Diplomová práca

Analýza a návrh riešenia

Samuel Baran

Názov práce:	Nové techniky učenia sa bez učiteľa pre klasifikáciu a predikciu molekulárnych vlastností
Zadávateľ:	Tachyum s.r.o.
Vedúci práce (UPJŠ):	RNDr. Ľubomír Antoni, PhD.
Konzultant (Tachyum):	RNDr. René Derian, PhD.

Úvod a motivácia

Strojové učenie je podmnožinou umelej inteligencie, pričom sa zaoberá metódami a algoritmami učenia sa stroja z údajov. Cieľom strojového učenia je modelovanie algoritmov učenia sa pomocou stroja na základe vstupných dát v definovanom priestore riešení. Strojové učenie bolo v poslednej dobe úspešne aplikované na riešenie úloh v mnohých oblastiach vedy. V oblasti predikcie a klasifikácie molekulárnych vlastností sa stretávame s nižšou dostupnosťou údajov a vyššou heterogénnosťou vstupných údajov. Tieto skutočnosti vplývajú na nevhodnosť používania metód kontrolovaného učenia v tejto oblasti. Z týchto dôvodov vzniká priestor pre využitie metód učenia sa bez učiteľa, ktorého prednosťou je primárne využitie neoznačených dát na vytvorenie skrytých reprezentácií vstupov, ktoré poskytujú lepšiu východiskovú pozíciu pre modely učenia sa s učiteľom.

Východiská

Predikcia molekulárnych vlastností

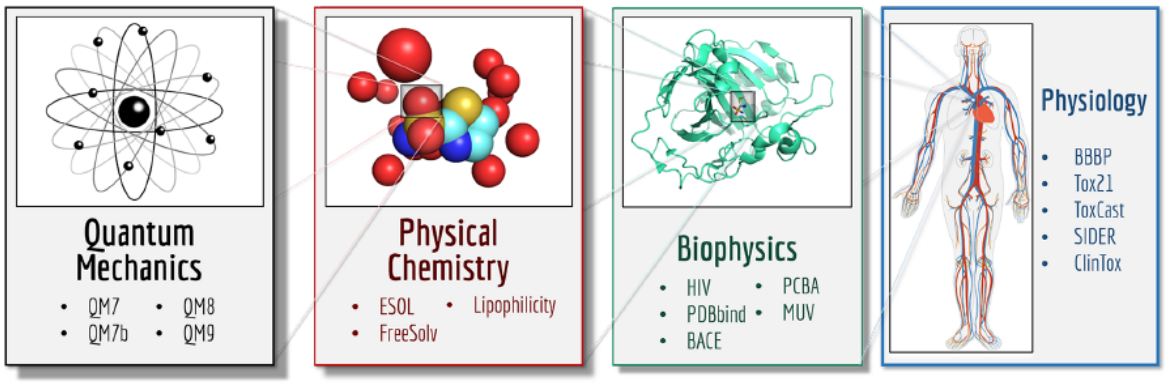
Tak ako v mnohých iných oblastiach, aj v chémii možno v posledných rokoch sledovať nárast využitia metód strojového učenia. Nové metódy v kombinácii s dostupnosťou väčších datasetov umožnili algoritmom strojového učenia uplatniť sa aj v oblasti predikcie molekulárnych vlastností.

Spočiatku bol vývoj v tejto oblasti limitovaný neprítomnosťou jednotného prístupu k vyhodnocovaniu efektívnosti jednotlivých metód. Na tento nedostatok reagovali autori článku MoleculeNet: A Benchmark for Molecular Machine Learning, ktorí vrámci knižnice DeepChem združili viacero rôznych datasetov z oblasti predikcie molekulárnych vlastností, určili metriky vhodné na evaluáciu modelov trénovaných na týchto datasetoch a sprístupnili open source implementácie algoritmov vyvinutých práve pre túto oblasť.

Dáta

Dáta pre úlohy strojového učenia spracúvajúce molekuly sú vysoko heterogénne, keďže obsahujú molekuly variabilnej dĺžky pozostávajúce z rôznych navzájom prepojených komponentov. Získavanie týchto dát je vzhľadom na potrebu špecializovaných zariadení a dohľadu odborníkov náročné, čo spôsobuje, že molekulárne datasety sú oveľa menšie ako tie, ktoré sú využívané pri úlohách strojového učenia z iných oblastí.

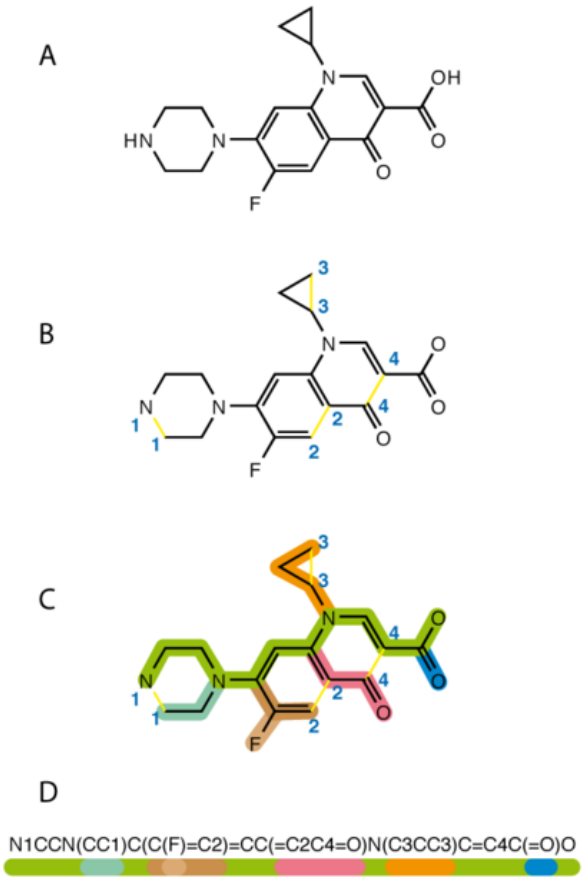
MoleculeNet obsahuje informácie o vlastnostiach vyše 700 000 chemických zlúčenín. Vzhľadom na oblasti do ktorých spadajú skúmané vlastnosti je možné rozdeliť datasety do štyroch základných kategórií: kvantová mechanika, fyzikálna chémia, biofyzika a fyziológia. Ako je znázornené na obrázku 1, jednotlivé datasety skúmajú rôzne úrovne molekulárnych vlastností, od energií excitovaných stavov (QM8), cez rozpustnosť molekúl vo vode (ESOL) až po skúmanie vlastnosti zabraňujúcej rozmnožovaniu vírusu HIV (HIV).



Obr. 1: Úlohy prislúchajúce jednotlivým datasetom sa zameriavajú na rôzne kategórie molekulárnych vlastností.

Reprezentácia molekúl

Molekuly sú v datasetoch kódované pomocou SMILES notácie, čo je jednoriadkové textové označenie molekuly, ktoré vznikne linearizáciou grafu molekuly pomocou očíslovania vrcholov a hrán a následného prechádzania grafu podľa topologického usporiadania. Proces linearizácie grafu je znázornený na obrázku 2.



Obr. 2: Linearizácia grafu. V grafe sa detekujú cykly a vymažú sa hrany, ktorých odstránením sa cykly rozpoja. K názvom vrcholov hrany sa pridá jednoznačný číselný identifikátor hrany. (B) Následne sa vyberie počiatočný vrchol a generuje sa SMILES prehľadávaním do hĺbky (C).

Zo spôsobu ktorým je vytváraná SMILES reprezentácia grafu plynie, že pre jednu molekulu môže existovať viacero validných SMILES reprezentácií a ich počet rastie s komplexnosťou molekuly.

Za účelom dosiahnutia bijektívneho zobrazenia z množiny molekúl do množiny ich SMILES reprezentácií boli vyvinuté algoritmy, ktoré pre každú molekulu vyberajú

jednu reprezentáciu nazývanú kanonický SMILES. Neexistuje však globálny kanonický SMILES, keďže rôzne nástroje využívajú rôzne kanonizačné algoritmy.

Samokontrolované učenie

Ako sme vyššie spomínali, pri úlohách predikcie molekulárnych vlastností sa stretávame so zhoršenou dostupnosťou anotovaných dát, čo znemožňuje efektívne využitie metód učenia sa s učiteľom, keďže modely trénované na takto malých datasetoch zle generalizujú a sú náchylné na preučenie. Preto sa v poslednej dobe vývoj zameral aj na oblasť samokontrolovaného učenia, ktorého prednosťou je primárne využitie neoznačených dát na vytvorenie skrytých reprezentácií vstupov, ktoré poskytujú lepšiu východiskovú pozíciu pre modely učenia sa s učiteľom. Predstavíme si dva takéto prístupy: ChemBERTa a MolCLR.

ChemBERTa

Väčšina konvenčných metód strojového učenia vyžaduje vstup rovnakej dĺžky, čo je pri SMILES reprezentácii molekúl vzhľadom na ich variabilnú veľkosť netriviálna úloha.

Hlavnými prístupmi pri riešení tohto problému je využitie grafových neurónových sietí (GNN) a metóda chemických odtlačkov (chemical fingerprints). Vzhľadom na fakt, že sa transformery stali štandardným nástrojom na učenie reprezentácií v oblasti spracovania prirodzeného jazyka (NLP), vyvstala otázka na preskúmanie prínosu tejto metódy aj pri spracovávaní reprezentácií chemických zlúčenín. Práca ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction ukázala, že transformery sú minimálne konkurencieschopným nástrojom využiteľným pri predikcii molekulárnych vlastností.

V rámci evaluácie bola metóda ChemBERTa predtrénovaná na datasete obsahujúcom 10 miliónov chemických zlúčenín porovnávaná s najpoužívanejšími modelmi v oblasti predikcie molekulárnych vlastností, ktorými sú náhodné lesy (RF), metóda podporných vektorov (SVM), a orientovaná neurónová sieť s preposielaním správ (directed message passing neural network - D-MPNN). Výsledky porovnaní sú zhrnuté v tabuľke 1.

Tab. 1: Porovnanie modelu ChemBERTa predtrénovaného na 10 miliónovom datasete PubChem a úspešných modelov na vybraných úlohách predikcie molekulárnych vlastností

	BBBP 2,039		**ClinTox (CT_TOX)** 1,478		**HIV** 41,127		**Tox21 (SR-p53)** 7,831	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
ChemBERTa 10M	0.643	0.620	0.733	0.975	0.622	0.119	**0.728**	0.207
D-MPNN	**0.708**	0.697	**0.906**	**0.993**	0.752	0.152	0.688	**0.429**
RF	0.681	0.692	0.693	0.968	**0.780**	**0.383**	0.724	0.335
SVM	0.702	**0.724**	0.833	0.986	0.763	0.364	0.708	0.345

Aj keď ChemBERTa neprekonal vybrané modely, ukázali transformery potenciál pre ďalšie využitie v tejto oblasti, jednak tým že sú dobre škálovateľné vzhľadom na veľkosť datasetu určeného na predtréning, ale aj možnosťou využitia vizualizácie attention mechanizmu.

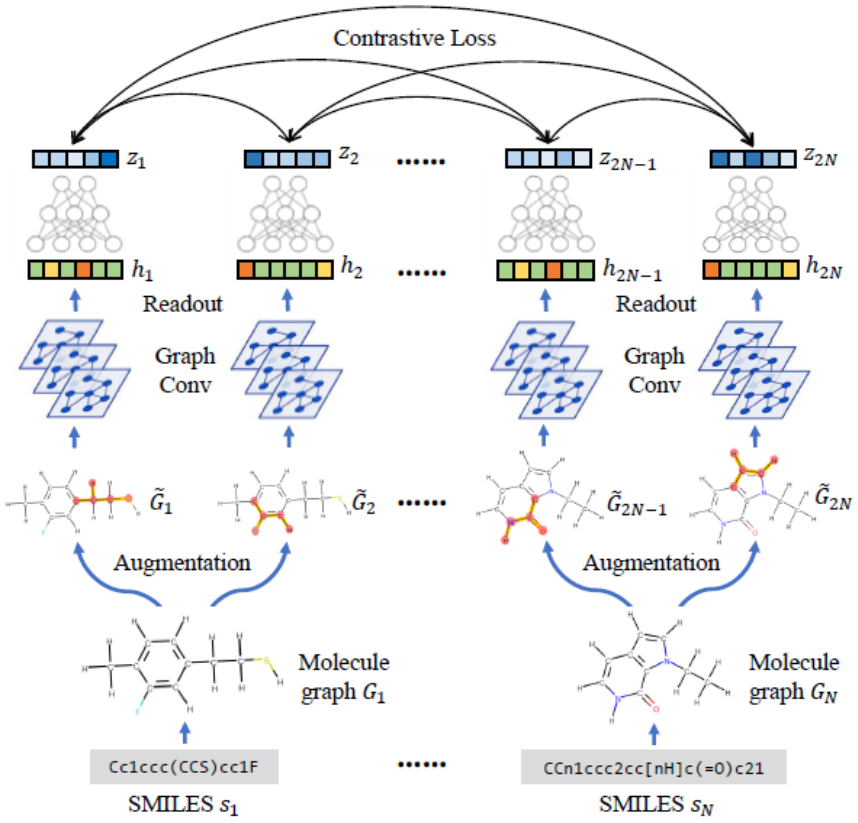
MolCLR

Jedným z ďalších prístupov samokontrolovaného učenia je kontrastívne učenie, kde sú reprezentácie vstupu trénované pomocou augmentácií vstupu a následnej snahy dosiahnuť rovnaké reprezentácie pre takto augmentované vstupy.

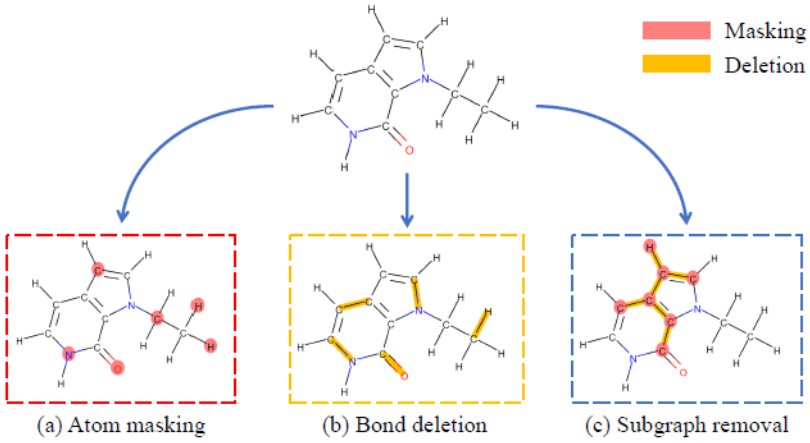
Využitím grafových neurónových sietí a aplikovaním kontrastívneho predtréningu vznikol model MolCLR: Molecular Contrastive Learning of Representations via Graph Neural Networks pomenovaný podľa jedného z prvých modelov kontrastívneho učenia SimCLR. Architektúru modelu opisuje obrázok 3.

GNN sú v tomto prípade využívané na kódovanie grafu molekúl a oproti SMILES reprezentácii sú schopné uchovávať aj informácie o topológii molekúl.

Aby bolo možné využiť princípy kontrastívneho učenia, boli predstavené 3 stratégie augmentácie aplikovateľné na GNN: maskovanie atómu, odstránenie väzieb a odstránenie podgrafu. Pri maskovaní atómu je vopred určený počet náhodných atómov nahradený špeciálnym znakom. Odstraňovanie väzieb narozdiel od maskovania atómov odstraňuje náhodne vybrané chemické väzby. Na odstránenie podgrafu je možné nahliadať ako na kombináciu predchádzajúcich dvoch augmentácií. Vizualizáciu týchto augmentácií zachytáva obrázok 4.



Obr. 3: Molekulárne kontrastívne učenie reprezentácií s použitím GNN. SMILES reprezentácia s_i je transformovaná na graf molekuly G_i ,na ktorý sú následne aplikované dva náhodne vybrané grafové augmentácie. Tým dostávame dva augmentované grafy G'_{2i-1}, G'_{2i}, ktoré sú vstupom pre sieť počítajúcu reprezentácie molekúl. Na takto vypočítané reprezentácie je aplikovaná kontrastívna stratová funkcia s cieľom maximalizovať zhodu reprezentácií dvojíc augmentácií.



Obr. 4: Tri augmentácie grafu prislúchajúceho molekule. (a) Maskovanie atómov náhodne nahradí vybrané atómy maskovacou značkou. (b) Odstránenie väzby náhodne odstráni

väzby medzi dvojicami atómov. (c) Odstránenie podgrafu náhodne odstráni indukovaný podgraf.

Experimenty zachytené v tabuľke 2 ukázali, že metóde MolCLR sa na viacerých úlohách predikcie vlastností molekúl podarilo prekonať najlepšie modely z oblasti kontrolovaného učenia akými sú náhodné lesy (RF), metóda podporných vektorov (SVM), orientovaná neurónová sieť s preposielaním správ (directed message passing neural network - D-MPNN) a grafová neurónová sieť MGCNN vyvinutá práve na predikciu molekulárnych vlastností.

Tab. 2: Porovnanie modelov na základe ROC-AUC (%) testovacej vzorky, kde prvé štyri modely sú modely učenia s učiteľom a zvyšné tri sú modely samokontrolovaného učenia.

Dataset	BBBP	Tox21	ClinTox	HIV	BACE	SIDER	MUV
# Molecules	2039	7831	1478	41127	1513	1478	93087
# Tasks	1	12	2	1	1	27	17
RF	71.4±0.0	76.9±1.5	71.3±5.6	78.1±0.6	**86.7±0.8**	**68.4±0.9**	63.2±2.3
SVM	72.9±0.0	**81.8±1.0**	66.9±9.2	**79.2±0.0**	86.2±0.0	**68.2±1.3**	67.3±1.3
MGCN [74]	**85.0±6.4**	70.7±1.6	63.4±4.2	73.8±1.6	73.4±3.0	55.2±1.8	70.2±3.4
D-MPNN [28]	71.2±3.8	68.9±1.3	**90.5±5.3**	75.0±2.1	85.3±5.3	63.2±2.3	**76.2±2.8**
HU. et.al [60]	70.8±1.5	78.7±0.4	78.9±2.4	80.2±0.9	85.9±0.8	65.2±0.9	81.4±2.0
N-Gram [75]	**91.2±3.0**	76.9±2.7	85.5±3.7	**83.0±1.3**	87.6±3.5	63.2±0.5	81.6±1.9
MolCLR	73.6±0.5	**79.8±0.7**	**93.2±1.7**	80.6±1.1	**89.0±0.3**	**68.0±1.1**	**88.6±2.2**

Model zaznamenal lepšie výsledky aj oproti metódam samokontrolovaného učenia, kde bol porovnávaný s modelmi grafových neurónových sietí navrhnutými Hu a kolektívom a tiež aj s modelom generujúcim reprezentácie molekúl pomocou N-gramového grafu (N-gram graph: Simple unsupervised representation for graphs, with applications to molecules).

Samokontrolované učenie v oblasti počítačového videnia

Oblasť počítačového videnia sa venuje viacerým úlohám, ako sú napríklad klasifikáciu obrázkov, detekciu objektov, segmentáciu obrazu, ale aj mnoho ďalších.

Pri úlohách spracovávania obrazu sa osvedčili konvolučné neurónové siete, avšak aj tie vzhľadom na dimenzionalitu a rozmer vstupov potrebujú veľké datasety, aby predišli preučeniu. Ako alternatíva sa ukázalo využitie transformerov inšpirované úspešnými aplikáciami v oblasti spracovania prirodzeného jazyka (NLP), čo vyústilo

do modelu nazývaného visual transformer (ViT). Aj napriek tomu, že modely ViT boli konkurencieschopné konvolučným sieťam, nepodarilo sa preukázať jednoznačné prednosti týchto modelov. ViT vyžadujú viac tréningových dát a sú výpočtovo náročnejšie.

Vzhľadom na náročnosť manuálnej anotácie veľkých množstiev vstupných dát aj v tejto oblasti vznikla potreba využitia samokontrolovaných metód strojového učenia. Tieto metódy ukázali potenciál v spojení s konvolučnými neurónovými sieťami, kde príkladom môže byť SimCLR, jedna z metód kontrastívneho učenia alebo samokontrolované metódy učenia sa reprezentácií obrázkov MoCo (Momentum Contrast for Unsupervised Visual Representation Learning) a BOYL (Bootstrap Your Own Latent).

Úspech nekontrolovaného predtréningu transformerov v oblasti NLP (BERT, GPT) bol motiváciou pre prispôsobenie a aplikáciu týchto princípov aj v oblasti počítačového videnia. Článok Generative pretraining from pixels opisuje model učiaci sa reprezentácie obrazu vhodné na ďalšie úlohy predikcie a klasifikácie, pomocou generatívneho predikovania časti obrazu. Ďalším prístupom je metóda DINO, ktorej sa budeme venovať v nasledujúcej kapitole.

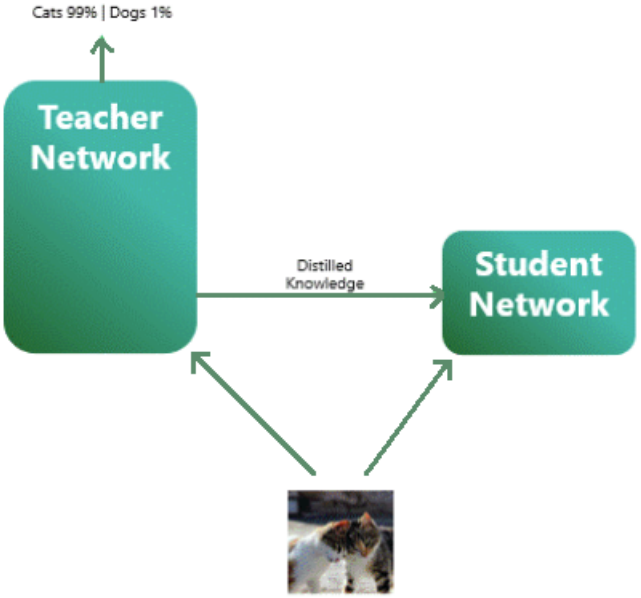
Dino

DINO (self distilation with no labels) je samokontrolovaná metóda učenia sa reprezentácií obrazu predstavená v práci Emerging Properties in Self-Supervised Vision Transformers, ktorá ako základ používa destiláciu znalostí.

Destilácia znalostí je proces (obrázok 5), do ktorého vstupuje veľký, nie nutne označkovaný dataset, a dve siete: učiteľská, zvyčajne zložitejšia, ktorá je natrénovaná na nejakej konkrétnej úlohe a študentská s jednoduchšou štruktúrou. Myšlienka destilácie znalostí spočíva v tom, že predikcie získané pomocou učiteľskej siete sú prezentované študentskej sieti ako značky, ktoré má predikovať. Výsledkom takéhoto procesu je študentská sieť s rovnakými znalosťami ako učiteľská.

Metóda DINO pozostáva z dvoch sietí s názvami študentská a učiteľská, ktoré majú rovnakú architektúru a dostávajú na vstup dve reprezentácie rovnakého obrázku. Z

každého obrázku z tréningovej množiny sú vytvorené dva čiastočne sa prekrývajúce globálne výseky zaberajúce viac ako 50% plochy obrázku a niekoľko ďalších malých lokálnych výsekov s plochou menšou ako 50% pôvodného obrázku.



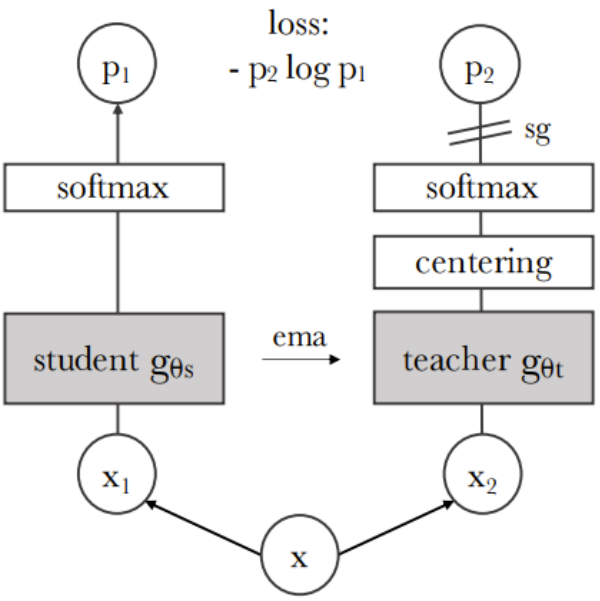
Obr. 5: Destilácia znalostí. Vstupné dáta sú poskytnuté na vstup obom sieťam a študentská sieť sa učí predikovať to čo predikuje učiteľská.

Obr. 6: Príklad globálnych a lokálnych výsekov obrázkov.

Na túto architektúru (obrázok 7) sa nahliada ako na destiláciu, počas ktorej sa študentská sieť učí na základe predikcií učiteľskej siete. Rozdiel oproti klasickej destilácii je v tom, že učiteľská sieť nie je netrénovaná na konkrétnej úlohe, ale

počas tréningu sa časť vedomostí nadobudnutých študentskou sieťou v predchádzajúcich iteráciach prenáša na učiteľskú.

Študentská sieť dostáva na vstup náhodne niektorý z globálnych alebo lokálnych výsekov, zatiaľ čo učiteľskej sieti sú poskytované len globálne výseky vstupných obrázkov. Obe siete majú rovnakú architektúru, ale rozličné parametre. Výstup učiteľskej siete je centralizovaný pomocou priemeru počítaného naprieč celou dávkou (batch). Výstupy sietí sú normalizované pomocou softmax funkcie s parametrom teploty a ich podobnosť je vypočítaná pomocou krížovej entropie. Na konci učiteľskej siete je operátor zabraňujúci spätnému šíreniu gradientu, čo spôsobuje, že chyba je spätne šírená len študentskou sieťou. Parametre učiteľskej siete sú zmenené pomocou ema funkcie (exponential moving average).



Obr. 7: Architektúra DINO. Model dá na vstup študentskej a učiteľskej siete dve náhodné transformácie vstupného obrázku. Obe siete majú rovnakú architektúru, ale rozličné parametre. Výstup učiteľskej siete je centralizovaný pomocou priemeru počítaného naprieč celou dávkou (batch). Výstupy sietí sú normalizované pomocou softmax funkcie s parametrom teploty a ich podobnosť je vypočítaná pomocou krížovej entropie. Na konci učiteľskej siete je operátor zabraňujúci spätnému šíreniu gradientu, čo spôsobuje, že chyba je spätne šírená len študentskou sieťou. Parametre učiteľskej siete sú zmenené pomocou ema funkcie (exponential moving average).

Samotná sieť pozostáva z dvoch komponentov: kostry, ktorú tvorí buď konvolučná sieť alebo visual transformer, za ktorou nasleduje projekčná hlava. Po natrénovaní sa ako reprezentácie obrázkov berú výstupy kostry siete.

Algoritmus 1: Pseudokód metódy DINO

```
# gs, gt: student and teacher networks
# C: center (K)
# tps, tpt: student and teacher temperatures
# l, m: network and center momentum rates
gt.params = gs.params
for x in loader: # load a minibatch x with n samples
    x1, x2 = augment(x), augment(x) # random views

    s1, s2 = gs(x1), gs(x2) # student output n-by-K
    t1, t2 = gt(x1), gt(x2) # teacher output n-by-K

    loss = H(t1, s2)/2 + H(t2, s1)/2
    loss.backward() # back-propagate

    # student, teacher and center updates
    update(gs) # SGD
    gt.params = l*gt.params + (1-l)*gs.params
    C = m*C + (1-m)*cat([t1, t2]).mean(dim=0)

def H(t, s):
    t = t.detach() # stop gradient
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) # center + sharpen
    return - (t * log(s)).sum(dim=1).mean()
```

Takto navrhnutá architektúra produkuje reprezentácie obrázkov, ktoré v sebe nesú explicitnú informáciu o sémantickej segmentácii obrazu a tiež sú vhodné ako vstup pre KNN algoritmus.

Ciele práce

Pre túto bakalársku prácu boli sformulované nasledujúce ciele:

- Spracovať prehľad metód klasifikácie a predikcie vlastností molekúl a najnovších techník učenia sa bez učiteľa alebo čiastočného učenia bez učiteľa, ktoré sú využívané v oblasti počítačového videnia.
- Navrhnúť novú metódu klasifikácie a predikcie vlastností molekúl, ktorá bude inšpirovaná existujúcimi technikami učenia sa bez učiteľa alebo čiastočného učenia bez učiteľa z oblasti počítačového videnia.
- Implementovať a experimentálne overiť navrhnuté riešenie testovaním na dátach pre klasifikáciu a predikciu molekulárnych vlastností.

- Porovnať dosiahnuté výsledky navrhnutého riešenia s doteraz najúspešnejšími technikami vyvinutými v oblasti molekulárnych vlasností

Návrh riešenia

Jedným z cieľov práce je návrh metódy klasifikácie a predikcie molekulárnych vlastností inšpirovanej najnovšími technikami učenia sa bez učiteľa z oblasti počítačového videnia. Ako základ našej metódy plánujeme využiť upravenú architektúru DINO.

Jedna zo zmien bude súvisieť s prácou s dátami. Vzhľadom na fakt, že pracujeme s textovými reprezentáciami namiesto obrazových dát, bude potrebné nájsť vhodné metódy vytvárania globálnych a lokálnych výsekov SMILES-ov, tak, aby bola zachovaná funkčná podstata tvorby týchto výsekov.

Generovanie výsekov

Prvým návrhom je vyberať súvislé podpostupnosti SMILES-ov a zaraďovať ich do množín globálnych alebo lokálnych výsekov pomocou relatívnej dĺžky vzhľadom na veľkosť danej reprezentácie. Pri návrhu augmentácie bude potrebné preskúmať aj nutnosť využitia doménových znalostí, teda či je možné hranice výsekov voliť náhodne, alebo je potrebné dbať na pravidlá, pomocou ktorých boli dané reprezentácie vytvárané.

Architektúra siete

Ďalšou zmenou súvisiacou s iným formátom vstupných dát je samotná architektúra študentskej a učiteľskej siete.

Priamočiarym riešením je v tomto prípade nahradiť ViT klasickým transformerom spracovávajúcim textové dáta a nahliadať na SMILES ako na text jazyka pozostávajúci zo základných znakov abecedy spájaných na základe gramatických pravidiel, čím by sme prinavrátili transformerom ich pôvodný význam v oblasti NLP. Bude však potrebné preskúmať možnosti tokenizácie SMIELS-ov, a teda či ako abecedu jazyka budeme chápať jednotlivé označenia chemických prvkov, znaky pre

väzby, žatvorky a číselné označenia prvkov, ktoré boli označené pri rozbíjaní cyklov v grafe molekuly, alebo na nájdenie tokenov zvolíme automatizovaný spôsob založený na štatistických metódach nad datasetom.

Alternatívnym prístupom by bolo využitie grafových neurónových sietí, ktoré boli použité napr. aj v rámci metódy MolCLR. To však vzniká otázka, či by bolo možné využiť vyššie spomínaný spôsob generovania výsekov, keďže nie každý výsek musí odpovedať validnému SMILES-u, z ktorého je možné generovať graf. Do úvahy však prichádza použitie augmentácie vyberajúcej náhodný súvislý indukovaný podgraf.

Zdroje:

1. Zhenqin Wu and Bharath Ramsundar and Evan N. Feinberg and Joseph Gomes and Caleb Geniesse and Aneesh S. Pappu and Karl Leswing and Vijay S. Pande, . "MoleculeNet: A Benchmark for Molecular Machine Learning". CoRR abs/1703.00564. (2017).
2. DeepChem - https://deepchem.readthedocs.io/en/latest/index.html
3. Mathilde Caron and Hugo Touvron and Ishan Misra and Hervé Jégou and Julien Mairal and Piotr Bojanowski and Armand Joulin, . "Emerging Properties in Self-Supervised Vision Transformers". CoRR abs/2104.14294. (2021).
4. Mahmoud Assran and Mathilde Caron and Ishan Misra and Piotr Bojanowski and Armand Joulin and Nicolas Ballas and Michael G. Rabbat, . "Semi-Supervised Learning of Visual Features by Non-Parametrically Predicting View Assignments with Support Samples". CoRR abs/2104.13963. (2021).
5. Yuyang Wang and Jianren Wang and Zhonglin Cao and Amir Barati Farimani, . "MolCLR: Molecular Contrastive Learning of Representations via Graph Neural Networks". CoRR abs/2102.10056. (2021).
6. Seyone Chithrananda and Gabriel Grand and Bharath Ramsundar, . "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction". CoRR abs/2010.09885. (2020).
7. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, 2020.
8. Jean-Bastien Grill, Florian Strub, Florent Altch´e, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, R´emi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In NeurIPS, 2020.

9. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. preprint arXiv:2002.05709, 2020.
10. Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan and Ilya Sutskever. Generative Pretraining from Pixels. In PMLR, 2020
11. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.
12. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
13. On DINO, Self-Distillation with no labels https://towardsdatascience.com/on-dino-self-distillation-with-no-labels-c29e9365e382
14. ChemBERTa repository https://github.com/seyonechithrananda/bert-loves-chemistry