شبکه‌های گرافی توجه: گزارش فاز پیاده‌سازی

*محمد ابراهیمی، پارسا عباسی*

*دانشگاه علم و صنعت ایران، تهران* [*m\_ebrahimi74@comp.iust.ac.ir*](mailto:m_ebrahimi74@comp.iust.ac.ir)*,* [*parsa\_abbasi@comp.iust.ac.ir*](mailto:parsa_abbasi@comp.iust.ac.ir)

**1) لایۀ GAT**

**2) مدل GAT**

**3) مجموعه داده**

در مقالۀ اصلی GAT چهار مجموعه داده مورد آزمایش قرار گرفته است که دو مورد از آن به عنوان نمونه‌هایی از رویکرد یادگیری Transductive و Inductive جهت بررسی صحت عملکرد مدل در این پیاده‌سازی نیز مورد ارزیابی قرار گرفته‌اند. مشخصات این دو مجموعه داده به شرح زیر است:

**3-1) CORA**

مجموعه داده CORA شامل مقالات مربوط به یادگیری ماشین است. این مقالات در یکی از هفت کلاس زیر طبقه بندی می‌شوند:

Case Based, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning, Theory

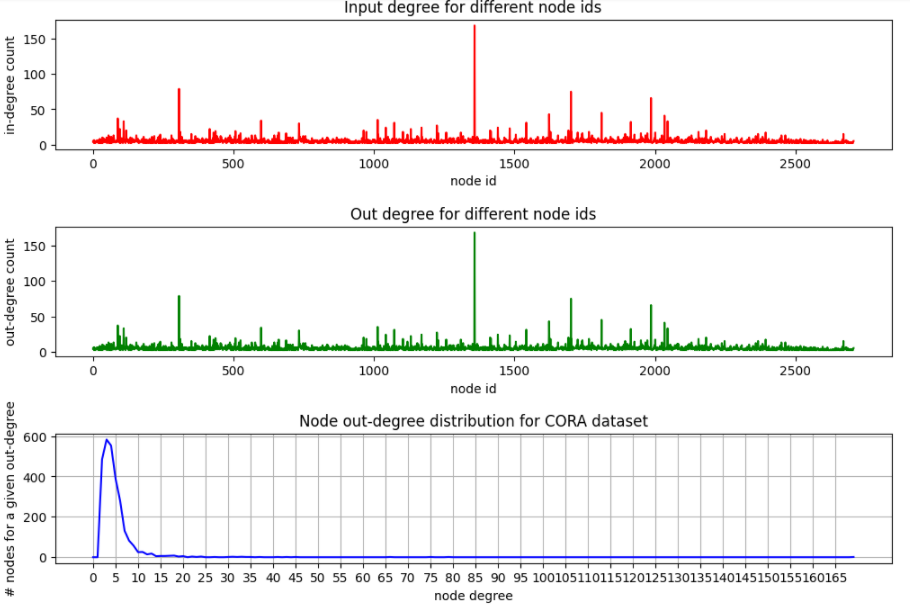
در این مجموعه داده ، گره ها به اسناد یا مقالات و یال‌ها به سایتیشن یا استناد‌ (بدون جهت) مربوط هستند. ویژگی‌های گره مطابق با اجزای نمایش bag-of-words یک مقاله هستند. پس از تبدیل کلمات به ریشه‌ی آن‌ها (stemming)و حذف کلمات توقف با یک مجموعه‌ی واژگان شامل 1433 کلمه منحصر بفرد مواجه هستیم. مقالات به گونه‌ای انتخاب شده‌اند که در پیکره نهایی هر مقاله حداقل یک مقاله دیگر به آن استناد کند یا از آن استناد شود. در کل پیکره 2708 مقاله وجود دارد.

هر گره دارای یک برچسب کلاس است. ما اجازه می‌دهیم فقط 20 گره در هر کلاس برای آموزش استفاده شود - با این وجود ، طبق به تنظیمات Transductive ، الگوریتم آموزش به بردارهای ویژگی همه گره‌ها دسترسی دارد.

آمار کلی این مجموعه داده مطابق با جدول زیر است:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **تعداد راس‌ها** | **تعداد یال‌ها** | **اندازۀ ویژگی هر راس** | **نمونه‌های آموزشی** | **نمونه‌های اعتبارسنجی** | **نمونه‌های آزمون** |
| 2708 | 5429 | 1433 | 140 | 500 | 1000 |

نمودارهای میزان درجه ورودی هر راس، درجه خروجی هر راس و توزیع درجات خروجی در زیر نمایش داده شده است. همانطور که مشاهده می‌شود، دو نمودار اول مشابه با همدیگر بوده و این به دلیل بدون جهت بودن گراف است. در نمودار سوم نیز به خوبی قابل مشاهده است که رئوسی با درجات بسیار بالا، نادر هستند و پراکندگی در تعداد یال‌های کمتر، بیشتر است.



**3-2) PPI**

علاوه بر این، مدل پیاده‌سازی شده بر روی دو مجموعه داده جدید نیز مورد آزمایش قرار گرفته است که جزئیات هرکدام از آن‌ها به شرح زیر است:

**3-3) Reddit**

شبکه اجتماعی Reddit یکی از بزرگترین انجمن‌های اینترنتی است که کاربران به انتشار محتوا و نظرات خود پیرامون بسیاری از موضوعات در آن می‌پردازند. در مسئله‌ای که برای این شبکه مطرح می‌شود سعی بر این است تا پیش‌پینی کنیم که هر پست Reddit به کدام جامعه (موضوع) تعلق دارد.

در سال 2014 یک مجموعه داده گرافی از پست‌های این شبکه ساخته شد. برچسب‌های هر راس در این گراف، جامعه یا subreddit ای است که پست بدان تعلق دارد. گراف ساخته شده به شکل پست-به-پست بوده و دو پست در صورتی با یک یال به یکدیگر متصل هستند که کاربر مشابهی بر روی هر دوی آن‌ها دیدگاه خود را ثبت کرده باشد.

در مجموع این مجموعه داده شامل 232965 پست با میانگین درجه 492 است که در مقایسه با سایر مجموعه داده‌های گرافی میزان درجه قابل توجهی به شمار می‌آید. داده‌های مربوط به 20 روز نخست به عنوان مجموعۀ آموزشی و سایر روزها به عنوان آزمایش (30% آن به عنوان اعتبارسنجی) در نظر گرفته شده است.

بردار ویژگی هر راس مطابق با بردارهای کلمات 300 بعدی GloVe CommonCrawl ساخته شده است. به ازای هر پست، ویژگی‌های زیر با یکدیگر الحاق شده‌اند و بردار نهایی را تشکیل داده‌اند:

* میانگین تعبیۀ عنوان پست
* میانگین تعبیۀ تمام دیدگاه‌های پست
* امتیاز پست
* تعداد دیدگاه‌های پست

آمار کلی این مجموعه داده مطابق با جدول زیر است:

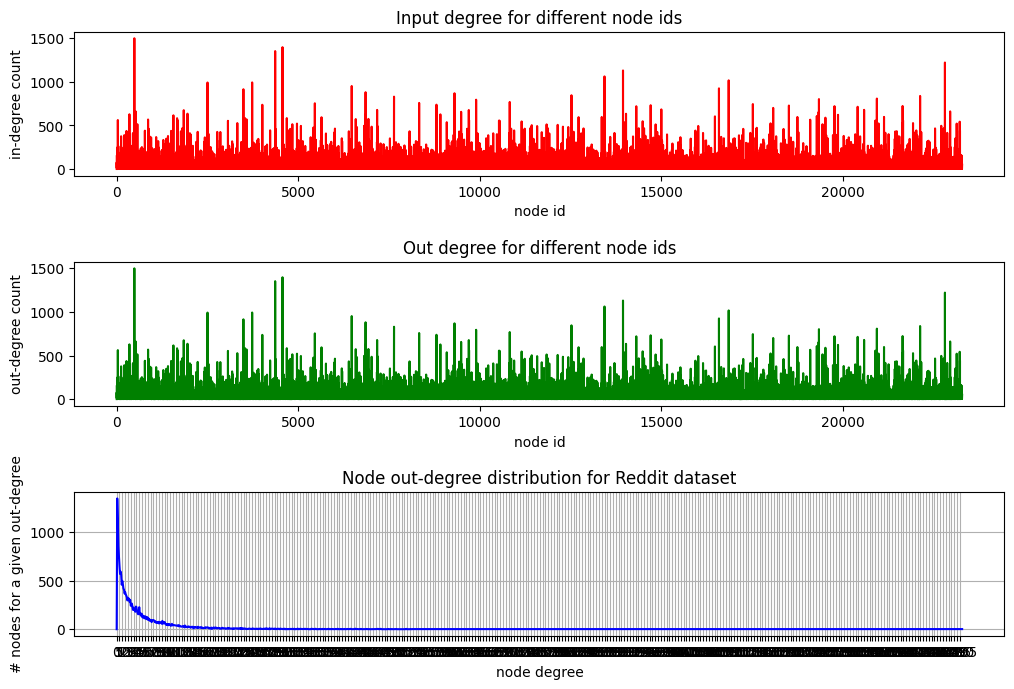
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **تعداد راس‌ها** | **تعداد یال‌ها** | **اندازۀ ویژگی هر راس** | **نمونه‌های آموزشی** | **نمونه‌های اعتبارسنجی** | **نمونه‌های آزمون** |
| 232965 | 114615892 | 602 | 153431 | 23831 | 55703 |

با این حال، به دلیل اندازۀ بسیار بزرگ مجموعۀ داده و عدم دسترسی به سخت‌افزار کافی برای آموزش مدل بر روی این حجم از داده، در آزمایش‌های صورت گرفته تنها به ده درصد آن اکتفا شده و آمار آن به شرح موجود در جدول زیر است:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **تعداد راس‌ها** | **تعداد یال‌ها** | **اندازۀ ویژگی هر راس** | **نمونه‌های آموزشی** | **نمونه‌های اعتبارسنجی** | **نمونه‌های آزمون** |
| 23296 | 1052560 | 602 | 15343 | 2383 | 5570 |

باید توجه داشت که در برخی مقالات پیشین که بر روی این مجموعه داده نیز آزمایش خود را انجام داده‌اند از رویکرد یادگیری Inductive استفاده شده است. این بدین معنی است که در طول فرآیند آموزش مدل، رئوس مربوط به نمونه‌های آزمایشی به کلی نادیده گرفته می‌شوند. با این حال از آنجا که در این آزمایش مجموعه داده کوچک شده، آموزش مدل به شکل Transductive و مشابه با آنچه که بر روی مجموعه CORA اتفاق افتاد، انجام خواهد شد.

نمودارهای میزان درجه ورودی هر راس، درجه خروجی هر راس و توزیع درجات خروجی در زیر نمایش داده شده است. همانطور که مشاهده می‌شود، دو نمودار اول مشابه با همدیگر بوده و این به دلیل بدون جهت بودن گراف است. در نمودار سوم نیز به خوبی قابل مشاهده است که رئوسی با درجات بسیار بالا، نادر هستند و پراکندگی در تعداد یال‌های کمتر، بیشتر است.



**3-4) Wiki\_CS**

مجموعه داده‌ Wiki-CS یک مجموعه داده‌ مبتنی بر ویکی‌پدیا برای آزمودن (benchmarking) شبکه های عصبی گرافی است. این مجموعه داده از دسته‌بندی‌های ویکی‌پدیا ، به طور خاص 10 کلاس مربوط به شاخه های علوم کامپیوتر ، با اتصال بسیار بالا ساخته شده است. ویژگی‌های گره از متن مقاله‌های مربوطه گرفته شده است. این ویژگی‌ها به عنوان میانگین تعبیه کلمه از پیش آموزش دیده GloVe محاسبه شدند. در نتیجه ویژگی‌های گره 300 بعدی حاصل می‌شود که می‌تواند برای آموزش مدل‌های بزرگ روی GPU یک مزیت باشد. مقالاتی که می‌توانستند به چند کلاس تعلق داشته باشند از این مجموعه داده حذف شده‌اند.

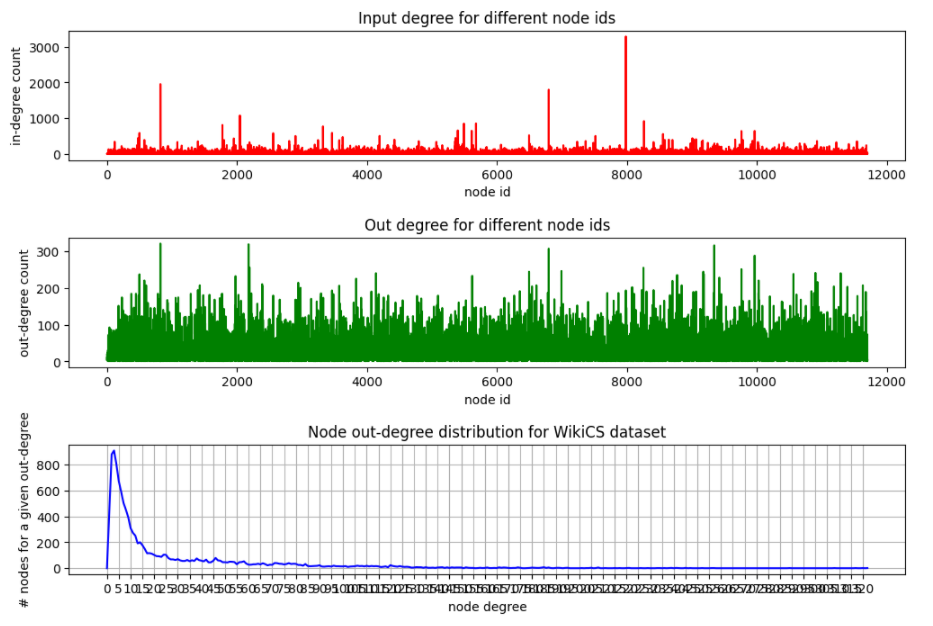
گره‌ها را در هر کلاس به دو مجموعه تقسیم می‌شوند، 50٪ برای مجموعه آزمون و 50٪ به طور بالقوه قابل مشاهده. از مجموعه قابل مشاهده ، 20 تقسیم مختلف برای آموزش ، اعتبارسنجی و توقف زودرس ایجاد شده‌است: در هر تقسیم 5٪ از گره‌های هر کلاس برای آموزش ، 22.5٪ برای ارزیابی معیار توقف زودرس مورد استفاده قرار گرفت و 22.5٪ به عنوان مجموعه اعتبارسنجی برای تنظیم هایپر پارامتر استفاده شده‌است.

ما در آزمایشات خود فقط از یکی از 20 تقسیم مختلف برای برای آموزش و اعتبارسنجی استفاده کرده و از داده‌های مربوط به توقف زودرس استفاده نمی‌کنیم.

آمار کلی این مجموعه داده مطابق با جدول زیر است:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **تعداد راس‌ها** | **تعداد یال‌ها** | **اندازۀ ویژگی هر راس** | **نمونه‌های آموزشی** | **نمونه‌های اعتبارسنجی** | **نمونه‌های آزمون** |
| 11701 | 216123 | 300 | 580 | 1769 | 5847 |

نمودارهای میزان درجه ورودی هر راس، درجه خروجی هر راس و توزیع درجات خروجی در زیر نمایش داده شده است. همانطور که مشاهده می‌شود، نمودار اول و دوم متفاوت هستند و این به دلیل جهتدار بودن گراف مربوطه است. در نمودار سوم نیز به خوبی قابل مشاهده است که رئوسی با درجات بسیار بالا، نادر هستند و پراکندگی در تعداد یال‌های کمتر، بیشتر است.



**4) نتایج**

در مقالۀ