

Dirbtinio intelekto modelių kolapsas

Atliko: Greta Virpšaitė

Darbo vadovas: j. asist. Boleslovas Dapkūnas

Vilniaus universitetas

Matematikos ir informatikos fakultetas

Programų sistemų bakalauro studijų programa

2025

Dirbtinio intelekto (DI) modelių kolapsas - degeneratyvus procesas, kuris atsiranda, kai generatyviniai modeliai mokomi naudojant jų pačių sugeneruotus duomenis praranda gebėjimą tiksliai atspindėti pradinę duomenų informaciją.

DI modelių kolapsas



pav. 1: **FFHQ-StyleGAN2 kolapso pavyzdys.** Didėjant generacijų (angl. *Generation*) skaičiui t, matomas išvesčių suvienodėjimas.

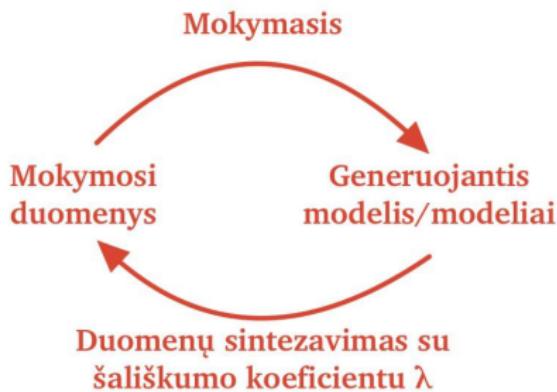
- Giliau išnagrinėti DI modelių kolapso mechanizmus, šiuolaikinius mokslinius metodus, skirtus šios problemos identifikavimui ir sprendimui, su tikslu pateikti įžvalgas apie galimas prevencines priemones ir jų implementavimą kasdienėje praktikoje.

- Išanalizuoti literatūrą, susijusią su DI modelių kolapsu, apimant pagrindinius mechanizmus, rizikas ir galimus sprendimus. Aprašyti DI modelių kolapso atpažinimo technikas, remiantis šiuolaikiniais tyrimais.
- Apibrėžti pagrindines DI modelių kolapso prevencijos strategijas, įtraukiant Europos Sajungos DI aktą ir jo įtaką DI modelių stabilumui.

- Atlikti eksperimentą su VAE modeliu, mokant jį MNIST duomenų rinkiniu, siekiant ištirti modelio kolapso reiškinj.
- Aptarti etines implikacijas, susijusias su DI modelių kolapsu ir jo prevencija.
- Parengti rekomendacijas, pagrįstas atliktos analizės rezultatais, kurios galėtų būti pritaikomos moksliniuose ir praktiniuose kontekstuose.

Save valgantys ciklai - Mokymo ciklai, kurie parodo, kaip modelio mokymas su sugeneruotais duomenimis gali lemti kolapsą. Save valgantys ciklai priskiriami trims kategorijoms: Sintetinių duomenų ciklai, sintetinių augmentacijų ciklai ir ciklai su šviežiais duomenimis.

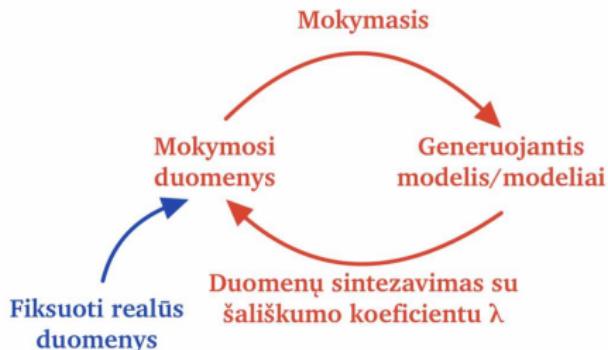
Save valgantys ciklai



pav. 2: Save valgančio ciklo iliustracija.

- **Pilnai sintetinių duomenų ciklas**: apibūdina ciklą, kuriame naudojami tik pilnai sintetiniai duomenys.

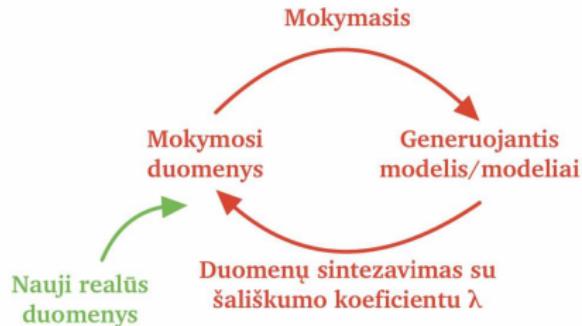
Save valgantys ciklai



pav. 3: Save valgančio ciklo iliustracija.

- **Sintetinių duomenų augmentacijų ciklas**: šiame cikle naudojami **fiksuoti realūs duomenys** kartu su **pilnai sintetiniai duomenimis**.

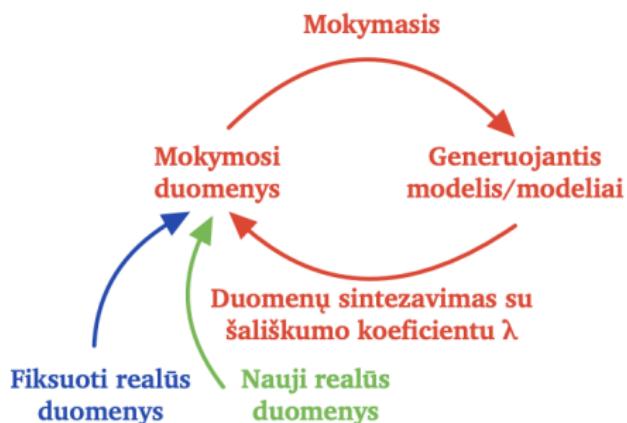
Save valgantys ciklai



pav. 4: Save valgančio ciklo iliustracija.

- **Ciklas su šviežiais duomenimis:** ciklas, apima **naujus realius duomenis** kartu su **pilnai sintetiniais duomenimis**.

Save valgantys ciklai



pav. 5: Save valgančio ciklo iliustracija.

- šališkumo parametras λ nusako, kaip atrodys generuojamų pavyzdžių kokybės ir įvairovės pasiskirstymas.

Kodėl pasirinkta VAE neuroninio tinklo architektūra?

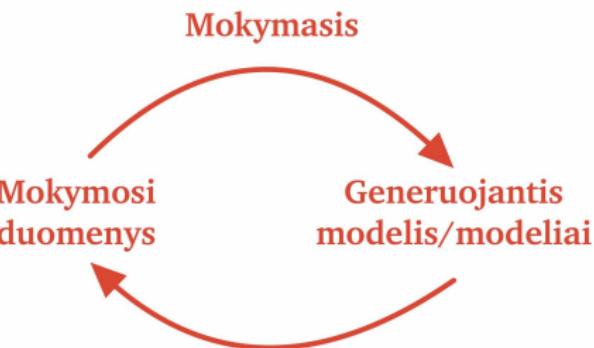
- Paprastos architektūros, kolapsui iliustruoti tinkamas tinklas, kurio veikimo mechanizmai aprašyti nagrinėjamoje mokslinėje literatūroje.
- Leidžia vizualiai stebėti modelio išvesčių pokyčius skirtingose treniravimo strategijose.

Kodėl pasirinktas MNIST duomenų rinkinys?

- Plačiai naudojamas kaip pavyzdinis duomenų rinkinys dirbtinio intelekto modelių mokymui.
- MNIST buvo naudotas viename iš nagrinėtų tyrimų, kitame - analizuotas λ koeficiente poveikis kolapsui, o šiame tyime tiriamas ciklo tipo parinkties efektas.

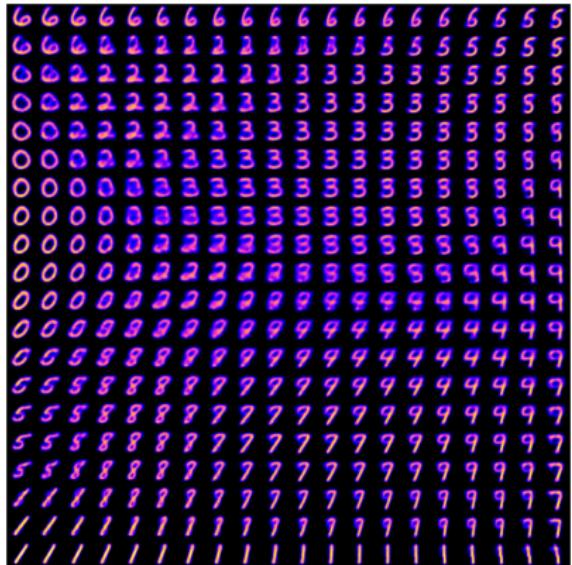
Rezultatai - sintetinių duomenų ciklas

Sintetinių duomenų ciklas:

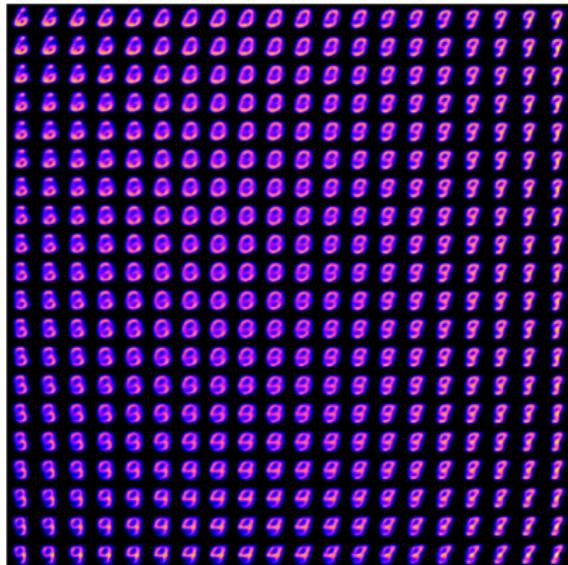


- Po penktosios generacijos VAE modelis parodė stiprų kolapso efektą: prarastos daugelio skaitmenų įvairovės savybės.
- Tik keli skaitmenys (0, 3, 6, 8, 9) buvo atpažįstami, tačiau ir jie buvo labai blankūs.

Rezultatai - sintetinių duomenų ciklas



(a) Pirma genereracija

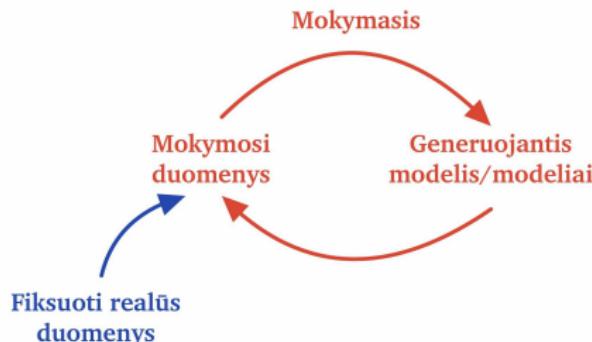


(b) Penkta genereracija

pav. 6: Pirmosios ir penktosios sintetinių duomenų ciklo generacijos palyginimas.

Rezultatai - sintetinių augmentacijų ciklas

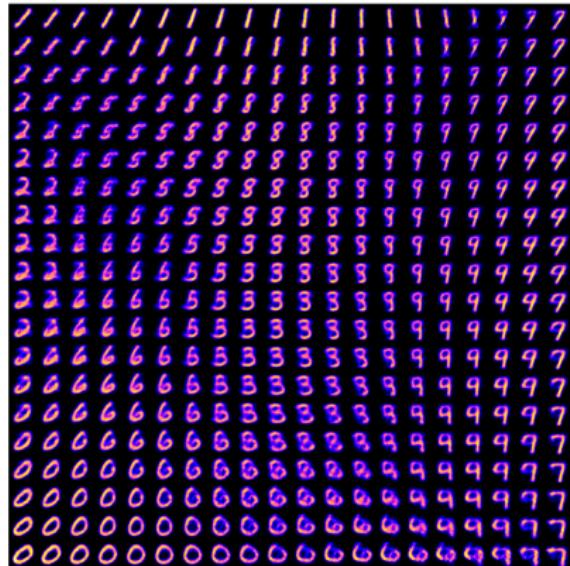
Sintetinių augmentacijų ciklas:



- Vizualiai rezultatų kokybė nuo pirmos iki penktos generacijos žymiai nesiskyrė.
- Pastebėtas nežymus vaizdų suliejimas.

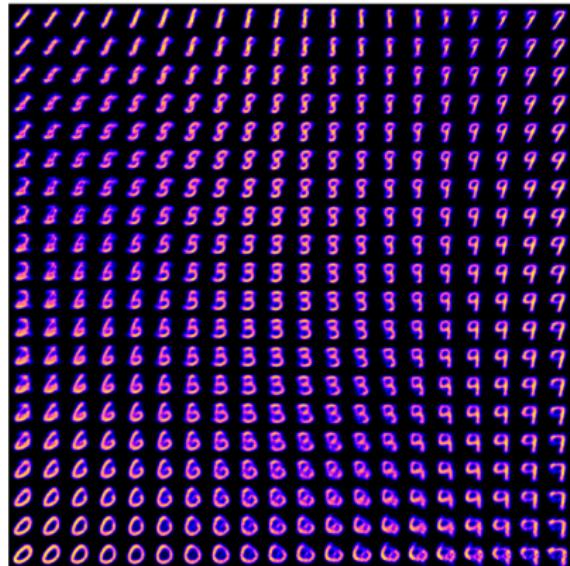
Rezultatai - sintetinių augmentacijų ciklas

Balanced Generation 1



(a) Pirma genereracija

Balanced Generation 5

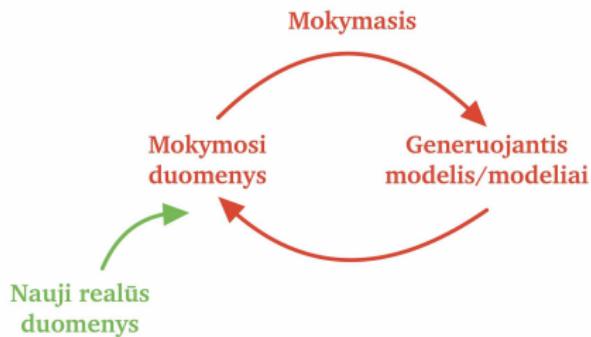


(b) Penkta genereracija

pav. 7: Pirmosios ir penktosios sintetinių augmentacijų ciklo generacijos palyginimas.

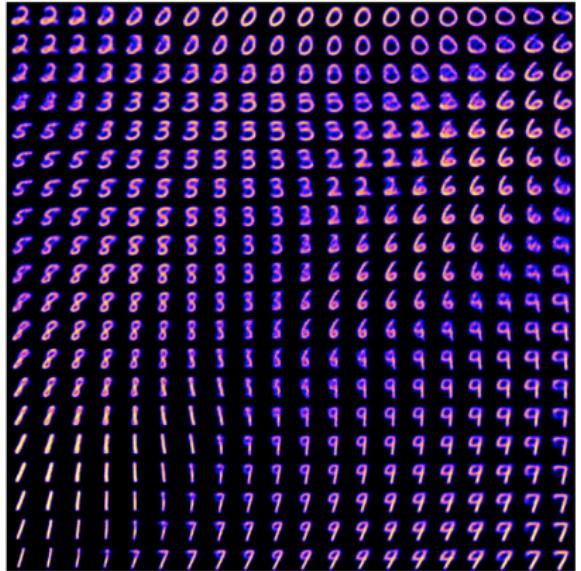
Rezultatai - ciklas su šviežiais duomenimis

- **Ciklas su šviežiais duomenimis:**

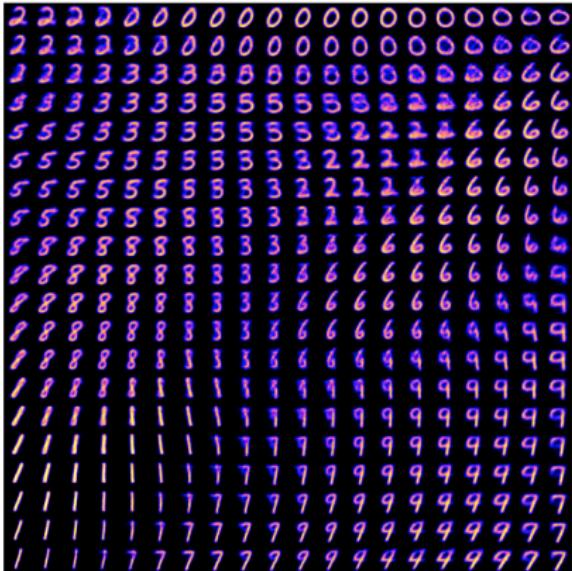


- Po penktos generacijos tinklo išvestys tapo šiek tiek ryškesnės.
- Jvairovės praradimas nebuvo pastebėtas.

Rezultatai - ciklas su šviežiais duomenimis



(a) Pirma genereracija



(b) Penkta genereracija

pav. 8: Pirmosios ir penktosios ciklo su šviežiais duomenimis generacijos palyginimas.

Standartinio modelio mokymas:

- Po penktos generacijos duomenų kokybė pagerėjo, vaizdai tapo ryškesni ir aiškesni.
- Modelis parodė gebėjimą mokytis.

Rezultatai - standartinio modelio mokymas



(a) Pirma generacija



(b) Penkta generacija

pav. 9: Pirmosios ir penktosios generacijų su standartiniu apmokymu palyginimas

DI modelių kolapso pobūdis:

- DI modelių kolapsas atsiranda, kai modeliai mokomi naudojant jų pačių sugeneruotus duomenis, o tai lemia duomenų įvairovės praradimą.
- Eksperimentai patvirtino, kad kolapsas sintetinių duomenų cikluose yra neišvengiamas ir pasireiškia jau po kelių generacijų.

Maišymo strategijos efektyvumas:

- Realių ir sintetinių duomenų maišymas (augmentacijos ir šviežių duomenų cikluose) sumažina kolapso riziką, tačiau netinkamas santykis šviežių duomenų cikluose gali sukelti kokybės praradimus, o sintetinių augmentacijų cikluose tik nutolinti DI modelių kolapsą.
- Naujų realių duomenų įtraukimas į mokymo procesą išlieka efektyviausiu būdu mažinti kolapso riziką.

Reguliacijos svarba:

- Europos Sajungos DI reglamentas, apimantis duomenų žymėjimą ir šališkumo kontrolę, gali netiesiogiai padėti DI modeliams išvengti kolapso.
- Reglamentas įsigalios tik 2026 metais, todėl kolapso prevencijai būtina plėtoti alternatyvius sprendimus iki to laiko.

Mokslininkams:

- Rekomenduojama testi tyrimus, siekiant geriau suprasti sintetinių ir šviežių realių duomenų santykio poveikį DI modelių stabilumui bei jų kokybei.
- Gilintis į šališkumo kontrolės paramетro λ reikšmę skirtinose generatyvinėse architektūrose ir mokymosi cikluose.
- Tyrinėti ilgalaikį sugeneruotų duomenų maišymo poveikį tarp skirtinų modelių ir jų rezultatų kokybę

Praktikams:

- Užtikrinti, kad duomenų rinkiniai būtų reguliariai atnaujinami šviežiais realiais duomenimis
- Užtikrinti, kad sintetiniai duomenys būtų naudojami subalansuotai, neviršijant saugią proporciją

Politikos kūrėjams ir vykdytojams:

- Siūloma Europos Sajungos DI akte toliau stiprinti reikalavimus sugeneruotų duomenų žymėjimui ir kontrolės mechanizmams.
- Svarstyti įtraukti daugiau nuostatų, skirtų mažesnės rizikos DI sistemoms.
- Valdomiesiems kūnams kurti ir kitus reglamentus ar įstatymus skirtus DI kolapso mažinimo tematikai

Bendruomenei:

- Didinti informuotumą apie DI modelių kolapso rizikas ir jų prevencijos svarbą.
- Atkreipti dėmesį į tai, kaip sistemų naudotojai prisideda prie duomenų kokybės, ir skatinti juos neteršti interneto modelių sugeneruotais duomenimis