بسم الله الرحمن الرحیم

نام مقاله: Proximal Policy Optimization Algorithms

نویسندگان : John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov (OpenAI)

ژورنال و ناشر: arXiv Preprint (OpenAI) / CoRR

سال چاپ : ۲۰۱۷

فیلد

مرجع کامل	J. Schulman et al., "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347, 2017
مسئله	واگرایی و **به‌روزرسانی های مخرب** در روش‌های سنتی گرادیان سیاست و پیچیدگی بالای پیاده سازی روش های مقید مانند TRPO
روش	**PPO** با استفاده از تابع هدف **Clipped Surrogate** جهت محدود کردن تغییرات سیاست در یک بازه امن
ورودی (INPUT)	بردارهای وضعیت (States) مستخرج از تله متری (یا GNN)، پاداش‌ها (Rewards) و تخمین های مزیت (Advantages)
خروجی	سیاست بهینه (π) و تابع ارزش (V) برای اتخاذ کنش های مدیریت ترافیک
فرضیات کلیدی مقاله	مقید کردن نسبت احتمال سیاست (r_t) یک حد پایین بدبینانه (Pessimistic Bound) برای بهبود عملکرد فراهم می‌کند
نوآوری	ابداع مکانیزم **Clipping** برای جایگزینی قیدهای سخت ریاضی با یک تابع هدف ساده و کارآمد
محدودیت	حساسیت به انتخاب هایپرپارامتر ϵ و ضریب بونوس انتروپی در سناریوهای بسیار پیچیده
پیچیدگی محاسباتی و زمان اجرا	بهینه سازی مرتبه اول (First-order) با پیچیدگی بسیار کمتر از TRPO و کارایی نمونه برداری بالا
معیارهای ارزیابی و بسترهای شبیه‌سازی	میانگین پاداش تجمعی در محیط‌های MuJoCo و بازی های Atari
ربط به نورو-باتلنک	موتور تصمیم‌گیر در ***xApp***؛ تضمین پایداری شبکه همراه اول با جلوگیری از تغییرات ناگهانی و مخرب در مسیرها
فصل پایان‌نامه	فصل ۲ (مبانی DRL) و فصل ۳ (طراحی عامل هوشمند مقید)

ایده قابل استفاده (در کدنویسی یا فرمول بندی)	فرمول بندی تابع زیان ترکیبی ($L^{CLIP+VF+S}$) و استفاده از **GAE** برای کاهش واریانس در شبکه اپراتوری

❖ **خلاصه برداشت شده از سند** :

این سند، معرف الگوریتم **PPO** به عنوان یکی از پایدارترین و کارآمدترین روش های یادگیری تقویتی عمیق است که هدف اصلی آن حل چالش "به‌روزرسانی‌های افراطی سیاست" می‌باشد. در محیط های پویایی مانند شبکه همراه اول، اعمال یک سیاست مسیریابی جدید که تفاوت فاحشی با وضعیت قبلی داشته باشد، می‌تواند منجر به سقوط ناگهانی کیفیت سرویس (**Performance Collapse**) شود. PPO با معرفی مکانیزم **Clipped Surrogate Objective**، نسبت تغییرات سیاست جدید به قدیم را در یک بازه امن (معمولاً بین ۰.۸ تا ۱.۲) مقید می‌کند.

در چارچوب **نوروباتل نک**، این الگوریتم به عنوان مغز متفکر لایه تصمیم‌گیر عمل می‌کند. این سند تبیین می‌کند که چگونه می‌توان با استفاده از **معماری مشترک** (**Shared Anatomy**)، ویژگی‌های گرافی استخراج شده توسط **GraphSAGE** را همزمان برای ارزیابی وضعیت (**Critic**) و صدور فرامین اصلاحی (**Actor**) به کار گرفت. همچنین، وجود **بونوس انتروپی** در فرمول بندی این مقاله، تضمین‌کننده این است که عامل هوشمند ما در مواجهه با شرایط جدید ترافیکی (مانند Flash Crowd)، از کاوش برای یافتن مسیرهای جایگزین باز نماند.