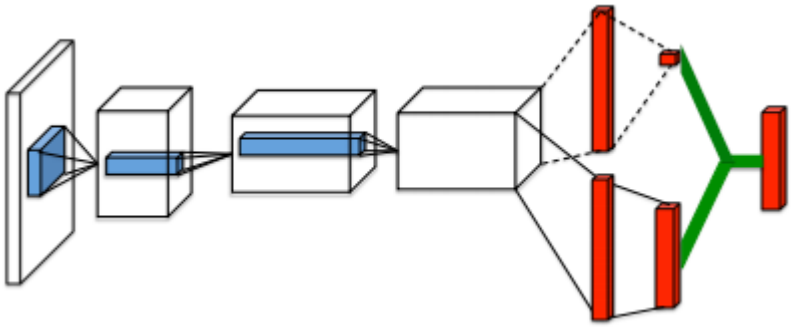
در این مقاله یک معماری جدید به اسم dueling ارائه شده است که به صورت صریح بازنمایی state values و state-dependent action advantage را از طریق دو استریم (stream) از هم جدا کرده است.
انگیزه اصلی پشت این معماری این است که در بعضی بازی ها، دانستن ارزش هر عمل یا اکشن در هر timestep ضروری نیست. برای مثال در atari game enduro نیازی نیست بدانیم چه عملی انجام دهیم تا زمانی که برخورد به اتومبیل ها قریب‌الوقوع شود.
این معماری بیشتر به task هایی مربوط است که اکشن ها همیشه اثر معنی‌داری در محیط ندارند.

معماری: مانند معماری DQN است که از لایه‌های کانولوشنی برای پردازش فریم‌های بازی استفاده می‌کند. از آنجا به بعد شبکه به دو استریم جدا از هم تقسیم می‌شود که قبلا به آن‌ها اشاره کردیم. بعد از این دو استریم آخرین ماژول برای ترکیب state-value و خروجی advantage است. معماری گفته شده به شکل زیر است. خروجی شبکه مجموعه ای از Q value ها میباشد که هر کدام برای یک اکشن است.



$V(s;\theta,\beta)$: همان state-value است که مقدار یکی از استریم‌ها را تشکیل می‌دهد.

$A(S,a;\theta,\alpha)$: همان advantage است که دومین استریم را تشکیل می‌دهد. این مقدار به صورت زیر به دست می‌آید:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s).$$

در فرمول بالا Q value ارزش یک اکشن مشخص در state داده شده را می‌دهد و V value ارزش state داده شده را بدون در نظر گرفتن عمل انجام شده می‌دهد. بنابراین به طور شهودی advantage value نشان می‌دهد که انتخاب یک عمل نسبت به بقیه، در حالت داده شده چقدر سودمند است.

حالا برای ترکیب یا تجمیع این دو استریم باید چیکار کرد؟
با توجه به تعریف advantage value، می‌توان این دو را به صورت زیر جمع زد تا خروجی شبکه به دست آید:

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + A(s,a;\theta,\alpha),$$

اما این کار مسئله ایجاد میکند چون جمع این دو مقدار غیر قابل شناسایی است و با گرفتن Q value نمی‌توان مقادیر A و V را منحصر به فرد بازیابی کرد که منجر به عملکرد ضعیف شبکه می‌شود. بنابراین آخرین ماژول شبکه، forward mapping را پیاده‌سازی می‌کند (دلیل استفاده از رابطه 9)

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \max_{a' \in |\mathcal{A}|} A(s,a';\theta,\alpha)\right).$$

این رابطه function estimator advantage را وادار می‌کند تا advantage در عمل انتخاب شده برابر 0 شود.

البته میتوان به جای عملگر max از average استفاده کرد که معادله آن به صورت زیر است:

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \frac{1}{|\mathcal{A}|}\sum_{a'} A(s,a';\theta,\alpha)\right). \quad (9)$$