های کامپیوتری روز به روز در حال افزایش است و همین امر سبب ظهور انواع پروتکل‌های مختلف و جدید شده است. تحلیل این دادگان عظیم در شبکه‌های تجاری بزرگ به امری مهم برای صاحبان آن شبکه‌ها تبدیل شده‌است. شرکت‌هایی نیز با خدماتی مبتنی بر شناخت پروتکل و یا کشف ناهنجاری، به وجود آمده‌اند که این خود بیانگر اهمیت و ارزش اینگونه تحلیل‌ها در شبکه اینترنت کنونی است.

روش‌های مرسوم برای این کار، مانند روش‌های مبتنی بر نشانه[۱]، روش‌های آماری[۲] و روش‌های مبتنی‌بر درگاه[۳] در عمل دقت بالایی از خود نشان نمی‌دهند. علاوه بر این، در این روش‌ها به تجربه یک فرد خبره نیاز است. هدف محققان در این حوزه، آن است که سامانه‌های تشخیص ترافیک و ناهنجاری نه‌تنها به صورت خودکار و بدون نیاز به دخالت عامل خارجی، وظیفه خود را به خوبی انجام بدهند، بلکه سامانه‌ها در عمل نیز دقت بالایی در تشخیص داشته‌باشند.

یافتن ویژگی‌های مناسب در ترافیک شبکه‌های کامپیوتری به امری چالش‌برانگیز در شبکه‌های ارتباطی تبدیل شده است. این امر ارتباط بسیار زیادی با فهم دقیق از ساختار و پویایی ترافیک اینترنت دارد که خود کاربردهای بسیاری در مدیریت و نظارت در شبکه‌های فراهم‌کنندگان اینترنت[۴] دارد.

به طور کلی ویژگی‌های ترافیک، ورودی بسیاری از سامانه‌های مدیریت شبکه مانند برنامه‌ریزی برای ظرفیت شبکه، مهندسی ترافیک، تشخیص خطا، تشخیص ناهنجاری و تعیین قیمت اینترنت است.

تلاش‌های بسیای برای تحلیل و اندازه‌گیری ترافیک اینترنت انجام شده‌است. بسیاری از این تحقیقات نشان‌دهنده این

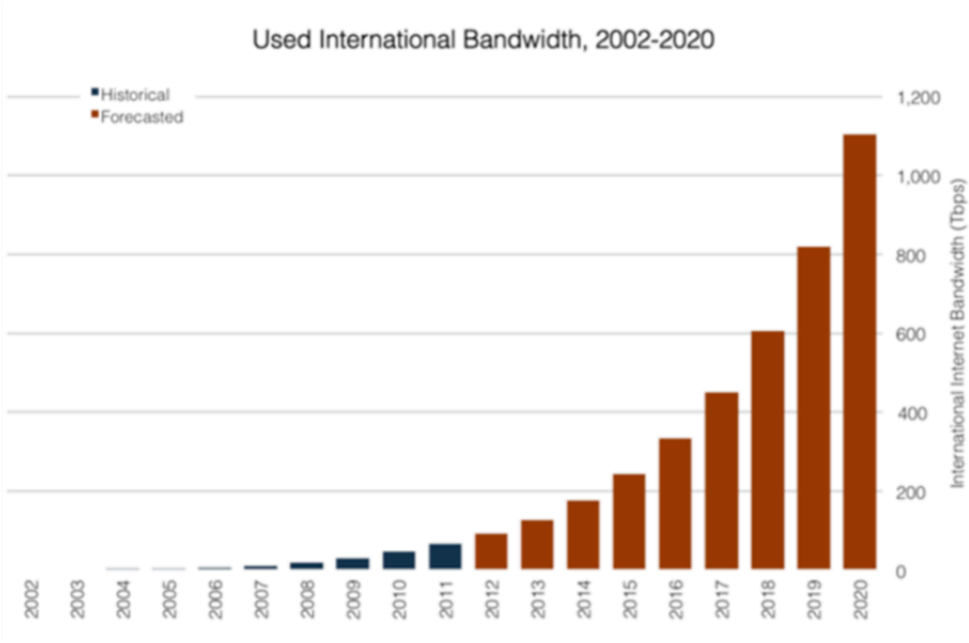
1. Signature-based Methods
2. Statistical Methods
3. Port-based Methods
4. Internet Service Providers

امر است که حدود ۸۰ درصد ترافیک جاری در شبکه اینترنت متعلق به برنامه‌ها و کاربردهای نظیر به نظیر[۱] است. همچنین در بررسی‌های اخیر [۹] این امر مشخص شده‌است که اشتراک فایل‌های ویدئویی بسیار افزایش پیدا کرده و ترافیک جاری در شبکه حاصل از این پدیده از ترافیک جاری متعلق به کاربردهای اشتراک‌گذاری فایل و برنامه‌های نظیر به نظیر پیشی گرفته‌است. البته باید به این نکته توجه داشت که یک اندازه‌گیری دقیق از ترافیک جاری در اینترنت کاری بسیار سخت و طاقت‌فرسا است. تحلیل‌ها و بررسی‌های گذشته هم در این راستا از محدودیت‌هایی همچون طول زمان اندازه‌گیری و پوشش جغرافیایی، هدررفت اطلاعات هنگام فرآیند اندازه‌گیری و شکست در شناسایی کاربرد مربوطه رنج می‌برند.

علاوه براین، دسترسی مردم به ارتباطات پرسرعت به‌ویژه فناوری‌های مبتنی بر ADSL [۲] و تلویزیون‌های کابلی روزبه‌روز در حال افزایش است. با توجه به این که این ارتباطات پرسرعت همواره در دسترس هستند از این رو کاربران به استفاده از کاربردها و برنامه‌های موجود بر روی شبکه اینترنت کنونی مانند VoIP[۳]، تجارت الکترونیکی، بانکداری اینترنتی و سیستم‌های نظیر به نظیر برای به اشتراک گذاری ویدئو و موسیقی ترغیب می‌شوند. به بیانی دیگر این افزایش ظرفیت و سرعت به پیچیده شدن رفتار کاربران نسبت به سایر روش‌های قدیمی اتصال به اینترنت منجر شده است. همان‌طور که در شکل ۱-۱ (برگرفته از [۱۰]) معلوم است مقدار پهنای باند در دسترس برای کاربران در دهه گذشته تقریبا دو برابر شده است. این در حالی است که تعداد کاربران هم به طور نمایی در حال افزایش است. از دیدگاه مدیریت شبکه، این حجم از افزایش کاربران و پهنای باند پیچیدگی نظارت بر شبکه را بسیار افزایش می‌دهد. از دیدگاه کیفی افزایش مقدار ترافیک می‌تواند به کاهش کیفیت دریافت اطلاعات مخصوصا در برنامه‌های حساس منجر شود. اگر این ترافیک با حجم زیاد مورد کنترل قرار نگیرد موجب می‌شود که عملکرد زیرساخت شبکه دچار اختلال شود. با توجه به بحث‌های مطرح شده، فراهم‌کنندگان خدمات اینترنتی باید توجه بیشتری به این رفتار پیچیده کاربران جدید داشته باشند. علاوه بر این گرایشی که برای انتقال تماس‌های تلفنی از شبکه‌های موبایل به شبکه‌های مبتنی بر VoIP وجود دارد، عاملی است که شرکت‌های مخابراتی را تهدید می‌کند.

ساختار سلسله‌مراتبی پروتکل TCP/IP یکی از عوامل موفقیت اینترنت است. برخی مواقع فراهم‌کنندگان سرویس‌های اینترنتی می‌خواهند بدانند که بسته‌ای که در لایه‌ی سوم از ساختار پروتکل TCP/IP در حال انتقال است، متعلق به چه کاربردی است. با این کار آن‌ها می‌توانند شبکه‌های خود را مدیریت و خدمات بهتری را به مشتریان

1. Peer-to-Peer
2. Asymmetric Digital Subscriber Line
3. Voice over IP



شکل ۱-۱: سیر تکامل ترافیک اینترنت (برگرفته از [۱]).

خود عرضه نمایند. دسته‌بندی ترافیک در کاربردهای امنیتی هم بسیار حیاتی است. برای مثال فیلتر کردن ترافیک‌های ناخواسته، یا هشدار دادن به هنگام وقوع یک ناهنجاری خاص از مسائلی هستند که از خروجی سامانه‌های دسته‌بندی ترافیک همواره استفاده می‌کنند.

اطلاعاتی که سامانه‌های تشخیص ترافیک به ما می‌دهند بسیار حائز اهمیت است. برای مثال یک دانش کلی از ترکیب کلی ترافیک همراه با اطلاعات کلی از تمایل کاربران به کاربردهای مختلف، در طراحی و نگهداری شبکه‌های اطلاعاتی بسیار مهم است. سامانه‌های کیفیت خدمات[1] که وظیفه اولویت‌بندی و برخورد با ترافیک جاری در شبکه را با توجه به معیارهای مختلف دارند، نیاز است که ترافیک را ابتدا به دسته‌های مختلف تقسیم کنند.

برخی از کشورها از فراهم‌کنندگان سرویس اینترنت انتظار دارند که اجازه ورود یا خروج برخی از انواع ترافیک خاص را به شبکه‌های خود ندهند. این امر مستلزم شناسایی ترافیک و دسته‌بندی آن است. شناسایی ترافیک اولین قدم برای شناسایی اقداماتی همچون استفاده خرابکارانه از منابع در شبکه است.

همان‌طور که مشهود است دسته‌بندی ترافیک دارای کاربردهای بسیاری است. از سوی دیگر چالش‌هایی که دسته‌بندها برای دسته‌بندی ترافیک با آن مواجه هستند را نباید فراموش کرد. آن‌ها باید با حجم عظیمی از اطلاعات و نرخ انتقال

1. Quality of Service

بسیار بالا به کار بپردازند. برای حل مشکل سرعت و حجم اطلاعات، محققان به دنبال الگوریتم‌هایی هستند که حداقل بار محاسباتی را داشته باشند. چالش دیگری که بسیار مهم است این است که بسیاری از توسعه‌دهندگان نرم‌افزار به دنبال این هستند که ترافیک خود را مخفی کنند تا بتوانند با هر سرعتی و حجمی که می‌خواهند ترافیک خود را عبور دهند. رمزنگاری ترافیک و جاسازی ترافیک در قالب سایر پروتکل‌ها مثال‌های ساده‌ای هستند که به ذهن خطور می‌کند. بنابراین محققان در این حوزه تحقیقاتی باید الگوریتم‌ها و روش‌های بدیعی را برای تشخیص ترافیک با توجه به روش‌های غیرمنتظره و جدید طراحی کنند.

با نظر گرفتن مسائل مطرح‌شده استفاده از روش‌های مبتنی بر یادگیری ماشین در امر شناسایی ترافیک به دلیل وجود الگوهای مختلف در هریک از انواع مختلف ترافیک قابل توجیه است. روش‌های سابق در تشخیص ترافیک عمدتاً مبتنی بر ویژگی‌هایی بودند که افراد خبره در زمینه شبکه استخراج‌کرده‌اند. استخراج این ویژگی‌ها چون توسط انسان انجام می‌شود دارای خطا و هزینه است. همچنین در مواردی که در ترافیک مورد نظر رمزنگاری و یا کپسوله‌سازی انجام شده باشد، دقت دسته‌بندی که از این ویژگی‌ها استفاده می‌کند به شدت کاهش می‌یابد. از طرفی شناسایی تغییرات در ترافیک هم بسیار سخت است. از این رو استفاده از روش‌های یادگیری ماشین به خصوص روش‌های آماری بسیار مفید واقع می‌شود. از طرفی باید به این نکته توجه کرد که در روش‌های آماری با زیاد شدن ابعاد داده استخراج ویژگی‌های مناسب، بسیار پیچیده و سخت می‌شود. با این توصیفات، دسته‌بند ما هرچقدر هم قوی باشد اگر ویژگی‌های خوبی در ورودی ارائه نشود نمی‌تواند به دقت بالایی برسد.

یکی از جدیدترین زمینه‌ها در حوزه علم یادگیری ماشین، یادگیری ژرف است. این روش مبتنی بر شبکه‌های عصبی مصنوعی است. در این چهارچوب ما استخراج ویژگی‌ها را به صورت سلسه مراتبی انجام می‌دهیم، بدین صورت که فرضاً اگر بخواهیم ویژگی‌هایی را استخراج کنیم که با آن شناسایی چهره را انجام دهیم، شبکه ابتدا ویژگی‌های بسیار پایه‌ای مانند لبه‌ها را استخراج می‌کند. در مرحله بعد ویژگی‌های سطح بالاتر یعنی گوشه‌ها را که حاصل اتصال لبه‌ها هستند را می‌سازد. با این اوصاف می‌بینیم که استخراج ویژگی به صورت سلسله مراتبی است. هرچه به لایه‌های بالاتر برویم ویژگی‌های ما از ترکیب ویژگی‌های سطح پایین‌تر ساخته شده‌اند. این فرآیند استخراج ویژگی خودکار دیگر ما را از وجود یک خبره برای استخراج ویژگی بی‌نیاز می‌کند. علاوه بر این، با توجه به این که استخراج ویژگی به صورت نظام‌مند انجام می‌گیرد، لذا امکان خطاهای انسانی نیز کاهش پیدا می‌کند.

مسئله دیگری که اینجا مطرح می‌شود اندازه متفاوت بسته‌های جاری در شبکه است. با توجه به این که ابعاد ورودی

سامانه استخراج ویژگی ثابت فرض می‌شود (مگر در شبکه‌های خاص مانند شبکه‌های پیچشی[1]) و این نکته که طول بسته‌ها متفاوت است، لذا تعیین ورودی مناسب برای شبکه عصبی هم خود یک چالش بزرگ محسوب می‌شود. برای برطرف کردن این چالش هم روش‌های بسیاری وجود دارد که انتخاب مناسب هر یک از این روش‌ها می‌تواند در استخراج بهتر ویژگی و در نتیجه، دسته‌بندی بهتر ترافیک تاثیرگذار باشد.

در بخش‌های بعدی به معرفی مفاهیم اصلی دسته‌بندی ترافیک شبکه، شبکه‌های عصبی مصنوعی و یادگیری ژرف می‌پردازیم.

۱-۱ دسته‌بندی ترافیک

همان‌طور که از قبل اشاره شد دسته‌بندی ترافیک [۱] شامل نسبت دادن نمونه‌های ترافیک و یا عناصر آن به کاربردها[2] و نوع خاصی از کابردها که آن‌ها را تولید کرده است، می‌باشد. با توجه به عنصری که باید دسته‌بندی شود این کار به سه دسته تقسیم می‌شود: سطح جریان[3]، سطح بسته و سطح میزبان. در سطح اول هدف این است که جریانی که در درون شبکه در حال برقراری ارتباط است را شناسایی کنیم. در حالت کلی جریان را می‌توان تبادل بسته‌ها در یک بازه زمانی معین بین دو میزبان ثابت و از پیش مشخص‌شده، تعریف نمود. در سطح دوم عنصر مورد کنکاش همان بسته‌های ارسالی در شبکه که در واقع کوچکترین واحد ارتباطی بین دو میزبان هستند. در سطح آخر هدف ما شناسایی کاربردهایی است که مورد استفاده یک میزبان خاص قرار می‌گیرد.

از طرفی مهمترین ویژگی‌هایی که می‌توان به دسته‌بندهای ترافیک با توجه به عملی [۱۱] که باید انجام دهند نسبت داد، به شرح زیر است:

دانه‌بندی[4] : ما باید به این نکته توجه کنیم که باید بین الگوریتم‌هایی که توانایی شناسایی خانواده بزرگی از پروتکل‌ها مانند: P2P ، HTTP و Streaming را دارند با الگوریتم‌هایی که توانایی شناسایی الگوریتم‌های خاص مانند: Skype، Torrent و edonky را دارند، تفاوت قائل شویم.

به‌هنگام بودن[5]: الگوریتم‌هایی که قادر هستند پس از دیدن چندین بسته به شناسایی پروتکل مربوطه اقدام ورزند دارای قابلیت دسته‌بندی سریع هستند. این دسته از الگوریتم‌ها در کاربردهایی همچون کنترل کیفیت و امنیت شبکه

1. Convolutional Networks
2. Applications
3. Flow
4. Granularity
5. Timeliness

که نیاز به تصمیم‌گیری سریع دارند مورد استفاده قرار می‌گیرند. از سوی دیگر برخی از الگوریتم‌های مبتنی بر جریان به این نیاز دارند که تعدادی مشخص از بسته‌ها در شبکه جابجا شوند تا بتوانند پروتکل مربوطه را شناسایی کنند. از این الگوریتم‌ها در کاربردهایی همچون نظارت[1] بلندمدت بر شبکه استفاده می‌شود.

هزینه محاسباتی[2]: یکی از مهمترین معیارها در انتخاب نوع الگوریتم دسته‌بندی، پیچیدگی محاسباتی و درصد منابع مصرفی است. پرهزینه‌ترین عملیات را می‌توان دسترسی به حافظه برای بررسی بسته و تطبیق عبارات منظم دانست. رایج‌ترین نوع دسته‌بندی همان دسته‌بندی مبتنی بر جریان است. دلیل این امر این است که تقریبا بیشترین ارتباط با کیفیت سرویس و شکل‌دهی به ترافیک را دارد. شایان ذکر است که جریان طبیعی‌ترین راه ارتباط بین کاربردها هم محسوب می‌شود. از سویی دیگر می‌توان اطلاعاتی با غنای بیشتر را از مجموعه‌ای از بسته‌ها استخراج کرد. این موضوع در مقایسه با یک بسته ساده در شبکه به خوبی محسوس است.

عبارت شناسایی ترافیک[3] دیگر عبارت مشابهی است که در این حوزه استفاده می‌شود. با اینکه عبارت دسته‌بندی ترافیک و شناسایی ترافیک بسیار شبیه به هم هستند ولی شناسایی ترافیک بیشتر وقتی مورد استقاده قرار می‌گیرد که ترافیک را در سطح ریزتری مورد بررسی قرار دهیم. البته در این پایان‌نامه ما این دو عبارت را به جای همدیگر به کار می‌بریم.

با توجه به تعداد کلاس‌های مورد نظر برای تشخیص ترافیک، مسئله دسته‌بندی ترافیک ابعاد مختلفی پیدا می‌کند. برای مثال در دسته‌بندی دودویی هدف این است که بفهمیم عنصر متعلق به کلاس مورد نظر هست یا خیر. از سوی دیگر در مسائل دسته‌بندی چنددسته‌ای[4] یک نمونه باید به یکی از چندین کلاس مختلف نسبت داده شود. آخرین گروه مسائل را می‌توان دسته‌بندی چند برچسبه[5] به شمار آورد. این حالت معمولا وقتی قصد دسته‌بندی میزبان را داریم به وقوع می‌پیوندد. علت این امر هم این است که هر میزبان می‌تواند همزمان از چندین کاربرد استفاده کند. در حالت کلی هم تعداد برچسب‌ها از پیش تعیین‌شده نیست. این سناریو سخت‌ترین حالت از دیدگاه دسته‌بندی است، یعنی حالتی که دسته‌بندی ما چندبرچسبه است و تعداد برچسب‌هایی که به هر کلاس هم می‌توان انتساب داد نامشخص است. در این جا باید فرق بین دسته‌بندی چنددسته‌ای با دسته‌بندی چند برچسبه را مورد توجه قرار دهیم. بدیهی است که در هر دو حالت تعداد کلاس‌ها بیشتر از یک است. ولی در حالتی که ما دسته‌بندی چنددسته‌ای را

1. Monitoring
2. Computational Cost
3. Traffic Identification
4. Multiclass Classification
5. Multilabel Classification

انجام می‌دهیم، در پایان دسته‌بندی به هر نمونه تنها یک برچسب از میان تعداد کلاس‌های موجود تخصیص می‌دهیم. در حالت چند برچسبه ممکن است به هر نمونه بیش از یک کلاس منتسب شود و سختی این کار هم از این‌جا ناشی می‌شود.

۱-۱-۱ چالش‌ها

دسته‌بندی ترافیک به خودی خود امری چالش‌برانگیز و سخت است. دلایل بسیار زیادی برای توجیه سختی این مسئله وجود دارد. مثلا شناسایی برخی از کاربردها بسیار سخت است و این ناشی از استفاده از پروتکل‌های پویا و پیچیده است (video streaming، P2P) دیگر مورد استفاده از تکنیک‌های مبهم سازی ترافیک[1] [۱۲]، تغییر درگاه پویا و استفاده از تونل است که کار شناسایی ترافیک را بسیار سخت می‌کند. علاوه بر این موارد می‌توان از رمزنگاری [۱۳] و یا NAT [2] برای جلوگیری از شناسایی ترافیک استفاده نمود. باید توجه داشت که رمزنگاری کار شناسایی ترافیک را با استفاده از روش‌هایی که بسته را مورد بررسی قرار می‌دهند بسیار سخت می‌کند. علت این امر هم این است که این روش‌ها برای شناسایی یک نوع ترافیک خاص یک الگوی خاصی را مورد جستجو قرار می‌دهند و هر بسته را با الگو تطبیق می‌دهند و اگر بسته رمزشده باشد دیگر امکان تطبیق الگو وجود ندارد. برخی دیگر از کاربردها هم دارای پروتکل‌های اختصاصی هستند که از دسترس عموم خارج است و لذا شناسایی این نوع از ترافیک هم مشکلات خاص خود را دارد. با این اوصاف با پیشرفته شدن تکنیک‌های مبهم‌سازی و همچنین ظهور کاربردهای جدید، تشخیص انواع مختلف ترافیک روز به روز سخت‌تر می‌شود.

لازم به ذکر است که، موارد دیگری هم در سخت‌تر شدن این امر دخیل هستند. از جمله این موارد رعایت کردن حریم خصوصی و همچنین مقابله کردن با مشکل سرعت روزافزون اینترنت و ارتباطات پرسرعت است. باید این مورد را در نظر داشت که در برخی از کشورها استفاده از بررسی‌کننده‌های عمیق بسته[3] برای بررسی بسته را علی‌رغم دقت بالایی که دارند ممنوع اعلام کرده‌اند. علت ممنوع شدن استفاده از این تکنیک که کاملا محتوای بسته را مورد بررسی قرار می‌دهد چیزی نیست جز حفظ حریم خصوصی کاربران که بسته‌های ارسالی آنها در شبکه جاری است. مورد دیگر طراحی دسته‌بندی است که بتواند با سرعت بالای ورود اطلاعات به کار خود ادامه دهد که این امر مستلزم

1. Traffic Obfuscation
2. Network Address Translation
3. Deep Packet Inspection

این است که دسته‌بند قادر به دسته‌بندی نمونه‌های آزمون در زمان بسیار کمی باشد.

با وجود اینکه تحقیقات بسیار زیادی در این زمینه انجام شده است ولی همه این تحقیقات از یک مشکل خاص رنج می‌برند و آن کاربرد دسته‌بند عرضه شده توسط محققان در یک زمینه خاص است. این امر بدین معنی است که مقالات فاقد خاصیت تعمیم[1] هستند. همین موضوع مقایسه نتایج را با سایر کارهای انجام‌شده سخت کرده‌است. به همین منظور در این کار تحقیقی ما سعی کرده‌ایم که دسته‌بندی را طراحی کنیم که تنها مناسب یک تنظیمات از پیش تعیین‌شده خاص نباشد. برای مثال علاوه بر شناسایی ترافیک کاربردهای خاص مانند Skype و Youtube قادر به شناسایی دسته بزرگتری از انواع ترافیک مانند Streaming نیز باشد.

علاوه بر این، بیشتر دسته‌بندهای موجود از مشکلات عدیده‌ای رنج می‌برند که این امر بازبینی در روش‌های موجود را می‌طلبد. مشکلات ذاتی که در این دسته‌بندها وجود دارد، شامل عملکرد غیرقابل قبول در زمینه دقت، حریم خصوصی، نبود قابلیت تعمیم و حتی در مواردی نداشتن امکان به‌کارگیری در سناریوهای واقعی است. با توجه به موارد گفته شده باید انتخاب یک دسته‌بند خوب را یک نوع مصالحه بین این عوامل در نظر گرفت. هرچه ما به سمت دقت بالاتر برویم امکان به مخاطره افتادن حریم خصوصی کاربران بیشتر می‌شود. از سوی دیگر اگر خود را ملزم به رعایت حریم خصوصی کاربر بدانیم باید به جای استفاده از بسته از ویژگی‌های جریان استفاده کنیم. استفاده از این ویژگی‌ها که معمولاً ورودی الگوریتم‌های یادگیری ماشین هستند موضوع دقت و سرعت این الگوریتم‌ها را به‌پیش می‌کشد.

۱-۱-۲ دسته‌بند بهینه

همان‌طور که در قسمت‌های قبل بحث شد، اکثر دسته‌بندها به دلیل این‌که تنها بر روی دسته خاصی از کاربردها آزموده می‌شوند فاقد قابلیت تعمیم هستند. با این حال باید به این سوال پاسخ داد که آیا کاربردهای خاصی وجود دارند که بتوان به عنوان معیار سنجش از آنها برای ارزیابی دسته‌بندها استفاده کرد؟ دلیل این که این کاربردها برای ارزیابی الگوریتم مورد استفاده مناسب هستند چیست؟

در ادبیات کنونی، موضوع دسته‌بندی ترافیک نظیر به نظیر[۱۴] به دلیل سختی‌های خاصی که در شناسایی این نوع خاص از ترافیک وجود دارد، بیشتر مورد توجه قرار می‌گیرد.

1. Generality

حال باید به این سوال پاسخ دهیم که چرا ترافیک نظیر به نظیر می‌تواند به عنوان معیار برای ارزیابی دسته‌بندها مورد استفاده قرار بگیرد. در واقع می‌توان گفت ترافیک نظیر به نظیر شامل همه تکنیک‌های مبهم‌سازی ترافیک، رمزگذاری و استفاده از تونل است. به این نکته هم باید توجه کرد که اغلب الگوریتم‌های رمزنگاری برای ترافیک نظیر به نظیر به صورت آشکارا منتشر نمی‌شوند. به همین دلیل شناسایی ترافیک نظیر به نظیر یکی از مهمترین معیارهای ارزیابی دسته‌بندهای جدیدی هستند که عرضه می‌شوند. برای این که به فهم بهتری از اینکه چرا ترافیک نظیر به نظیر را به سختی می‌توان کشف کرد دست یابیم، ما در بخش بعدی توضیحاتی را در مورد ویژگی‌های ترافیک نظیر به نظیر ارائه می‌دهیم.

۱-۱-۳ اهمیت ترافیک نظیر به نظیر

امروزه استفاده از کاربردهای نظیر به نظیر به امری رایج در اینترنت کنونی تبدیل شده‌است. کاربردهای زیادی مثل اشتراک فایل[1]، VoIP و تلویزیون‌های نظیر به نظیر از جمله موارد مشهور در این زمینه هستند. شناسایی ترافیک نظیر به نظیر به امری بسیار مهم در امنیت و مدیریت شبکه بدل شده‌است. از سویی سهم بزرگی از ترافیک جاری در شبکه‌های امروزی را ترافیک نظیر به نظیر تشکیل می‌دهد. پس مقدار زیادی از منابع ارتباطی برای جابجایی این نوع ترافیک مصرف می‌شود. طراحی شبکه اینترنت کنونی بر مبنای جریان نامتقارن[2] ترافیک طراحی شده‌است. این بدین معنی است که در تخصیص منابع، بنا بر این بوده است که ترافیک دریافتی از ترافیک ارسالی بیشتر باشد. در حالی که ترافیک نظیر به نظیر این امر را نقض می‌کند و به چالشی بزرگ برای فراهم‌کنندگان خدمات اینترنت بدل شده‌است. گرچه استفاده از کاربردهای نظیر به نظیر برای جابجایی فایل کاهش یافته است ولی هنوز کاربردهای نظیر به نظیر بیشترین حجم ترافیک شبکه اینترنت کنونی را شامل می‌شود [۱۵].

از سوی دیگر ترافیک نظیر به نظیر به علت ذات باز بودن[3] و اینکه دارای توزیع‌شدگی خاص خود است، می‌تواند به عنوان یک تهدید بالقوه مطرح شود. علت تهدید بودن هم در این است که این ترافیک می‌تواند به عنوان وسیله‌ای برای جابجایی نرم افزارهای مخرب و یا جابجایی فایل‌هایی که حق مؤلف را نقض می‌کنند، مورد استفاده قرار گیرد. همان‌طور که اشاره شد سامانه‌های نظیر به نظیر را می‌توان با ویژگی بازبودن و توزیع‌شدگی توصیف کرد. کاربردهای نظیر به نظیر از چندین تکنیک مبهم‌سازی ترافیک استفاده می‌کنند [۱۶]. این تکنیک در سطح درگاه[4]، بسته و جریان

1. File Sharing
2. Asymmetric
3. Openness
4. Port

مورد استفاده قرار می‌گیرد. مورد مهم دیگر این است که معمولاً این کاربردها به قدری جدید هستند که هنوز نمی‌توان طبق یک‌سری از قوانین از پیش تعیین‌شده به شناسایی آن‌ها پرداخت. به طور خلاصه تمامی این کارها برای این است که این ترافیک در سامانه‌های تشخیص ترافیک، شناسایی نشوند.

۱-۲ مبانی شبکه‌های عصبی

شبکه‌های عصبی یک روش محاسباتی الهام‌گرفته‌شده از ساختار مغز است، که با هدف حل مسائل مختلف به تقلید از ذهن انسان به وجود آمده‌است. مغز انسان ساختار بسیار پیچیده‌ای دارد و لذا پیاده‌سازی دقیق ساختار آن برای حل مسائل هنوز میسر نیست. شبکه‌های عصبی از واحدهای پایه‌ای به نام نورون ساخته شده‌اند. هر یک از نورون‌ها به وسیله اتصالاتی وزن دار به هم مرتبط شده‌اند. در طی آموزش شبکه، تعداد زیادی از نمونه‌های آموزشی به عنوان ورودی به شبکه داده می‌شود. الگوریتمی که برای یادگیری از آن استفاده می‌شود پس‌انتشار[1] [۱۷] نام دارد. این الگوریتم هر بار با دیدن نمونه‌های جدید ورودی سعی می‌کند که وزن‌های شبکه را به گونه‌ای تنظیم کند که خروجی شبکه همان خروجی مورد نظر گردد. بدیهی است که هر چه دادگان آموزشی بیشتر باشد، شبکه خروجی بهتری خواهد داشت. لذا همواره توصیه می‌شود که در صورت داشتن دادگان آموزشی کافی از رویکرد شبکه‌ی عصبی استفاده شود. ساده‌ترین نوع شبکه‌های عصبی شبکه‌های جلوسو[2]، هستند. این نوع شبکه‌های عصبی عموماً برای دسته‌بندی مورد استفاده قرار می‌گیرند.

در شکل ۱-۲ یک شبکه جلوسو نشان داده‌شده‌است. همان‌طور که در تصویر مشخص است، شبکه از تعدادی واحد پردازشی به نام نورون در چندین لایه مختلف ساخته‌شده‌است. نورون‌های موجود در هر لایه با هم هیچ گونه ارتباطی ندارند. به هریک از این نورون‌ها گره هم گفته می‌شود. به منظور حل مسئله توسط مدل شبکه عصبی، نیاز است تا وزن‌های مناسب برای هر لایه تعیین شوند. پس از تعیین پارامترهای مدل مانند نرخ یادگیری، تعداد لایه‌های پنهان، تعداد نورون‌های هر لایه، تابع فعال‌سازی و نظایر این‌ها، وزن‌های بهینه مرتبط با هر لایه به کمک دادگان آموزش محاسبه می‌شوند. در نهایت شبکه با دادگان آزمون مورد ارزیابی قرار می‌گیرد.

هدف ما در اینجا تمرکز بر روی الگوریتم‌های مورد استفاده در شبکه‌های عصبی نیست، بلکه بیشتر بر کاربرد این شبکه‌ها متمرکز می‌شویم.

1. Back Propagation
2. Feed Forward Networks