پروژه نهایی بیوانفورماتیک

سیاوش پورفلاح

۱۰ بهمن ۱۴۰۲

فهرست مطالب

۱ مقدمه

پروتئین‌ها نقش حیاتی در عملکردهای موجودات زنده دارند. در حالی که پیش‌بینی سکانس اصلی پروتئین‌ها کار آسانی است، پیشبینی خواص آنها بر اساس این سکانس‌ها(مانند ساختار دوم و دسترسی به حلال) مسئله اصلی ماست که باید حل شود. پیشفرض ما این است که این خواص به طور کلی از سکانس پروتئین نشات می‌گیرند، اما مشخص کردن آنها به روش محاسباتی همچنان کار سختی است. در زمان نگارش این مقاله، متودهای پیشبینی ساختار دوم پروتئین معمولا بر شبکه‌های MLP یا Multilayer Perceptron متکی بودند اما این روش‌ها به دلیل نیاز به قدرت پردازش بالا و فرآیند آموزش طولانی محدود بودند.

در این ارائه به بررسی مقاله MUST-CNN: A Multilayer Shift-and-Stitch Deep Convolutional Architecture for Sequence-based Protein Structure Prediction می‌پردازیم که در آن، با استفاده از مدل جدیدی بر اساس شبکه‌های عصبی کانولوشنی(Convolutional Neural Networks) به نام MUST-CNN و استفاده از تکنیکی به نام Multilayer Shift-and-Stitch، پیش بینی داده‌های سکانسی پروتئین‌ها انجام گرفته است.

برای اینکه بتوانیم از محدودیت‌های شبکه‌های MLP عبور کنیم، این مقاله استفاده از CNNها را پیشنهاد می‌دهند که می‌توانند برای خواص هر آمینواسید در تمام سکانس برچسب‌گذاری انجام دهند. CNNها در بقیه حوزه‌ها مثل بینایی کامپیوتری و پردازش زبان طبیعی به خاطر خواص محاسباتی‌شان به طور گسترده مورد استفاده قرار گرفته‌اند. اما از طرفی استفاده از CNN در داده‌های متوالی (Sequential) محدودیت‌های خودش را دارد. به دلیل اینکه در عملیات pooling وضوح داده‌های خروجی پایینتر است و سبب کاهش دقت می‌شود. برای حل این مشکل از تکنیک Shift-and-Stitch استفاده می‌شود. همچنین روش معرفی شده در این مقاله با تکنیک‌های انتقالی در که در کارهای دیگر در پیشبینی ساختار پروتئین و وظایف پردازش زبان طبیعی انجام می‌شود شبیه است.

۲ روش

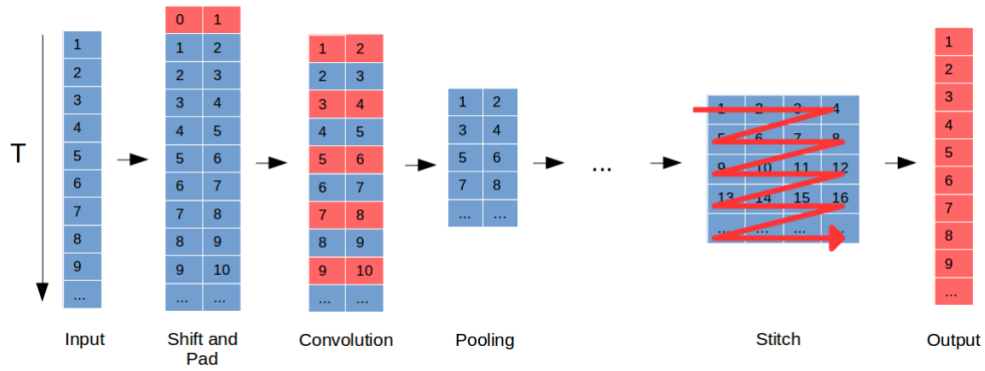
۱.۲ CNN

شبکه‌های عصبی کانولوشنی در اصل برای تصاویر دوبعدی استفاده می‌شدند اما در مسئله پشبینی توالی‌های پروتئینی در ابعاد تک‌بعدی استفاده می‌شوند. عملیات کانولوشن برروی تانسورهای داده‌های توالی و با استفاده از توابع مختلف غیر خطی از جمله tanh، ReLU و PReLU انجام می‌شوند که توابع ReLU و PReLU به دلیل عملکرد و کارایی محاسباتی مورد نظر ما هستند. پس از انحراف و عملیات غیر خطی، maxpooling روی خروجی‌ها اعمال می‌شود و به دنبال آن عملیات dropout برای جلوگیری از overfitting انجام می‌گیرد. برای ایجاد یک چارچوب چندلایه عمیق از این مدل از چندین لایه از شبکه کانولوشنی استفاده می‌شود.

۲.۲ تکنیک MUST

شبکه‌های کانولوشنی در اصل برای تصاویر دوبعدی استفاده می‌شدند اما در مسئله پیشبینی توالی‌های پروتئین به صورت یک بعدی استفاده می‌شوند. فرآیند pooling عملیاتی برای کاهش ابعاد است که چندین مقدار نزدیک به هم را با هم ترکیب کرده و یک مقدار تحویل می‌دهد، فرآیند maxpooling از تابع max برای انجام این کار استفاده می‌کند. به این دلیل maxpooling مهم است چرا که همه مقدارهای نزدیک به هم به یک مقدار رسیده‌اند و از این طریق، متود طبقه‌بندی تشویق می‌شود که تغییرناپذیری را یاد بگیرد. اما مشکلی که اینجا ایجاد می‌شود این است که با اعمال maxpooling با اندازه m بر روی یک توالی ورودی X با طول T، طول توالی خروجی maxpool به اندازه T/m است.

بنابراین از آنجایی که ابعاد هر توالی با یک فاکتور m تقسیم شده‌است، دیگر نمی‌توانیم همه پوزیشن‌های توالی اصلی را پیشبینی کنیم. برای حل این مشکل تکنیکی به نام shift-and-stitch در پژوهش‌های قبلی معرفی شده بود. این تکنیک در تصاویر دوبعدی هزینه محاسبات را بالا می‌برد، اما در پیشبینی توالی باید توالی‌های کمتری را به هم بخیه بزنیم (که همان stitch کردن است) که محاسبات را آسان‌تر می‌کند. این پروسه در شکل ۱ به تصویر کشیده شده است.



شکل ۱: تکنیک shift-and-stitch به ما اجازه می‌دهد که با اینکه maxpooling روی نمونه‌ها انجام شده هر المان در نمونه را برچسب‌گذاری کنیم.

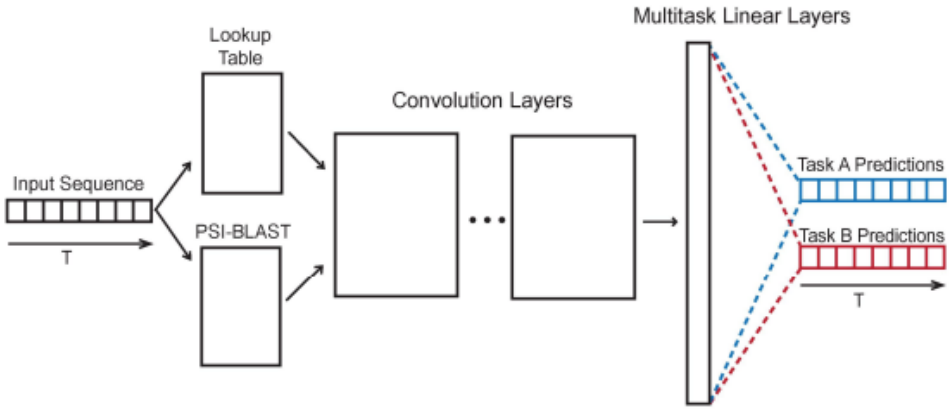
۳.۲ معماری شبکه

در این بخش به معماری شبکه MUST-CNN می‌پردازیم. در این شبکه ورودی به صورت توالی‌های آمینواسیدی کدگذاری شده دریافت می‌شود و بعد از عملیات پیش‌پردازش تکنیک shift-and-stitch روی آن اعمال شده که توالی‌های آمینواسیدی را بر اساس میزان pooling در هر لایه شیفت می‌دهد. هر توالی شیفت داده شده از همه لایه‌های کانولوشنی عبور می‌کند و نتایج آن به هم بخیه می‌شوند. این کار باعث می‌شود مدل بتواند همه توالی را با هم پردازش کند که باعث می‌شود سرعت کانولوشن بالا رفته و قابلیت استفاده از مدل‌های بزرگتر داشته باشیم.

بعد از این که توالی کدگذاری شده از لایه‌های کانولوشنی دریافت شد از چند لایه خطی متصل به هم عبور می‌کند. عملیات multitask یک رابطه خطی بین توالی‌های زنجیره پروتئینی و خواص مورد پیشبینی ایجاد می‌کند. خروجی این لایه‌ها نیز از یک لایه طبقه بندی softmax عبور می‌کند و فرآیند آموزش انجام می‌شود. در نهایت با fine-tuning وزن‌های آموزش داده شده اصلاح می‌شوند.

معماری این شبکه در شکل ۲ به تصویر کشیده شده است. برخی جزییات شبکه مانند استفاده آن از Lookup Table و PSI-BLAST و همچنین تحلیل وظایف یا taskهای پروتئین

نیز در شکل نمایش داده شده‌اند.



شکل ۲:

۴.۲ ارتباط با مطالعات قبلی

روش MUST-CNN به سه مدل در مطالعات گذشته مربوط است: مدل OverFeat، مدل Gen-erative Stochastic Networks یا GSNها، و مدل Conditional Neural Fields یا CNFها.

روش MUST-CNN از CNFها در عمق نمایش عملکرد بالاتری دارد. در حالی که CNF با استفاده از شبکه‌های MLP از یک نسخه پنجره‌ای و یک Conditional Random Field استفاده می‌کند، روش MUST-CNN از چندین لایه کانولوشنی و لایه‌های multitasking برای طبقه بندی استفاده کرده که باعث می‌شود برای ساختارهای پیچیده‌تر بهتر آموزش ببیند.

در مطالعات دیگر، شبکه MUST-CNN از شبکه‌های GSN یا شبکه‌های تصادفی مولد بهتر عمل می‌کند. GSNها شبیه ماشین‌های محدود بولتزمن هستند اما برای آموزش به یک الگوریتم تسویه نیاز دارند که باعث می‌شود کمتر قابل درک باشند. از طرفی MUST-CNN با لایه‌های کانولوشنی خود بیشتر در صنعت استفاده شده و فرآیندهای آموزش و آزمایش سریع‌تری را ارائه می‌دهد.

در نهایت MUST-CNN خود را از OverFeat در حوزه عملیات و رویکردی محلی‌سازی اشیا متمایز می‌کند. مدل OverFeat عموما بر طبقه‌بندی تصاویر تمرکز کرده و از تکنیک

shift-and-stitch صرفا در لایه آخر و برای افزایش وضوح استفاده می‌کند. در مقابل آن، MUST-CNN در طبقه‌بندی تک بعدی توالی‌های پروتئینی عمل می‌کند و shift-and-stitch را کاملا انتها به انتها اجرا می‌کند که همانطور که قبلا اشاره شد این کار در طبقه بندی تصاویر هزینه محاسبات را به شدت بالا می‌برد.

۳ پایگاه داده

در آزمایشات این مقاله از دو پایگاه داده بزرگ خواص پروتئینی استفاده شده است:

دیتاست 4prot که نویسندگان مقاله از آن به گونه‌ای استفاده کرده‌اند که مدل بر روی داده‌های train آموزش داده شده و سپس توسط داده‌های validation انتخاب شده و بهترین نتایج با امتحان کردن بر روی داده‌های test به دست می‌آیند.

همچنین دیتاست CullPDB که در این مقاله داده‌های train و validation از CullPDB و داده‌های test از CB513 استخراج شده‌اند که بخشی از CullPDB است.

۴ صحت

در فرآیند انتخاب مدل می‌توان دید که این مدل برای پارامترهای خود بسیار قوی است و با استفاده از maxpooling در کنار تکنیک shift-and-stitch در این مدل میانگین دقت تقریبا ۵.۰ درصد افزایش می‌یابد که این افزایش تقریبا بدون هیچ کاهش محاسباتی انجام می‌گیرد.

نتایج کار در دیتاست 4prot در جدول نشان داده می‌شود. مدل کوچکی که از طریق بهینه‌سازی بیزی ساخته شده تقریبا به اندازه مدل‌های پیشرفته قبلی پارامتر دارد اما می‌بینم که از شبکه‌های مورد مقایسه در مورد تمام وظایف بهتر عمل می‌کند.

۵ جمع بندی

در این مقاله یک معماری کانولوشنی با استفاده از تکنیک shift-and-stitch بررسی شد که با الهام از تکنیک‌های طبقه‌بندی تصاویر ساخته شده است. این معماری در برچسب‌گذاری

Model	Q_8
CNF (Wang et al. 2011)	.649
GSN (Zhou and Troyanskaya 2014)	.664
LSTM (Kaae Sønderby and Winther 2014)	.674
MUST-CNN (Ours)	**.684**

شکل ۳: اندازه‌گیری دقت با آموزش داده‌ها در CullPDB و تست برروی CB513. مدت زمان تست با داده‌های 4prot تقریبا یکسان است.

پوزیشن به پوزیشن یک توالی استفاده می‌شود و از مدل‌های دیگر عملکرد بالاتری دارد. سرعت این مدل به ما اجازه می‌دهد آموزش با ظرفیت بالاتری انجام دهیم و به نتایج بهتری برسیم. علاوه بر این، آزمایشات نشان می‌دهد این مدل با کمترین تنظیمات دستی از استحکام بالایی برخوردار می‌شود. همچنین انعطاف‌پذیری عملکرد مدل در برچسب‌گذاری توالی‌ها پیداست و در عملیاتی مثل استخراج ویژگی‌ها عملکرد بالایی دارد.