TRANSZFORMER MODELLEK JELENTŐSÉGE AZ IDŐSOROK ELŐREJELZÉSÉNÉL ÉS OSZTÁLYOZÁSÁNÁL

Szerző:

Boros Gerzson Dávid
Data Science Europe Kft. és Redivivum

E-mail:
gerzson.boros@datascienceeurope.ai

Lektorok:

Mező Ferenc (Ph.D.)
Eszterházy Károly Katolikus Egyetem

Szabóné Balogh Ágota (Ph.D.)
Gál Ferenc Egyetem

…és további két anonim lektor

Absztrakt

A transzformátor modellek jelentős előrelépéseket hoztak az idősor-elemzésben, mivel képesek hosszú távú függőségeket megragadni. A tanulmány a transzformátor modellek (mélytanulási architektúrák) alapvető felépítését, erősségeit és korlátait vizsgálja az idősorok modellezésében. Bemutatja a különféle transzformátor variánsokat idősorokhoz, és ismerteti alkalmazásukat előrejelzésre, osztályozásra és anomália-felismerésre, valamint az elmagyarázhatóságot és interpretálhatóságot.

Kulcsszavak: transzformátor, idősor-elemzés, előrejelzés, osztályozás, mélytanulás

Diszciplínák: informatika, matematika, adattudomány

Abstract

IMPORTANCE OF THE TRANSFORMER MODELS IN FORECASTING AND CLASSIFICATION OF TIME SERIES

Transformer models have brought significant advancements in time series analysis due to their ability to capture long-term dependencies. This study examines the fundamental structure, strengths, and limitations of transformer models (deep learning architectures) in time series modeling. It presents various transformer variants for time series and discusses their applications in forecasting, classification, and anomaly detection, as well as the explainability and interpretability.

Keywords: transformer, time series analysis, forecasting, classification, deep learning

Discipline: informatics, mathematics, data science

Boros Gerzson Dávid (2024): Transzformer modellek jelentősége az idősorok előrejelzésénél és osztályozásánál. *Mesterséges Intelligencia – interdiszciplináris folyóirat, VI. évf. 2024/1. szám.* 35-45. DOI: https://www.doi.org/10.35406/MI.2024.1.35

A transzformátor modelleknek nevezett mélytanulási architektúrák, forradalmasították a természetes nyelvfeldolgozást, a számítógépes látást, és jelentős érdeklődést váltottak ki az idősor-elemzéssel foglalkozók körében is. Előnyük, hogy hosszú távú függőségeket és interakciókat tudnak megragadni, ami különösen alkalmassá teszi őket az idősorok modellezésére, és jelentős előrelépéseket eredményeznek ezen a területen.

Az évek során nagy számú különféle transzformátor variánst fejlesztettek ki az idősorok modellezésében felmerülő specifikus kihívások kezelésére. Ezeket a modelleket sikeresen alkalmazták olyan feladatokban, mint az előrejelzés, az osztályozás és az anomália-felismerés. A transzformátorok bevezetése jelentősen javította az élvonalbeli teljesítményt ezeken a területeken.

Ez a tanulmány a transzformátor modellek alapvető felépítését, alkalmazhatóságát, erősségeit és korlátait vizsgálja az idősorok alkalmazásában. Először röviden bemutatásra kerül az első transformer modell működése, a gyakran alkalmazott modulok különböző transformer modell architektúrákban, a szóba jövő gépi tanulási feladatok és néhány alkalmazott modell architektúra. Végül megvitatjuk a transzformátorok értelmezhetőségét és magyarázhatóságát.

Vanilla Transformer

A vanilla transzformátor modell (Vaswani és tsai, 2017) egy kódoló-dekódoló struktúrát alkalmaz. A kódoló és a dekódoló mindegyike több azonos rétegből áll. Minden kódolóréteg egy többszörös fejű önfigyelő mechanizmust és egy pozíció szerinti előrecsatolt hálózatot tartalmaz. A dekóderben keresztfigyelő rétegek kerülnek a többszörös fejű önfigyelő mechanizmus és a pozíció szerinti előrecsatolt hálózat közé.

A jobb megértés céljából, most tekintsük át a transzformátor modellek összetevőit az idősor adatok kontextusában! Gyakran használt modulok az idősorok transzformátor architektúrájában:

Kódoló: A kódoló blokk többszörös fejű önfigyelésből (multi-head self-attention) és egy előrecsatolt rétegből (feed-forward layer) áll, maradék összeköttetésekkel és normalizációs rétegekkel együtt. A mély neurális hálózatok gyakran használnak maradék összeköttetéseket a képzés stabilitásának és tanulási folyamatának javítása érdekében (Szegedy és tsai, 2015). A réteg normalizáció, amely gyakori a szekvencia-feldolgozó neurális hálózatokban, felgyorsítja a képzés konvergenciáját (Ba és tsai, 2016). Az előrecsatolt réteg két lineáris rétegből és egy ReLU aktivációs

függvényből áll (Agarap, 2018). Az egyik kódoló kimenete a következő bemenete. Az első kódoló blokk bemeneteként a szóbeágyazások (word embeddings) és a pozíciós kódolás (PE) vektorainak összege szolgál.

Dekódoló: A dekódoló blokk hasonló az encoder blokkhoz, hasonló rétegekkel és műveletekkel, de két bemenettel rendelkezik: az egyik az előző dekódoló kimenetéből, a másik pedig az utolsó kódolóból származik. Három réteget tartalmaz: többszörös fejű önfigyelés, kódoló-dekódoló figyelés és egy előrecsatolt réteg, maradék összeköttetésekkel és réteg normalizációval. A kódoló-dekódoló figyelem réteg az utolsó kódoló kimenetét használja kulcs és érték vektorok létrehozásához, valamint az előző többszörös fejű önfigyelem réteg lekérdezési vektorait. Egyes idősor alkalmazásokban a dekódoló teljesen kihagyható és a következő szakaszokban tárgyalt sűrű interpolációs rétegre helyezhető át.

Kódoló és dekóder halmozás: Egy transzformátor modell több egymásra rakott kódoló és dekódoló blokkból állhat, a probléma természetétől függően (Tschannen, Bachem és Lucic, 2018). Vizuálisan ezek a rétegek egy neurális hálózat rétegeiként jelennek meg, tetszőleges számú rejtett réteggel, mind azonos reprezentációs dimenzióval, például amikor kódoló vagy dekódoló információt dolgozunk fel.

Pozíciós kódolás: Az időbeli adatokat pontosan kell feldolgozni, beleértve a szekvencia sorrendjét is, mivel az önfigyelés nem rendelkezik a bemeneti szekvencia sorrendjének belső ismeretével (Ahmed és tsai, 2023). Azaz, a Vanilla transzformátor nem rekurzív, ami azt jelenti, hogy pozíciós kódolást használ minden elem pozíciójának beágyazására a bemeneten, ahelyett, hogy explicit módon modellezné ezeket a szekvenciákat, mint az LSTM vagy RNN esetében. Ezért a transzformátorok párhuzamosan működhetnek. A pozíciós kódolás véletlenszerűen generál pozíció kódolt vektorokat, majd kiszámítja és összefűzi ezeket a vektorokat a bemenettel. A transzformátorok különféle típusú pozíciós kódolásokat használnak, melyekre most nem térünk ki.

Figyelem modul: A transzformátor központi eleme az önfigyelem modul, amely egy teljesen összekapcsolt réteg, ahol a súlyokat – vagy azokhoz hasonlókat – a pont-pont hasonlóság alapján generálják, és ezek az alsóbb réteg nem normalizált súlyainak függvényei. Ez lehetővé teszi számunkra a hosszú távú függőségek megragadását, hasonlóan a teljesen összekapcsolt rétegekhez, de sokkal kevesebb paraméterrel.

A Vanilla Transformer például idő- és térkomplexitása O(N^2) (ahol N az idősorok bemeneti dimenziója), ami hosszú szekvenciák esetén számítási szempontból nem megvalósítható. Számos hatékony transzformátort javasoltak ennek a kvadratikus komplexitásnak a csökkentésére, amelyek két fő kategóriába sorolhatók (Li és tsai, 2019; Liu és tsai, 2021; Zhou és tsai, 2021; Zhou és tsai, 2022):

- Ritkaság torzítás hozzáadása a figyelem mechanizmushoz, ezzel nem minden tokenre figyel a modul
- Az önfigyelem mátrix alacsony rangú természetének kihasználása a számítási igény csökkentése érdekében

Néhány munka átalakítja a transzformátor architektúráját az idősorok modellezésére. Például az idősorok többszintű természetének kezelésére a közelmúltban bemutattak hierarchikus architektúrákat. Az Informer (Zhou és tsai, 2021) max-pooling rétegeket ad hozzá 2-es lépéssel az önfigyelem blokkok közé, hogy felére csökkentse a szekvencia hosszát. A Pyraformer (Liu és tsai, 2021) egy C-ary fa alapú figyelem mechanizmust alkalmaz, ahol a különböző szintű csomópontok különböző felbontású szekvenciáknak felelnek meg, és segítenek a modelleknek jobban feltárni az időbeli függőségeket. A hosszú idősorok is számítási szempontból hatékonyak hierarchikus architektúrákkal.

Konvolúció: A transzformátor architektúra nem használ konvolúciós rétegeket, azonban a konvolúciós rétegek hozzáadása kifizetődő lehet az idősorok előrejelzése esetében. Konvolúció alkalmazható, akár az önfigyelem mechanizmus előtt, akár közvetlenül vele együtt, és sok idősorokra szánt transzformátor hasonló technikákat alkalmaz. Például TCCT (Shen és Wang, 2022), LogSparse Transformers (Li és tsai, 2019), TabAConvBERT (Shankaranarayana és Runje, 2021), Traffic Transformers (Cai és tsai, 2020). A konvolúciókat általában rövid távú függőségek vagy térbeli függőségek megragadására használják.

Sűrű interpolációs algoritmus: A transzformátor rétegek, a transzformátor megvalósításokban, kódoló és dekódoló blokkokkal vannak kitöltve, lineáris és softmax rétegekkel, amelyek döntéseket hoznak. A Simply Attend and Diagnose (SAnD) megközelítés sűrűbb interpolációt biztosít azzal, hogy a dekódoló blokkot egy sűrű interpolációs réteggel helyettesíti, ami segíti az időbeli sorrend figyelését (Song és tsai, 2018). Ennek eredményeként csökkenthető a rétegek száma az output embedding kihagyásával. Ebben az esetben a modell közvetlenül a kódoló blokk kimenetét lineáris rétegekre fordítja le, mielőtt a végső eredményt osztályként adná meg. Egy naiv összefűzés rossz predikciós teljesítményhez vezetne, ezért egy sűrű interpolációs algoritmust alkalmaznak hangolható hiperparaméterekkel a predikció javítása érdekében.

A Transformer modellek alkalmazhatósága különböző gépi tanulási feladatok esetében

A transzformátor modellek például megoldhatják az idősor előrejelzését, segíthetik a téridő előrejelzést, az esemény előrejelzést, az osztályozást és anomália észlelést.

Előrejelzés: Az idősorok előrejelzésének egyik legelterjedtebb felhasználási esete különböző valós élethelyzetekben fordul elő, például: mezőgazdasági termelés előrejelzése, kórházi kapacitás előrejelzése vagy éppen gyártási kapacitás előrejelzése. Az idősor előrejelzés a domináns feladat idősor adatok esetében, ezért több modellvariánst is bemutatunk, míg

a többi esetben kifejezetten koncentrálunk a jelenleg elérhető legjobb eredményeket biztosító modellekre.

Az élvonalbeli előrejelzési módszerek, amelyek pontos előrejelzéseket tudnak generálni nagyon hosszú szekvenciákra, olyan modelleken alapulnak, amelyek képesek hosszú távú függőségeket modellezni a bemenetek és a hosszú kimeneti szekvenciák között. Azonban a kvadratikus időkomplexitás, a túlzott memóriafogyasztás és a kódoló-dekódoló szerkezet hibái megakadályozzák, hogy a modell közvetlenül hosszú szekvenciákat jósoljon. A nagy teljesítményű, transzformátor-szerű modell, az Informer (2021) három fő mechanizmust alkalmaz ezen problémák kezelésére:

1) Egy figyelem mechanizmus a ProbSparse szinten az önfigyelem szintjén, amely támogatja a szekvenciafüggőségi igazítást (és így optimális O(L \log L) idő- és térkomplexitást).

2) Önálló figyelés tömörítése, amely előnyben részesíti a figyelmet, hogy a bemenetet felére csökkentse a kaszkádoló réteg számára, és így jobban kezelje a nagyon hosszú bemeneti szekvenciákat.

3) Egy generatív, batch-módban működő dekódoló, amely egyetlen előre haladással képes hosszú idősor szekvenciákat előre jelezni, jelentősen javítva az előrejelzési skálázhatóságot.

Négy nagyméretű adathalmazon végzett kísérleti eredmények azt mutatják, hogy az Informer felülmúlja a korábban elérhető módszereket, megerősítve a javasolt megoldás hatékonyságát a hosszú idősor szekvencia adatok előrejelzésében.

Az idősorok esetében a transzformátorok gyakran nehezen tudják megragadni a globális mintázatokat (például az összes időpont trendjeit). Ezt a FEDformer (Zhou és tsai, 2022) kezeli, amely a transzformátort szezonális-trend dekompozíciós módszerrel kombinálja. Bár a dekompozíciós módszer megragadja az idősorok általános profilját, a transzformátorok inkább a részletes struktúrákra koncentrálnak. Továbbá, a FEDformer kihasználja a legtöbb idősor ritka természetét a standard bázisokban, például a Fourier-transzformációban, hogy egy frekvenciafokozott transzformátort fejlesszen ki. Ez nemcsak pontosabb, hanem számítási szempontból is hatékonyabb, mint a Vanilla transzformátor, mivel lineáris a szekvencia hosszában. Hat benchmark adathalmazon végzett kísérleti eredmények azt mutatják, hogy a FEDformer 14,8%-os csökkenést ér el a többváltozós idősorok és 22,6%-os csökkenést az egyváltozós idősorok előrejelzési hibájában a korábbi módszerekhez képest.

A PatchTST (Nie és tsai, 2022) megmutatta, hogy lehetséges a számítási és memóriaigények csökkentése, valamint a hosszabb előzményekkel való munkavégzés, hogy jobb teljesítményt érjen el. Két fő elemet tartalmaz:

• Idősorok szegmentálása kisebb darabokra (patch) alszekvencia szinten, amelyek bemeneti tokenekké formálják a transzformátort.

• Csatorna-függetlenség; minden csatorna egyetlen egyváltozós idősor, közös beágyazással és transzformátor súlyokkal az összes sorozat között.

Ez a daraboló kialakítás három belső előnnyel rendelkezik: segít megőrizni a helyi infor-

mációkat a beágyazásban, lineárisan csökkenti a figyelem térképek számítási és memóriaigényét, és lehetővé teszi a modell számára, hogy szélesebb figyelem ablakot használjon. A PatchTST jelentősen javíthatja a hosszú távú előrejelzési teljesítményt a korábbi transzformátoron alapuló modellekhez képest.

Az időpontok közötti kapcsolat elsősorban numerikus, nem pedig szemantikai természetű. A kutatók ezzel magyarázzák, hogy egyszerű lineáris rétegek teljesítményben és hatékonyságban gyakran felülmúlják a Transformer modelleket. Pedig valószínűleg a probléma nem a modellre, hanem annak a nem optimális használatára vezethető vissza. Az iTransformer (Liu és tsai, 2023) minden idősorozatot variáns tokenekként ágyaz be. Egy előrecsatolt hálózatot használ a sorozat reprezentációjára, valamint többváltozós korrelációkra alkalmaz figyelmet.

Térbeli és időbeli előrejelzés: A tér-időbeli előrejelzés során mind az időbeli, mind a tér-időbeli függőségeket figyelembe kell venni az időszakos transzformátorokban a pontos előrejelzések elérése érdekében. Bizonyos modellek, mint például a Traffic Transformer (Cai és tsai, 2020), egy kódoló-dekódoló struktúrával rendelkeznek, amely egy önfigyelmi modult használ az időpontok közötti időbeli kapcsolatok megragadására, miközben egy gráf neurális hálózatot is beépítenek a helyek közötti térbeli kapcsolatok figyelembevételére.

Egy másik megközelítés, a Tér-időbeli Gráf Transzformátor (Yu és tsai, 2020), egy figyelemalapú gráf konvolúciós mechanizmust tartalmaz, amely képes bonyolult idő-térbeli figyelmi mintázatokat megtanulni a gyalogosok pályáinak előrejelzésének javítása érdekében.

Az Earthformer (Gao és tsai, 2022) a kuboid figyelmet vezeti be a hatékony tér-idő modellezés érdekében. Az adatokat cella kuboidokra bontja, és egyidejűleg alkalmaz cella-szintű önfigyelmet az egy kuboid összes cellájára. Az Earthformer kiváló teljesítményt nyújt időjárási és éghajlati előrejelzésekben.

A közelmúltban az AirFormer (Liang és tsai, 2023) egy dartboard térbeli önfigyelmi modult és egy oksági időbeli önfigyelmi modult dolgozott ki, amelyek hatékonyan képesek megragadni a térbeli korrelációkat és az időbeli függőségeket. Továbbá a transzformátorokat látens változókkal bővíti, hogy képviselje az adatok bizonytalanságát és javítsa a levegőminőség előrejelzését.

Az akciófelismerés hagyományosan az emberi tevékenységek osztályozására összpontosított videókban, és olyan területeken talált alkalmazásra, mint az ember-robot interakció, az egészségügy és a videómegfigyelés. A mélytanulás előrehaladásával az akciófelismerés három megközelítésre oszlott: videóalapú, csontvázalapú és több forrást kombináló (keresztmodális) megközelítések. Egy új módszer, a STAR-transformer (Ahn és tsai, 2023), ötvözi a keresztmodális tanulást és a tér-időbeli keresztezett transzformátor figyelmi mechanizmusát, és kiváló teljesítményt mutatott különféle kísérletek során.

Esemény előrejelzés: Az időben szabálytalanul mintavételezett adatsorok rendkívül gyakoriak különféle valós alkalmazásokban, mint például

a mezőgazdasági termés, a hurrikán vagy akár megbetegedések előrejelzése. Az ilyen típusú adatsorok, ellentétben a szabályos időbeli adatsorokkal, szabálytalan időközönként kerülnek megfigyelésre, jellegükből fakadóan. Emellett folyamatos, összetett kapcsolatokat mutatnak, amelyeket nehéz megérteni a hagyományos statisztikai megközelítésekkel. Ezek a tulajdonságok azt jelentik, hogy a hagyományos időbeli adatelemzési módszerek nehezen birkóznak meg ezzel az adattípussal. Az egyik legújabb és legpontosabb megoldás erre a problémára a ContiFormer (Chen és tsai, 2024)., amely ötvözi a neurális ODE-k (neural ordinary differential equations) folyamatos dinamikáját a Transformers figyelem mechanizmusával. Ez a módszer hatékonyan modellezi az információszolgáltatás bonyolult, időbeli dinamikáját - és pontosabbnak bizonyul, mint más modellek.

Osztályozás: Idősor osztályozási feladatok elég gyakoriak a mindennapokban, ilyen például az EKG vagy EEG jelek osztályozása a szív- és agyi rendellenességek felismerésére. A ConvTran (Foumani és tsai, 2024) az egyik legújabb és legkorszerűbb többváltozós időbeli transformer osztályozási modell. A pozíció enkódolás idősorok esetében nem elég mélyen kutatott, ebben hozott áttörést a ConvTran, amely egyesíti az időbeli abszolút pozíciókódolást (tAPE), a hatékony relatív pozíciókódolást (eRPE) és a konvolúció alapú bemeneti kódolást. A modell célja az időbeli adatok pozíció- és adatbeágyazásának javítása. A tAPE figyelembe veszi a sorozat hosszát és a bemeneti beágyazási dimenziót, míg az eRPE növeli az időbeli adatok általánosítási képességét. A ConvTran könnyen integrálható transformer blokkokba, és hatékony osztályozáson túl más időbeli feladatokhoz, például előrejelzés, regresszió és rendellenes mintázatok észlelése esetén. A kísérletek azt mutatták, hogy a ConvTran modell jelentősen jobb, mint a konvolúciós és transformer-alapú modellek, mindezt 32 időbeli adatsorokat tartalmazó adatbázison tesztelték.

Anomália detekció: Az idősorok esetében az anomália detekció egy gyakori feladat. Például, az élettani jelek (mint a szívritmus vagy vérnyomás) folyamatos monitorozása szükséges ahhoz, hogy időben felismerjük a potenciális egészségügyi vészhelyzeteket. Az Energy Transformer (Hoover és tsai, 2024) egy értelmező modellt javasol a Zajcsökkentő Variációs Transformer (DVT) használatával. Ez a DVT egy új figyelmi mechanizmust alkalmaz, amely a figyelmi súlyokat feltételes Gauss-eloszlásokként kezeli, azaz képes megtanulni az adateloszlásokat. A modell egy felügyelet nélküli anomália detektor, amely egy kódoló-dekódoló formulációt használ a bemenet rekonstrukciójára, mielőtt összehasonlítaná a maradékot az eredeti bemenettel, és eldöntené az anomáliákat. Továbbá, egy maradék értelmezőt fejlesztettek ki annak érdekében, hogy megbecsüljék, az egyes jellemzők milyen mértékben járulnak hozzá az anomália pontszámhoz. A tesztelések során a DVT modell jobb teljesítményt nyújtott más felügyelet nélküli anomália felismerési módszerekhez képest mind az F1, mind a görbe alatti terület pontszámokban. A modell értelmezhetősége

jobb volt más hasonló módszereknél, miközben drámaian csökkentette a számítási időt.

Értelmezés és magyarázhatóság

Sok esetben egy modell előrejelzéseit értelmező és magyarázó módszer utólagos, ami azt jelenti, hogy a modell előrejelzéseinek elkészítése után nyújtanak magyarázatokat. Ezek az utólagos megközelítések szinte bármely modellre alkalmazhatók, gyakran vizuálisan tetszetős eredményeket hozva. Azonban előfordulhat, hogy nem pontosan tükrözik a modell belső működését (Nielsen és tsai, 2022).

Ezzel szemben léteznek olyan módszerek, amelyek a magyarázatokat és az értelmezhetőséget közvetlenül a modellbe építik, nem pedig utólagos közelítésekre támaszkodnak. Számos idősorokra vonatkozó transzformátort fejlesztettek ki beépített értelmezhetőséggel (Tipirneni és Reddy, 2021; Liu és tsai, 2021; Lim és tsai, 2021). Ezek a modellek képesek olyan magyarázatokat generálni, amelyek növelik az eredmények értelmezhetőségét és nagyobb felhasználói bizalmat építenek.

Jövőbeni fejlesztési lehetőségek

A jövőbeli kutatásoknak arra kell összpontosítaniuk, hogy a Transformer modelleket az idősorteljesítmény adatokhoz igazítsák, integrálva az induktív torzításokat (Zhou és tsai, 2021) és a GNN-ekkel való kombinációt a pontosabb tér-idő modellezés érdekében (Yu és tsai, 2020). Emellett fontos fejleszteni az előtanított modelleket (Zerveas és tsai, 2021), és alkalmazni az architektúra szintű újításokat (Wu és tsai, 2021) és a neurális architektúra keresést (NAS) az optimalizált teljesítmény érdekében (Elsken és tsai, 2019). Ezek az előrelépések együttesen jelentősen javíthatják a Transformer modellek hatékonyságát és alkalmazhatóságát az idősorteljesítmény-elemzések terén.

Konklúzió

A transzformer modellek innovatív megközelítést jelentenek az idősor-elemzésben. A tanulmány röviden bemutatja a transzformer modellek alapvető építőelemeit, erősségeit és gyengeségeit, azok különböző kiterjesztéseit és az elért eredmények mértékét más korszerű technológiákhoz képest. Látható, hogy a transzformerek alkalmazása nemcsak nagyobb előrejelzési pontossághoz vezethet, hanem az adatok értelmezését és magyarázhatóságát is javíthatja. Emellett fontos kiemelni, hogy ez a terület még mindig nagyon új és fejlődés alatt áll, így jelentős lehetőségek nyílnak további kutatásokra és fejlesztésekre.

Irodalom

Agarap, A. F. (2018). *Deep learning using rectified linear units (relu).* arXiv preprint arXiv:1803.08375.

Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Ramachandran, R. P., & Rasool, G. (2023). Transformers in time-series analysis: A tutorial. *Circuits, Systems,*

and Signal Processing, 42(12), 7433-7466. DOI: https://doi.org/10.1007/s00034-023-02454-8

Ahn, D., Kim, S., Hong, H., & Ko, B. C. (2023). Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3330-3339). DOI: https://doi.org/10.1109/WACV56688.2023.00333

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer normalization*. arXiv preprint arXiv:1607.06450.

Cai, L., Janowicz, K., Mai, G., Yan, B., & Zhu, R. (2020). Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting. *Transactions in GIS*, 24(3), 736-755. DOI: https://doi.org/10.1111/tgis.12644

Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., & Li, D. (2024). ContiFormer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems*, 36.

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55), 1-21. DOI: https://doi.org/10.1007/978-3-030-05318-5_11

Foumani, N. M., Tan, C. W., Webb, G. I., & Salehi, M. (2024). Improving position encoding of transformers for multivariate time series classification. *Data Mining and Knowledge Discovery*, 38(1), 22-48. DOI: https://doi.org/10.1007/s10618-023-00948-2

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y. B., Li, M., & Yeung, D. Y. (2022). Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35, 25390-25403.

Hoover, B., Liang, Y., Pham, B., Panda, R., Strobelt, H., Chau, D. H., ... & Krotov, D. (2024). Energy transformer. *Advances in Neural Information Processing Systems*, 36.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y. X., & Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y. X., & Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32.

Liang, Y., Xia, Y., Ke, S., Wang, Y., Wen, Q., Zhang, J., ... & Zimmermann, R. (2023, June). Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 12, pp. 14329-14337). DOI: https://doi.org/10.1609/aaai.v37i12.26676

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of*

Forecasting, 37(4), 1748-1764. DOI: https://doi.org/10.1016/j.ijforecast.2021.03.012

Liu, M., Ren, S., Ma, S., Jiao, J., Chen, Y., Wang, Z., & Song, W. (2021). *Gated transformer networks for multivariate time series classification*. arXiv preprint arXiv:2103.14438.

Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., & Dustdar, S. (2021, May). Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*.

Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., & Dustdar, S. (2021, May). Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*.

Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., & Dustdar, S. (2021, May). Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2023). *iTransformer: Inverted transformers are effective for time series forecasting*. arXiv preprint arXiv:2310.06625.

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). *A time series is worth 64 words: Long-term forecasting with transformers*. arXiv preprint arXiv:2211.14730.

Nielsen, I. E., Dera, D., Rasool, G., Ramachandran, R. P., & Bouaynaya, N. C. (2022). Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4), 73-84. DOI: https://doi.org/10.1109/MSP.2022.3142719

Shankaranarayana, S. M., & Runje, D. (2021, December). Attention augmented convolutional transformer for tabular time-series. In *2021 International Conference on Data Mining Workshops (ICDMW)* pp. 537-541. IEEE. DOI: https://doi.org/10.1109/ICDMW53433.2021.00071

Shen, L., & Wang, Y. (2022). TCCT: Tightly-coupled convolutional transformer on time series forecasting. *Neurocomputing,* 480, 131-145. DOI: https://doi.org/10.1016/j.neucom.2022.01.039

Song, H., Rajan, D., Thiagarajan, J., & Spanias, A. (2018, April). Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1). DOI: https://doi.org/10.1609/aaai.v32i1.11635

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9). DOI: https://doi.org/10.1109/CVPR.2015.7298594

Tipirneni, S., & Reddy, C. K. (2021). *Self-supervised transformer for multivariate clinical time-series with missing values*. arXiv preprint arXiv:2107.14293.

Transactions in GIS,24(3), 736-755.

Tschannen, M., Bachem, O., & Lucic, M. (2018). *Recent advances in autoencoder-based representation learning*. arXiv preprint arXiv:1812.05069.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, H., Xu, J., Wang, J., Long, M., Wang, X., & Jiang, J. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information*

Yu, C., Ma, X., Ren, J., Zhao, H., & Yi, S. (2020). Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision-ECCV 2020: 16th European Conference, Glasgow*, UK, August 23-28, 2020, Proceedings, Part XII 16 (pp. 507-523). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-58610-2_30

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021, May). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 12, pp. 11106-11115). DOI: https://doi.org/10.1609/aaai.v35i12.17325

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022, June). Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning* (pp. 27268-27286). PMLR.