دانشکده مهندسی کامپیوتر و فناوری اطلاعات

کنترل عصبی برای عملیات پیگردی با استفاده از یادگیری تقویتی Attention-Gated در BMI

Neural Control of Tracking Task via Attention-Gated Reinforcement Learning for Brain-Machine Interface

Yiwen Wang, ***Member, IEEE, Fang Wang, Student Member, IEEE, Kai Xu, Qiaosheng Zhang,***

Member, IEEE, Shaomin Zhang, Xiaoxiang Zheng, Member, IEEE

درس: یادگیری تقویتی

تهیه‌کننده: سیدسینا سجادپور

استاد درس: دکتر منصور فاتح

تاریخ ارائه:

۹۵/۱/۲۹

مقدمه

در ابتدای این گزارش توصیفی از BMI خواهیم داشت. BMI مخفف Brain Machine Interface است که با نام‌های BCI(Brain machine Inteface)، BMI(Brain Machine Interface)، MMI(Mind Machine Interface) و DNI (Direct Neural Interface) نیز شناخته شده است. هدف این سیستم‌ها ترجمه سیگنال‌های مغز به فرمان‌های عملی در سیستم‌های کامپیوتری است، که نتیجه آن استفاده از فرمان‌ها برای کنترل یک برنامه کامپیوتری و یا یک وسیله مکانیکی از جمله هدایت مکان‌نما، حرکت دادن دست رباتیکی و یا هدایت ویلچر می‌باشد. در دهه اخیر با پیاده سازی سیستم‌های BMI بر روی حیوانات از جمله موش‌ها و میمون‌ها و همچنین انسان‌ها برای کنترل ابزارهای مکانیکی و کامپیوتری میزان موفقیت این سیستم‌ها نشان داده شده‌اند.

پروسه عملیاتی سیستم‌های BMI به این گونه است که ابتدا سیستم سیگنال‌های تولید شده توسط مغز را دریافت کرده و بر روی آنها عملیات‌های پیش پردازش از جمله رفع نویز و نرمالسازی سیگنال‌ها را انجام می‌دهد. و پس از آن سیگنال‌ها وارد شبکه عصبی شده و طبقه بندی می‌شوند که هر طبقه مربوط به یک فرمان مغز می‌باشد.

این سیستم‌ها به راحتی قابل بهره‌برداری نیستند. چه بسا در اغلب موارد بیماران معلول قادر به تولید وکنترل برخی سیگنال-های مغزی در غشای حرکتی برای آموزش برنامه نیستند (به علت عدم وجود عضو اصلی). در چنین مواردی کاربران BMI نیازمند بازخوردهای بیولوژیکی همچون بازخوردهای دیداری (مشاهده نتایج خروجی و سعی بر تنظیم کردن سیگنال‌های تولید شده با آزمون و خطا) می‌باشند.

از طرفی دیگر، مطالعاتی در زمینه بهبود کارایی سیستم‌های BMI توسط تقویت دیکدرهای استفاده شده در این سیستم‌ها انجام شده است. هدف این مطالعات افزایش دقت و یادگیری دیکدرها برای عواملی چون رفع نویز و معیارهای دسته‌بندی بوده است.

هر دو روش ذکر شده می‌توانند بر بهبود عملکرد سیستم‌های BMI موثر واقع شوند. در واقع روشی می‌تواند نتیجه بهتری داشته باشد که هر دو روش فوق را پوشش دهد. یادگیری تقویتی یکی از روش‌هایی است که هر دو روش در آن حضور دارند. مطالعات زیادی در زمینه استفاده از یادگیری تقویتی برای دیکدرهای BMI صورت گرفته است، از جمله استفاده از Errp استخراج شده از EEG به عنوان سیگنال پاداش به منظور کاهش actionهای اشتباه [۱]، استفاده از متدهای اختلاف زمانی برای آموزش طبقه بندی کننده‌های عصبی بطور on-line در BCIهای EEG-Based [۲] و طراحی سیستم RL-Based BMI که با استفاده از Q(λ)-learning در آن موش ها برای کنترل دست مصنوعی آموزش داده شدند [۳].

مقایسه RL و SL در سیستم‌های BMI

Supervied learning

این روش یادگیری که بطور گسترده‌ای در شبکه‌های عصبی مصنوعی کاربرد دارند، برای مسائل neurobiological (مطالعه سلول‌های سیستم عصبی) حداقل به دو علت مناسب نیستند:

۱) در supervised learning در هر بار آموزش، سیگنال خطای لایه خروجی باید به لایه ورودی انتشار کند. این سیگنال‌های خطا در مسائل neurobiological قابل مشاهده نمی‌باشند.

۲) Supervised learning نیاز به یک " آموزش دهنده" دارد تا در طول آموزش الگوهای درست را در لایه خروجی تعیین کند. در چنین مسائلی تعیین تمام فعالیت های عصبی همچون قشر حرکتی مغز کار پیچیده ای است.

Reinforcement Learning

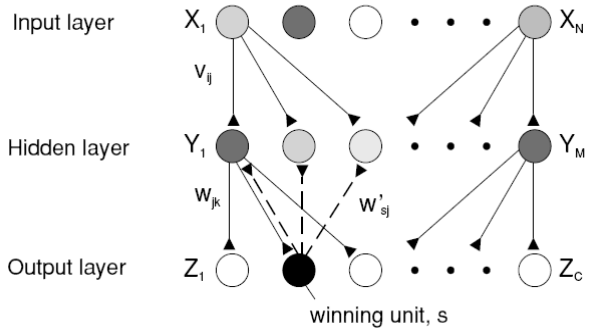
این روش یادگیری یکی از مدل های پرطرفدار در مسائل neurobiological می باشد. در مسائل neurobiological با استفاده از RL خروجی ها بطور تصادفی (stochastic) انتخاب می شوند. بنابراین شبکه عصبی میتواند خروجی های متنوعی را به ازای هر داده ورودی انتخاب کند(همانند عملکرد یادگیری انسان و حیوان). مدل های biological RL بر اساس میزان پاداش دریافتی در یک آزمایش، که میتواند بهتر یا بدتر از میزان پاداش پیش بینی شده باشد، تغییر رفتار دهد (action متفاوت انتخاب کند).

با تمام این اوصاف مدل های biological RL که از قوانین یادگیری biologic استفاده می کنند، به کارآمدی مدل های supervised learning در بهینه tune کردن نرون های لایه مخفی در شبکه عصبی نیستند. علت این امر عدم وجود مکانیزم بهینه برای ارزش دهی به نرون هایی که نقش مهمی در stimulus-response دارند.

معرفی Attention-Gated Reinforcement Learning (AGREL)

همانطور که گفته شد مدل های biological RL به کارآمدی مدل های supervised learning در tune کردن وزن‌های نرون های لایه مخفی در شبکه عصبی نیستند. در مدل AGREL این مشکل در مسائل neurobiological با اضافه کردن یک سیگنال attentional حل شده است. سیگنال attentional در واقع Feedback برای پردازش های شبکه برای انتخاب state می‌باشد. در غشای میانی مغز نرون‌های دوپامین وجود دارند که اطلاعات مربوط به موفقیت انسان(پاداش) در اعمال خود را حمل می‌کند. سیگنال attentional حاصل واکنشات این نرون ها می‌باشد.

AGREL یک شبکه عصبی سه لایه(شکل ۱) برای انتخاب رفتار یک حیوان یا انسان است که stimuli (سیگنال‌هایی که از مغز در حین انجام فعالیت خاص دریافت می‌شوند) ها را دسته بندی می‌کند. فرض می کنیم که P تا stimuli داریم که می‌خواهیم آنها را توسط شبکه عصبی به C کلاس مجزا و بدون اشتراک برای فرمان‌های مشخصی تقسیم بندی کنیم. در واقع ورودی‌های شبکه stimuli ها و خروجی‌های شبکه فرمان‌های مربوط به stimuli ها می‌باشند.



شکل ۱: تصویر شبکه AGREL

همانطور که در شکل ۱ مشاهده می‌شود، این شبکه همانند شبکه MLP دارای سه لایه ورودی، میانی و خروجی است. بین لایه ورودی و لایه میانی وزن‌های V_{ij} وجود دارند که مقادیر آنها در هر بار آموزش شبکه update می‌شوند. بین لایه میانی و لایه خروجی وزن‌های W_{ij} وجود دارند که مقادیر آنها نیز در طی آموزش update می‌شوند. تمام خروجی‌های این شبکه مقدار صفر دارند، بجز یک خروجی که به عنوان خروجی پیروز مقدار یک را اختیار می‌کند.

پس از وارد شدن یک ورودی جدید به شبکه و انتشار آن به لایه خروجی برای هر خروجی یک مقدار احتمالی برای پیروز شدن تعیین می‌شود. این مقدار احتمالی با استفاده از رابطه زیر محاسبه می‌شود:

$$\Pr\left(Z_k^p = 1\right) = \frac{\exp\left(a_k^p\right)}{\sum_{k'=1}^{C} \exp\left(a_{k'}^p\right)} \quad \text{with} \quad a_k^p = \sum_{j=0}^{M} w_{jk} Y_j^p. \qquad (۱)$$

پس از محاسبه مقدار Pr(Z) برای تمام خروجی‌های شبکه، بیشترین مقدار Pr مربوط به خروجی برنده است. در صورت انتخاب صحیح شبکه برای action مناسب برای stimuli مشخص، پاداش r دریافت میشود. در صورت انتخاب صحیح r=1 و در صورت عدم انتخاب صحیح هیچ گونه پاداشی دریافت نمی‌شود.

در هر مرحله آموزش، وزن های شبکه در لایه میانی و لایه خروجی طبق قاعده Hebbian، update می‌شوند. برای update این وزن‌ها از رابطه زیر استفاده می‌شود:

$$\Delta w_{jk} = \beta Y_j^p Z_k^p f(\delta). \qquad (۲)$$

این قاعده تنها برای عنصر خروجی برنده اعمال می‌شود. به این دلیل که مابقی Z_k ها مقدار صفر اختیار می‌کنند و حاصل دلتا نیز صفر خواهد بود. مقدار f(δ) توسط روابط ۳ و ۴ محاسبه می‌شود.

$$\delta = r - E^p(r). \qquad (۳)$$

مقدار r میزان پاداش دریافت شده است که برابر با یک یا صفر است. مقدار $E^p(r)$ برابر میزان پاداش مورد انتظار برای ورودی p است که همان مقدار احتمالی Pr(Z) محاسبه شده برای هر خروجی می‌باشد.

$$f(\delta) = \begin{cases} \delta/(1-\delta); & \delta \geq 0 \\ \delta; & \delta = -1 \end{cases}. \qquad (۴)$$

در مرحله آخر وزن های Vij که بین لایه ورودی و لایه میانی است نیز با استفاده از قاعده Hebbian، update می شوند.

$$\Delta v_{ij} = \beta X_i^p Y_j^p f(\delta) f b_{Y_j}^p \quad \text{with} \quad f b_{Y_j}^p = \left(1 - Y_j^p\right) \sum_{k=1}^{C} Z_k^p w'_{kj}. \qquad (۵)$$

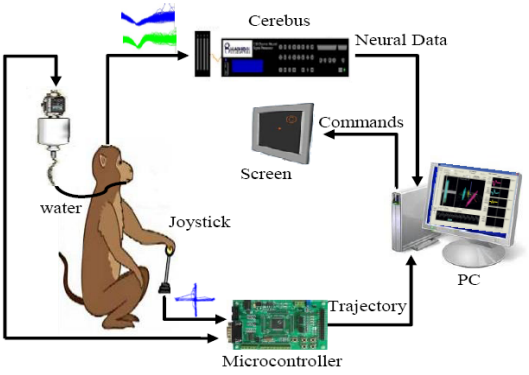
مقایسه AGREL با Q-Learning

در مقاله مورد مطالعه دو روش AGREL و Q-Learning بر میزان موفقیت سیستم BMI بررسی شده‌اند.

جمع آوری داده

در این آزمایش به یک میمون رزوس (rhesus) آموزش داده شد تا بتواند با یک دسته joystick بازی center-out task را انجام دهد. بازی به این گونه است که ابتدا یک هدف در مرکز تصویر ظاهر می‌شود. میمون ابتدا باید مکان‌نما را به مرکز تصویر هدایت کند. وقتی مکان‌نما روی هدف ثابت ماند، هدف در یکی از چهار جهت تصویر به صورت تصادفی ظاهر می‌شود. در صورتی که میمون مکان‌نما را به سمت هدف هدایت کند و روی آن نگه دارد، مقداری آب از طریق شلنگی که در دهان میمون گذاشته شده است فرستاده می‌شود. پس از دریافت پاداش مجدداً هدف در مرکز ظاهر می‌شود و بازی با همین روند ادامه دارد.

الکترودی در سر میمون کار گذاشته شده است تا سیگنال‌های مغز را دریافت کند و به سیستم ارسال کند. سیگنال‌های دریافتی پیش پردازش شده و وارد سیستم BMI می‌شوند.



شکل ۲: شمای اتصالات مربوط به میمون و سیستم BMI و پاداش

درحالی که cursor روی صفحه نمایش توسط حرکت دست میمون کنترل میشود، اطلاعات دریافتی از واکنش های عصبی بصورت online توسط سیستم دریافت، پیش پردازش و decode میشوند.

داده های دریافت شده توسط سیستم به ۷ دسته (action) تقسیم بندی میشوند.

۱) up

۲) down

۳) left

۴) right

۵) ثابت نگه داشتن جهت حرکت بر روی محور yها (holding1)

۶) ثابت نگه داشتن جهت حرکت بر روی محور xها (holding2)

۷) استراحت یا قرار دادن joystick در حالت بی حرکت (resting)

آموزش توسط AGREL و Q-Learning

در این مقاله از مدل AGREL برای پیش بینی تصمیم میمون جهت حرکت cursor استفاده شده است. سپس نتایج بدست آمده از این مدل را با مدل Q-learning و مقایسه شده اند. برای مقایسه روش AGREL با روش Q-learning از دو سیاست برای انتخاب actionبرنده در Q-Learning استفاده شده است:

۱) سیاست حریصانه (Q-greedy)

۲) سیاست stochastic softmax (Q-softmax)

انتخاب حریصانه (Q-greedy) و انتخاب Q-softmax به ترتیب طبق روابط (۶) و (۷) محاسبه می‌شوند:

$$\arg\max_{a_k \in A(s)} Q(s_t), \quad \text{where} \quad Q_k(s_t, a_t) = \frac{1}{1+\exp\left(-\sum_{j=0}^{M-1} w_{jk} Y_j\right)} \tag{۶}$$

$$Q_t(Z_k = 1) = \frac{\exp\left(\left(\sum_{j=0}^{M-1} w_{jk} Y_j\right)/\tau\right)}{\sum_{k'=1}^{C} \exp\left(\left(\sum_{j=0}^{M-1} w_{jk'} Y_j\right)/\tau\right)} \tag{۷}$$

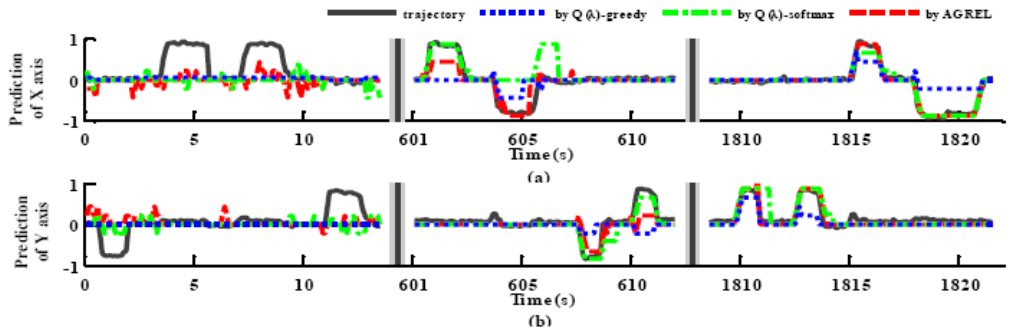
در هر سه روش از داده های ورودی یکسان، شبکه عصبی یکسان و خروجی های یکسان استفاده شده است اما سیاست متفاوت برای انتخاب action اتخاذ شده است.

نتایج

نمودار ۱ مربوط به پیش بینی‌های سیستم BMI با استفاده از سه روش AGREL، Q-greedy و Q-Softmax آورده شده است. نمودار بالا مربوط به پیش بینی‌ها برای انتخاب‌ها در جهت افقی(محور X) و نمودار پایین مربوط به پیش بینی‌ها برای انتخاب‌ها در جهت عمودی(محور Y) است. خط سیاه مربوط به انتخاب انجام شده، خط آبی مربوط به پیش بینی با استفاده از روش Q-greedy، خط سبز مربوط به پیش بینی با استفاده از روش Q-softmax و خط قرمز مربوط به پیش بینی با استفاده از روش AGREL می-باشد.

همانطور که دیده می‌شود در ۱۵ بار اول هر سه روش انتخاب‌های نامناسبی را انجام می‌دهند. در این قسمت هر سه در حال آموزش هستند و وزن‌های شبکه در حال tune شدن هستند. از دفعات ۶۰۰ به بعد الگوریتم‌های Q-softmax و AGREL واکنشات بهتری نشان می‌دهند، در حالی که Q-greedy همچنان انتخاب‌های اشتباهی را پیش بینی می‌کند. از مرحله ۶۰۲ به بعد AGREL

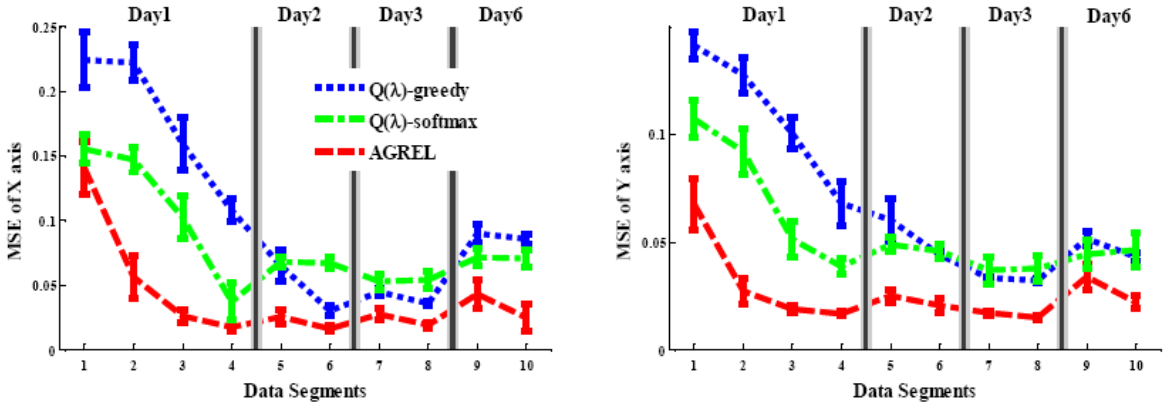
انتخاب‌های دقیقی را پیش بینی می‌کند در حالی که در Q-softmax اشتباهاتی دیده می‌شود. از مرحله ۱۸۱۰پیش بینی هر دو الگوریتم AGREL و Q-softmax انتخاب‌های مناسب را اتخاذ می‌کنند ولی Q-greedy همچنان واکنش مناسبی نمی‌دهد.



نمودار ۱: پیش بینی‌های سیستم BMI با استفاده از سه روش AGREL، Q-greedy و Q-Softmax

نمودار ۲ مربوط به MSE (Mean Square Error) های بدست آمده از پیش بینی های سه روش مورد بررسی می‌باشد. نمودار سمت راست مربوط به پیش بینی‌ها برای انتخاب‌ها در جهت افقی(محور X) و نمودار سمت چپ مربوط به پیش بینی‌ها برای انتخاب‌ها در جهت عمودی(محور Y) است.

با توجه به نتایج بدست آمده مشاهده می‌شود که در روز اول سرعت کاهش MSE در AGREL بیشتر از Q-softmax و Q-greey بوده، و سرعت کاهش MSE در Q-softmax بیشتر از Q-greedy بوده است. همچنین در روزهای ۲، ۳ و ۶ همچنان MSE AGREL پایین‌تر از Q-softmax و Q-greedy بوده است.



نمودار ۲: MSE های بدست آمده از پیش بینی های سه روش AGREL، Q-greedy و Q-Softmax

منابع

[1] R. Chavarriaga, and J. del R Millán, “Learning from EEG error-related potentials in noninvasive brain-computer interfaces,” *IEEE Trans. Neural Syst. Rehabil. Eng., vol. 18, no. 4, pp. 381-388, 2010.*

[2] J. Millan, "On the need for on-line learning in brain-computer interfaces." *IEEE International Joint Conference on Neural Networks, vol. 4, pp. 2877-2882,2004.*

[3] J. DiGiovanna, B. Mahmoudi, J. Fortes, J. C. Principe, and J. C. Sanchez, “Coadaptive brain–machine interface via reinforcement learning,” *IEEE Trans. Biomed. Eng., vol. 56, no. 1, pp. 54-64, 2009.*

[4] Reinforcement Learning: An Introduction. Richard Sutton and Andrew Barto. MIT Press, 1998.

[5] P. R. Roelfsema, and A. Ooyen, “Attention-gated reinforcement learning of internal representations for classification,” *Neural Comput., vol. 17, no. 10, pp. 2176-2214, 2005.*