GMAN: графовая сеть множественного внимания для прогнозирования трафика

# Стандарт Чжэн1,2,3, Сяолян Фань1,2,3∗, Чэн Ван1,2,3, Цзяньчжун Ци4

1Фуцзяньская ключевая лаборатория сенсорики и вычислений для умных городов, Сямыньский университет, Сямынь, Китай

2Digital Fujian Institute of Urban Traffic Big Data Research, Сямыньский университет, Сямынь, Китай

3Школа информатики, Сямыньский университет, Сямынь, Китай

4Школа вычислительной техники и информационных систем, Мельбурнский университет, Мельбурн, Австралия zhengchuanpan@stu.xmu.edu.cn, {fanxiaoliang, cwang}@xmu.edu.cn, jianzhong.qi@unimelb.edu.au

Абстрактный

Долгосрочное прогнозирование трафика является очень сложной задачей из-за сложности транспортных систем и постоянно меняющегося характера многих влияющих факторов. В этой статье мы сосредоточимся на пространственно-временных факторах и предложим графовую сеть множественного внимания (GMAN) для прогнозирования условий дорожного движения на временные шаги вперед в различных местах графа дорожной сети. GMAN адаптирует архитектуру энкодер-декодер, в которой и энкодер, и декодер состоят из нескольких пространственно-временных блоков внимания для моделирования влияния пространственно-временных факторов на условия движения. Кодировщик кодирует входные функции трафика, а декодер прогнозирует последовательность выходных данных. Между кодировщиком и декодером применяется преобразующий слой внимания для преобразования закодированных объектов трафика для создания последовательных представлений будущих временных шагов в качестве входных данных декодера. Механизм преобразования внимания моделирует прямые связи между историческими и будущими временными шагами, что помогает смягчить проблему распространения ошибок среди временных шагов прогнозирования. Экспериментальные результаты по двум реальным задачам прогнозирования трафика (т.е. прогнозированию объема трафика и прогнозированию скорости трафика) демонстрируют превосходство GMAN. В частности, в прогнозировании на 1 час вперед GMAN превосходит современные методы на 4% по показателю MAE. Исходный код доступен по адресу https://github.com/zhengchuanpan/GMAN.

# Знакомство

Прогнозирование дорожного движения направлено на прогнозирование будущих условий дорожного движения (например, объема или скорости движения) в дорожных сетях на основе исторических наблюдений (например, записанных с помощью датчиков). Он играет важную роль во многих реальных приложениях. Например, точное прогнозирование трафика может помочь транспортным агентствам лучше контролировать трафик для уменьшения заторов на дорогах (Lv et al. 2018; Zheng et al. 2019).

Ожидается, что условия дорожного движения в близлежащих местах будут влиять друг на друга. Для выявления таких пространственных корреляций  *широко используются сверточные нейронные сети* (CNN) (Zhang, Zheng, and Qi 2017; Yao et al. 2018; 2019). Между тем, дорожная обстановка в месте также коррелирует с его историческими наблюдениями. *Рекуррентные нейронные сети* (РНС)

arXiv:1911.08415v2 [eess. SP] 26 ноя 2019

∗

Автор, отвечающий за переписку

Copyright 2020, Ассоциация по развитию искусственного интеллекта (www.aaai.org). Все права защищены.

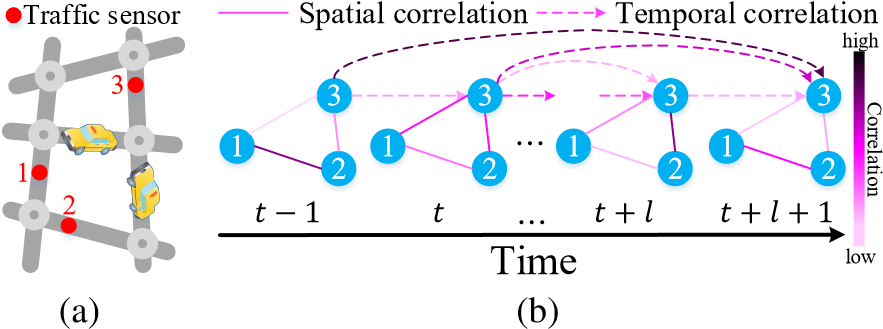


Рисунок 1: Сложные пространственно-временные корреляции. а) Датчики в дорожной сети. b) динамические пространственные корреляции: датчики 1 и 2 не всегда тесно коррелируют, хотя и находятся близко друг к другу в дорожной сети; Нелинейные временные корреляции: состояние трафика датчика 3 на временном шаге *t* + *l* +1 может быть в большей степени коррелировано с дальними временными шагами (например, *t* − 1), чем с недавними временными шагами (например, *t* + *l*).

широко применяется для моделирования таких временных корреляций (Ma et al. 2015; Song, Kanasugi, and Shibasaki 2016).

В недавних исследованиях прогнозирование трафика формулируется как задача графического моделирования, поскольку условия движения ограничены на графах дорожной сети (Li et al. 2018b; Юй, Инь и Чжу 2018; Wu et al. 2019b). Используя *графовые сверточные сети* (GCN) (Defferrard, Bresson, and Vandergheynst 2016), эти исследования достигают многообещающих результатов для краткосрочного прогнозирования трафика (на 5∼15 минут вперед). Тем не менее, долгосрочное (с точностью до нескольких часов вперед (Hou and Li 2016)) прогнозирование трафика все еще не получило удовлетворительного прогресса в литературе, в основном из-за следующих проблем: 1) *Сложные пространственно-временные корреляции.*

* Динамические пространственные корреляции. Как показано на рисунке 1, корреляция условий дорожного движения между датчиками в дорожной сети значительно меняется с течением времени (например, до и во время часов пик). Как динамически выбирать данные соответствующих датчиков для прогнозирования условий движения целевого датчика в долгосрочной перспективе — сложная задача.
* Нелинейные временные корреляции. Кроме того, на рисунке 1 состояние дорожного движения на датчике может сильно и внезапно колебаться (например, из-за аварии), влияя на корреляцию между различными временными шагами. Как адаптивно моделировать нелинейные временные корреляции, когда время уходит дальше в будущее, остается сложной задачей.

2) *Чувствительность к распространению ошибок.* На долгосрочном горизонте небольшие ошибки на каждом временном шаге могут усиливаться при составлении прогнозов на более отдаленное будущее. Такое распространение ошибок делает прогнозирование в далеком будущем очень сложным.

Для решения вышеупомянутых проблем мы предлагаем Graph *Multi-Attention Network* (GMAN) для прогнозирования условий дорожного движения на графе дорожной сети с течением времени на несколько шагов вперед. В данном случае условия дорожного движения относятся к наблюдениям за системой дорожного движения, которые могут быть представлены в числовых значениях. В качестве иллюстрации мы сосредоточимся на прогнозировании объема и скорости трафика, хотя наша модель может быть применена к прогнозированию других числовых данных о трафике.

GMAN следует архитектуре кодировщик-декодер, в которой кодировщик кодирует входные функции трафика, а декодер прогнозирует последовательность вывода. Между кодировщиком и декодером добавляется слой внимания преобразования для преобразования закодированных исторических объектов трафика для создания будущих представлений. Как энкодер, так и декодер состоят из стека *блоков ST-Attention*. Каждый блок ST-Attention формируется механизмом пространственного внимания для моделирования динамических пространственных корреляций, механизмом временного внимания для моделирования нелинейных временных корреляций и механизмом стробируемого слияния для адаптивного слияния пространственных и временных представлений. Механизм преобразования внимания моделирует прямые связи между историческими и будущими временными шагами, чтобы смягчить эффект распространения ошибок. Эксперименты на двух реальных наборах данных подтверждают, что GMAN достигает самых современных показателей.

Результаты этой работы можно резюмировать следующим образом:

* Предложены механизмы пространственного и временного внимания для моделирования динамических пространственных и нелинейных временных корреляций соответственно. Кроме того, мы разрабатываем стробированное слияние для адаптивного слияния информации, извлекаемой механизмами пространственного и временного внимания.
* Мы предлагаем механизм преобразования внимания для преобразования исторических объектов трафика в будущие представления. Этот механизм внимания моделирует прямые связи между историческими и будущими временными шагами, чтобы смягчить проблему распространения ошибок.
* Мы оцениваем нашу графовую сеть множественного внимания (GMAN) на двух реальных наборах данных о дорожном движении и наблюдаем улучшение на 4% и превосходную отказоустойчивость по сравнению с современными базовыми методами при прогнозировании на 1 час вперед.

# Связанная работа

Прогнозирование трафика Прогнозирование трафика широко изучалось в последние десятилетия. Подходы к глубокому обучению (например, долгосрочная кратковременная память (LSTM) (Ma et al. 2015)) демонстрируют более высокую производительность в улавливании временных корреляций в условиях дорожного движения по сравнению с традиционными методами временных рядов (например, авторегрессионная интегрированная скользящая средняя (ARIMA) (Makridakis and Hibon 1997)) и моделями машинного обучения (например, регрессия опорного вектора (SVR) (Wu, Ho, and Lee 2004), k-ближайший сосед (KNN) (Zheng and Su 2014)). Для моделирования пространственных корреляций исследователи применяют сверточные нейронные сети (СНС) для выявления зависимостей в евклидовом пространстве (Zhang, Zheng, and Qi 2017; Yao et al. 2018; 2019). В недавних исследованиях прогноз трафика формулируется на графах и используются графовые сверточные сети (РЦС) для моделирования неевклидовых корреляций в дорожной сети (Li et al. 2018b; Lv et al. 2018). Эти модели на основе графиков генерируют прогнозы на несколько шагов вперед с помощью пошагового подхода и могут страдать от распространения ошибок между различными шагами прогнозирования.

Глубокое обучение на графах Обобщение нейронных сетей на графово-структурированные данные является новой темой (Bronstein et al. 2017; Wu et al. 2019a). Ряд исследований обобщает СНС для моделирования произвольных графиков на спектральном уровне (Defferrard, Bresson, and Vandergheynst 2016; Kipf and Welling 2017; Li et al. 2018a) или пространственными (Atwood and Towsley 2016; Hamilton, Ying, and Leskovec 2017; Chen, Ma, and Xiao 2018). Еще одно направление исследований сосредоточено на встраивании графов, которое изучает низкоразмерные представления вершин, сохраняющие информацию о структуре графа (Grover and Leskovec 2016; Cui et al. 2019). (Wu et al. 2019b) интегрирует WaveNet (van den Oord et al. 2016) в GCN для пространственно-временного моделирования. По мере изучения статических матриц смежности этот метод сталкивается с трудностями при захвате динамических пространственных корреляций.

Механизм внимания Механизмы внимания широко применяются в различных областях благодаря их высокой эффективности и гибкости в моделировании зависимостей (Vaswani et al. 2017; Shen et al. 2018; Du et al. 2018). Основная идея механизмов внимания заключается в адаптивной концентрации на наиболее релевантных характеристиках в соответствии с входными данными (Cheng et al. 2018). В последнее время исследователи применяют механизмы внимания к структурированным графам данным (Velikovi et al. 2018) для моделирования пространственных корреляций для классификации графов. Мы расширяем механизм внимания на прогнозирование пространственно-временных данных на графике.

# Отборочные

Обозначим дорожную сеть в виде взвешенного ориентированного графа G = (V,E,A). Здесь V — это множество *N* = |V| вершины, представляющие точки (например, дорожные датчики) на дорожной сети; E — набор ребер, представляющих связность между вершинами, а A ∈ RN×N — взвешенная матрица смежности, где

Avi,vjпредставляет собой близость (измеряемую расстоянием по дорожной сети) между вершинами *vi* и *vj*.

Состояние трафика на временном шаге *t* представлено в виде графического сигнала *Xt* ∈ RN×C на графе G, где *C* — количество интересующих вас условий трафика (например, интенсивность движения, скорость движения и т. д.).

Изучаемая проблема Учитывая наблюдения в *N* вершинах исторических *временных* шагов P X = (*Xt* 1,Xt*2*,...,X*tP) ∈* RP×N×C, мы стремимся предсказать условия трафика следующих временных *шагов Q для всех вершин, обозначенных как* Yˆ *= (*XˆtP+1,XˆtP+2,...,X*ˆtP+Q) ∈* RQ×N×C.

# Графовая сеть множественного внимания

На рисунке 2 показана структура предложенной нами *графовой сети множественного внимания* (GMAN), которая имеет структуру энкодердекода. Как кодировщик, так и декодер содержат

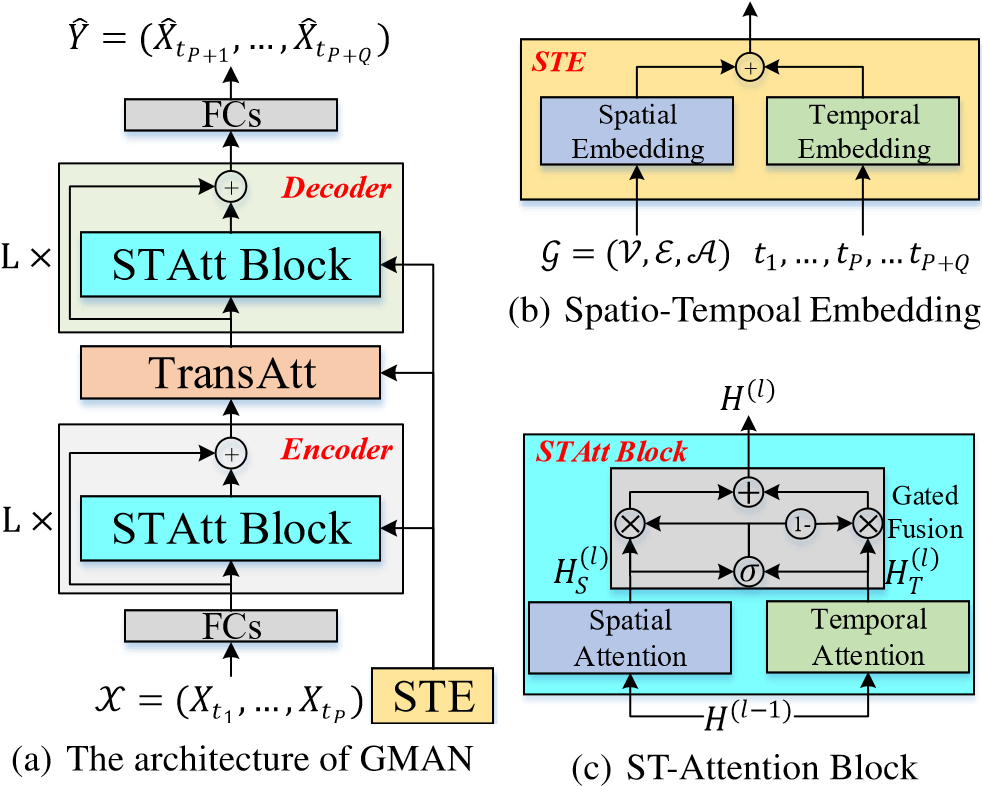


Рисунок 2: Структура сети множественного внимания Graph (GMAN). (a) GMAN состоит из пространственно-временного встраивания (STE), энкодера и декодера с блоками L STAttention (STAtt Block), преобразующим слоем внимания (TransAtt) и двумя полносвязными слоями (FC). (b) Пространственно-временное вложение включает пространственное вложение и временное вложение. (c) Блок ST-Attention сочетает в себе пространственные и временные механизмы внимания посредством стробируемого слияния.

*L* ST-блоки внимания (STAtt Block) с остаточными соединениями (He et al. 2016). Каждый блок ST-Attention состоит из пространственных и временных механизмов внимания с закрытым слиянием. Между кодировщиком и декодером в сеть добавляется преобразующий слой внимания для преобразования закодированных объектов трафика в декодер. Мы также включаем структуру графа и временную информацию в механизмы множественного внимания с помощью *пространственно-временного встраивания* (STE). Кроме того, чтобы облегчить остаточное соединение, все слои производят выходы размеров *D* . Модули подробно описаны ниже.

## Пространственно-временное встраивание

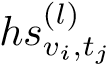
Поскольку эволюция условий дорожного движения ограничена базовой дорожной сетью (Lv et al. 2018), крайне важно включить информацию о дорожной сети в модели прогнозирования. С этой целью мы предлагаем пространственное *встраивание* для кодирования вершин в векторы, сохраняющие информацию о структуре графа. В частности, мы используем подход *node2vec* (Grover and Leskovec 2016) для изучения представлений вершин. Кроме того, для совместного обучения предварительно изученных векторов со всей моделью, эти векторы подаются в двухслойную полносвязную нейронную сеть. Тогда мы получаем пространственное вложение, представленное в виде, где *vi* ∈ V.

Пространственное встраивание обеспечивает только статические представления, которые не могут представлять динамические корреляции между датчиками дорожного движения в дорожной сети. Таким образом, мы далее предлагаем *временное встраивание* для кодирования каждого временного шага в вектор. В частности, пусть день будет с *T* временными шагами. Мы кодируем день недели и время суток каждого временного шага в R7 и RT с помощью одногорячего кодирования и объединяем их в вектор RT+7. Затем мы применяем двухслойную полносвязную нейронную сеть для преобразования временного признака в векторный RD. В нашу модель мы встраиваем временные признаки как для исторического P, так и для будущих временных шагов *Q*, представленных как *tj*  = *t*1*,...,tP,...,tP+Q*.

Чтобы получить изменяющиеся во времени представления вершин, мы объединяем вышеупомянутое пространственное вложение и временное вложение в пространственно-временное вложение (STE), как показано на рисунке 2(b). В частности, для вершины *vi* на временном шаге *tj*, метод

STE определяется как. Таким образом, STE *N* вершин с *временными шагами* P *+* Q представляется как *E* ∈ R(*P+Q*)×*N×D*. STE содержит как графическую структуру, так и временную информацию, и она будет использоваться в пространственных, временных и трансформационных механизмах внимания.

## ST-Блок внимания

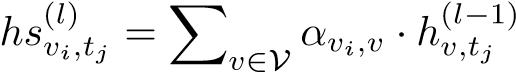
Как показано на рисунке 2(c), блок ST-Attention включает в себя пространственное внимание, временное внимание и стробированное слияние. Вход *l-го* блока обозначим как *H*(*l−1),* где скрытое состояние вершины *vi* на временном шаге *tj* представлено в виде. Выходы пространственных и временных механизмов внимания в *l-м* блоке представлены как and, где скрытые состояния вершины *vi* на временном шаге *tj* обозначаются как и

соответственно. После стробированного слияния мы получаем выход *l-го* блока, представленного как *H*(*l*).

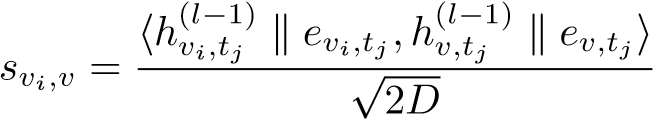
Для иллюстрации обозначим нелинейное преобразование как:

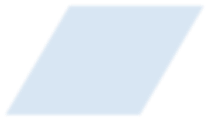
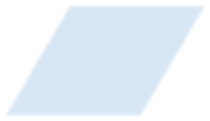
*f*(*x*) = ReLU(*xW+b*)*,* (1) где **W**, **b** – изучаемые параметры, а ReLU (Nair and Hinton 2010) – функция активации.

Пространственное внимание На состояние дороги влияют другие дороги с различными воздействиями. Такое воздействие очень динамично и меняется со временем. Чтобы смоделировать эти свойства, мы разрабатываем механизм пространственного внимания, чтобы адаптивно улавливать корреляции между датчиками в дорожной сети. Ключевая идея заключается в том, чтобы динамически присваивать разные веса различным вершинам (например, датчикам) на разных временных шагах, как показано на рисунке 3. Для вершины *vi* на временном шаге *tj* мы вычисляем взвешенную сумму по всем вершинам:

*,* (2)

где V обозначает множество всех вершин, *αvi,v* — оценка внимания, указывающая на важность вершины *v* до *vi*, а сумма оценок внимания равна 1: Pv∈V *αvi,v* = 1. На определенном временном этапе на корреляции между датчиками могут влиять как текущая дорожная обстановка, так и структура дорожной сети. Например, затор на дороге может существенно повлиять на условия движения на прилегающих к ней дорогах. Руководствуясь этой интуицией, мы учитываем как особенности трафика, так и структуру графика, чтобы узнать оценку внимания. В частности, мы объединяем скрытое состояние с пространственно-временным встраиванием и применяем подход масштабированного скалярного произведения (Vaswani et al. 2017) для вычисления релевантности между вершинами *vi* и *v*:

*,* (3)



α

1

,

1



3



1



2



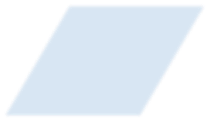
3



1



2



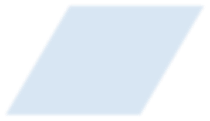
3



1



2



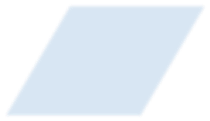
3



1



2



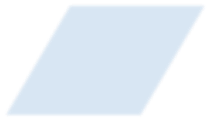
3



1



2



3



1



2

𝑡

𝑗

−

1

Время

𝑡

𝑗

−

2

𝑡

𝑗

𝐻

(

𝑙

−

1

)

𝐻

𝑆

(

𝑙

)

α

1

,

2

α

1

,

3

*...*

*...*

Рисунок 3: Механизм пространственного внимания фиксирует временные попарные корреляции между вершинами.

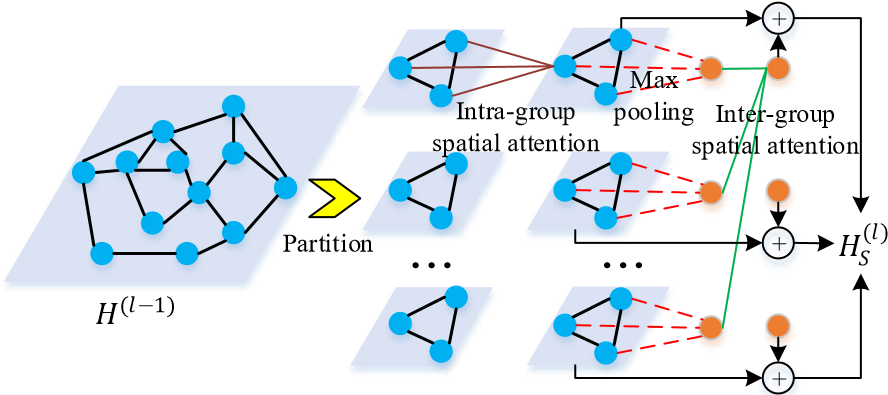
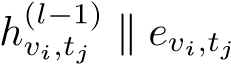
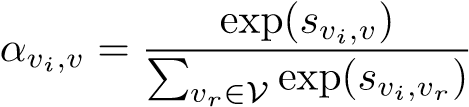


Рисунок 4: Групповое пространственное внимание вычисляет как внутригрупповое, так и межгрупповое внимание для моделирования пространственных корреляций.

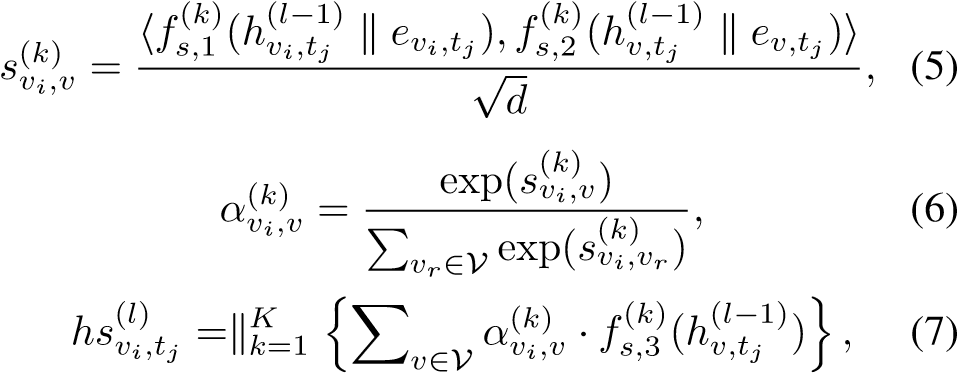
где k представляет операцию конкатенации, h•*,*•i обозначает внутреннего оператора произведения, а 2D — размерность

. Затем *svi,v*нормализуется через softmax как:

*.* (4)

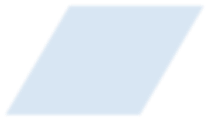
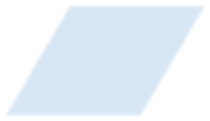
После того, как будет получена оценка внимания *αvi,v*, скрытое состояние может быть обновлено с помощью уравнения 2.

Чтобы стабилизировать процесс обучения, мы расширяем механизм пространственного внимания до многоголовых (Vaswani et al. 2017). В частности, мы объединяем *K* параллельных механизмов внимания с различными обучаемыми проекциями:



где и представляют три различные нелинейные проекции (уравнение 1) в *k-м* головном внимании, производящие *d* = *D/K* размерные выходы.

Когда количество вершин *N* велико, требуется много времени и памяти, так как нам нужно вычислить *N2* баллов концентрации внимания. Чтобы устранить это ограничение, мы также предлагаем групповое пространственное внимание, которое включает в себя внутригрупповое пространственное внимание и межгрупповое пространственное внимание, как показано на рисунке 4.



3



1



2



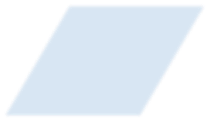
3



1



2



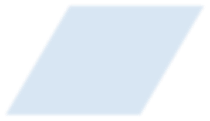
3



1



2



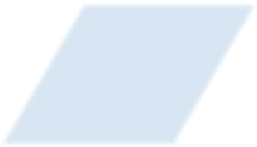
3



1



2



3



1



2



3



1



2

𝑡

𝑗

−

1

Время

𝑡

𝑗

−

2

𝑡

𝑗

𝐻

(

𝑙

−

1

)

𝐻

𝑇

(

𝑙

)

*...*

*...*

β

𝑡

𝑗

,

𝑡

𝑗

−

2

β

𝑡

𝑗

,

𝑡

𝑗

−

1

β

𝑡

𝑗

,

𝑡

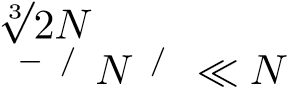
𝑗

Рисунок 5: Механизм временного внимания моделирует нелинейные корреляции между различными временными шагами.

Мы случайным образом разбиваем *N* вершин на *G* группы, где каждая группа содержит *M* = *N/G* вершин (при необходимости может быть применен отступ). В каждой группе мы вычисляем внутригрупповое внимание для моделирования локальных пространственных корреляций между вершинами с помощью уравнений 5, 6 и 7, где обучаемые параметры являются общими для всех групп. Затем мы применяем подход max-pooling в каждой группе, чтобы получить одно представление для каждой группы. Затем мы вычисляем межгрупповое пространственное внимание, чтобы смоделировать корреляции между различными группами, создавая глобальный признак для каждой группы. Локальный объект добавляется к соответствующему глобальному объекту в качестве конечного результата.

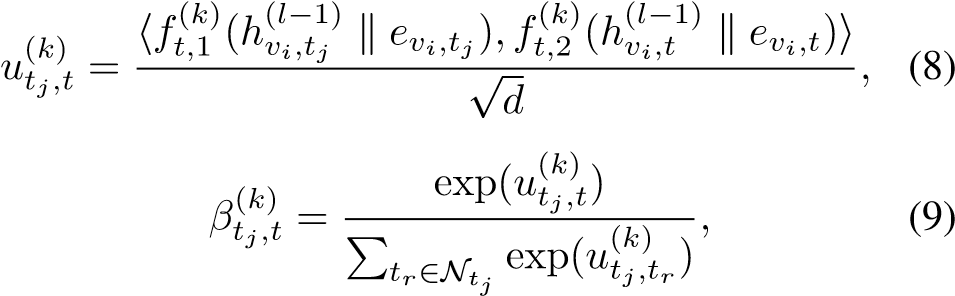
В групповом пространственном внимании нам нужно вычислить показатели *внимания GM2 + G2* = *NM* + (*N/M*)2 на каждом временном шаге.

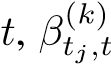
Приведя градиент к нулю, мы узнаем, когда *M* =

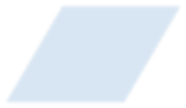
, количество оценок внимания достигает своего минимума

2 1 3 4 3 2.

Временное внимание Состояние дорожного движения в определенном месте коррелирует с его предыдущими наблюдениями, и корреляции изменяются с течением времени нелинейно. Чтобы смоделировать эти свойства, мы разрабатываем механизм временного внимания для адаптивного моделирования нелинейных корреляций между различными временными шагами, как показано на рисунке 5. Обратите внимание, что на временную корреляцию влияют как условия движения, так и соответствующий временной контекст. Например, затор, возникающий в утренние часы пик, может повлиять на движение в течение нескольких часов. Таким образом, мы учитываем как особенности трафика, так и время для измерения релевантности между различными временными шагами. В частности, мы объединяем скрытое состояние с пространственно-временным встраиванием и применяем многоголовый подход для вычисления оценки внимания. Формально, рассматривая вершину *vi*, корреляция между временным шагом *tj* и *t* определяется как:



где обозначает релевантность между временным шагом *tj* и  — оценка внимания в *k-й* голове, указывающая на важность временного шага и представляющая



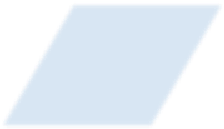
3



1



2



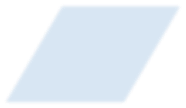
3



1



2



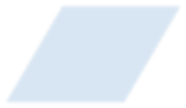
3



1



2



3



1



2

*...*

*...*

*...*

Время

𝐻

(

𝑙

−

1

)

𝑡

𝑗

𝑡

1

𝑡

𝑃

𝑡

𝐻

(

𝑙

)

γ

𝑡

𝑗

,

𝑡

1

γ

𝑡

𝑗

,

𝑡

γ

𝑡

𝑗

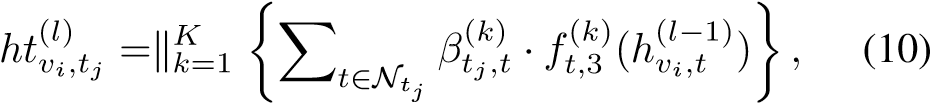
,

𝑡

𝑃

Рисунок 6: Механизм преобразования внимания моделирует прямые связи между историческими и будущими временными шагами.

два различных обучаемых преобразования, Ntj обозначает набор временных шагов перед *tj*, т.е. учитывает информацию только из временных шагов, предшествующих целевому шагу, чтобы обеспечить причинно-следственную связь. После того, как оценка внимания получена, скрытое состояние вершины *vi* на временном шаге *tj* обновляется следующим образом:



где представляет собой нелинейную проекцию. Обучаемые параметры в уравнениях 8, 9 и 10 являются общими для всех вершин и временных шагов с помощью параллельных вычислений.

Слияние ворот Дорожное состояние дороги на определенном временном шаге коррелирует как с ее предыдущими значениями, так и с условиями движения на других дорогах. Как показано на рисунке 2(c), мы разрабатываем стробированное слияние для адаптивного слияния пространственных и временных представлений. В *l-м* блоке выходы пространственного и временного механизмов внимания представлены как и, оба имеют формы RP×N×D в энкодере или RQ×N×D в декодере. и сплавляются как:

*,* (11)

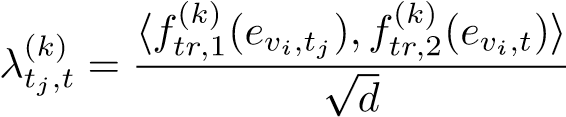
с

*,* (12)

где **Wz,1** ∈ RD×D, **Wz,2** ∈ RD×D и **bz** ∈ RD — изучаемые параметры, представляет произведение по элементам, *σ*(•) обозначает активацию сигмовидной кишки, *z* — вентиль. Механизм стробируемого слияния адаптивно управляет потоком пространственных и временных зависимостей на каждой вершине и временном шаге.

## Трансформация внимания

Чтобы уменьшить эффект распространения ошибок между различными временными шагами прогнозирования на длинном временном горизонте, мы добавляем преобразующий слой внимания между кодировщиком и декодером. Он моделирует прямую связь между каждым будущим временным шагом и каждым историческим временным шагом для преобразования закодированных объектов трафика для создания будущих представлений в качестве входных данных декодера. Как показано на рисунке 6, для вершины *vi* релевантность между временным шагом *предсказания tj* (*tj* = *tP+*1*,...,tP+Q*) и историческим временным шагом *t* (*t* = *t1,...,tP*) измеряется с помощью пространственно-временного встраивания:

*,* (13)

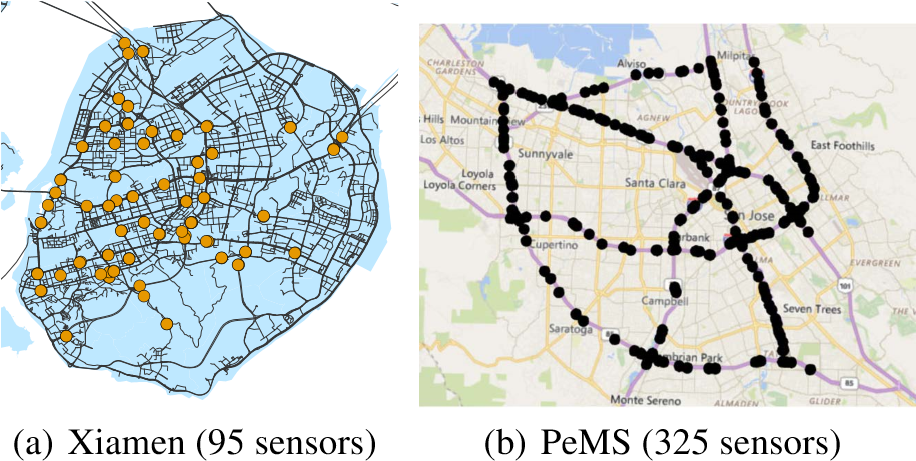
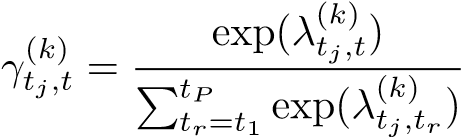
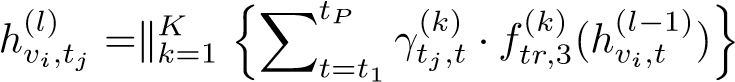


Рисунок 7: Распределение датчиков наборов данных Xiamen и PeMS.

*.* (14)

С помощью оценки внимания закодированный трафик преобразуется в декодер путем адаптивного выбора релевантных объектов на всех исторических *временных шагах* P:

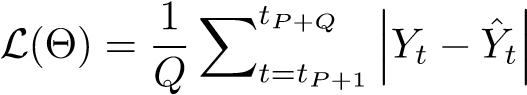
*.* (15)

Уравнения 13, 14 и 15 могут быть вычислены параллельно по всем вершинам и временным шагам, используя общие изучаемые параметры.

## Декодер

Как показано на рисунке 2(a), GMAN — это архитектура кодера-декодера. Перед вводом в энкодер историческое наблюдение X ∈ RP×N×C преобразуется в *H*(0) ∈ RP×N×D с помощью полносвязных слоев. Затем *H*(0) подается в энкодер с *помощью* блоков L ST-Attention и выдает на выходе *H*(*L*) ∈ RP×N×D. Вслед за энкодером добавляется слой внимания преобразования для преобразования закодированного признака *H*(*L*) для генерации будущего представления последовательности *H*(*L+1)* ∈ RQ×N×D. Затем декодер складывает блоки L STAttention на *H*(*L+1)* и выдает вывод в виде *H*(2L+1) ∈ RQ×N×D. Наконец, полностью связанные слои генерируют предсказание *Q* на несколько шагов вперед *Yˆ* ∈ RQ×N×C.

GMAN может быть обучен сквозным путем обратного распространения, сводя к минимуму *среднюю абсолютную ошибку* (MAE) между прогнозируемыми значениями и наземными истинами:

*,* (16)

где Θ обозначает все изучаемые параметры в GMAN.

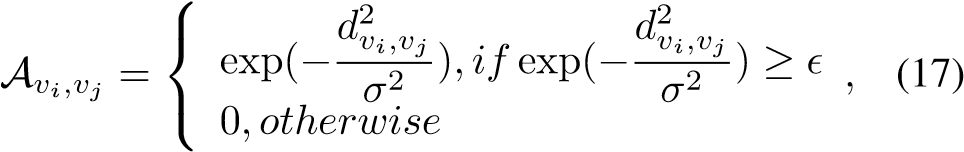
# Эксперименты

## Наборов данных

Мы оцениваем производительность GMAN в двух задачах прогнозирования трафика с различными масштабами дорожной сети: (1) прогнозирование объема трафика на наборе данных Xiamen (Wang et al. 2017), который содержит данные за 5 месяцев, записанные 95 датчиками дорожного движения в период с 1 августа 2015 года по 31 декабря 2015 года в Сямыне, Китай; (2) прогнозирование скорости движения на базе данных PeMS (Li et al. 2018b)), который содержит данные за 6 месяцев, записанные 325 датчиками дорожного движения в период с 1 января 2017 года по 30 июня 2017 года в районе залива. Распределение датчиков в двух наборах данных визуализировано на рисунке 7.

|  |
| --- |
| 15 мин 30 мин 1 час  Данные Метод  МАЕ RMSE МАПЕ МАЕ RMSE МАПЕ МАЕ RMSE МАПЕ  АРИМА 14.81 25.03 18.05% 18.83 33.09 22.19% 26.58 46.32 30.76%  СВР 13.05 21.47 16.46% 15.66 26.34 19.68% 20.69 35.86 26.24%  ФНН 13.55 22.47 16.72% 16.80 28.71 19.97% 22.90 39.51 26.19%  ФК-ЛСТМ 12.51 20.79 16.08% 13.74 23.93 17.23% 16.02 29.57 19.33%  СТГКН 11.76 19.94 14.93% 13.19 23.29 16.36% 15.83 29.40 18.66%  DCRNN 11.67 19.40 14.85% 12.76 22.20 15.99% 14.30 25.86 17.17%  График WaveNet 11.26 19.57 14.39% 12.06 21.61 15.39% 13.33 24.77 16.50%  ГМАН 11.50 19.52 14.59% 12.02 21.42 15.14% 12.79 24.15 15.84%  АРИМА 1.62 3.30 3.50% 2.33 4.76 5.40% 3.38 6.50 8.30%  СВР 1.85 3.59 3.80% 2.48 5.18 5.50% 3.28 7.08 8.00%  ФНН 2.20 4.42 5.19% 2.30 4.63 5.43% 2.46 4.98 5.89%  ПЭМС  ФК-ЛСТМ 2.05 4.19 4.80% 2.20 4.55 5.20% 2.37 4.96 5.70%  СТГКН 1.36 2.96 2.90% 1.81 4.27 4.17% 2.49 5.69 5.79%  DCRNN 1.38 2.95 2.90% 1.74 3.97 3.90% 2.07 4.74 4.90%  График WaveNet 1.30 2.74 2.73% 1.63 3.70 3.67% 1.95 4.52 4,63% GMAN 1.34 2.82 2.81% 1.62 3.72 3.63% 1.86 4.32 4.31%  Сямынь  Таблица 1: Сравнение производительности различных подходов к прогнозированию трафика на наборах данных Xiamen и PeMS. |

Предварительная обработка данных Мы применяем те же процедуры предварительной обработки данных, что и в (Li et al. 2018b). В обоих наборах данных временной шаг обозначает 5 минут, а данные нормализуются с помощью метода ZScore. Мы используем 70% данных для обучения, 10% — для валидации и 20% — для тестирования. Для построения графа дорожной сети каждый датчик дорожного движения рассматривается как вершина, и мы вычисляем попарные расстояния по дорожной сети между датчиками. Затем матрица смежности определяется следующим образом:



где dvi,vj *– расстояние по дорожной сети от датчика*  VI до VJ*,* σ *– стандартное отклонение, а (присвоено 0,1) – порог для управления разреженностью матрицы смежности* A.

## Экспериментальные настройки

Метрики Мы применяем три широко используемых показателя для оценки производительности нашей модели, а именно: *среднюю абсолютную ошибку* (MAE), *среднеквадратичную ошибку* (RMSE) и *среднюю абсолютную процентную ошибку* (MAPE).

Гиперпараметры Следуя предыдущим работам (Li et al. 2018b; Wu et al. 2019b), мы используем *P* = 12 исторических временных шагов (1 час) для прогнозирования условий трафика следующих *Q* = 12 шагов (1 час). Мы обучаем нашу модель с помощью оптимизатора Адама (Kingma and Ba 2015) с начальным коэффициентом обучения 0,001. При групповом пространственном внимании мы разделяем вершины на *G* = 19 групп в наборе данных Xiamen и *G* = 37 групп в наборе данных PeMS соответственно. Количество условий трафика в обоих наборах данных равно *C* = 1. Всего в нашей модели имеется 3 гиперпараметра: количество блоков STAttention *L*, количество голов внимания *K* и размерность *d* каждой головы внимания (канал каждого слоя *D* = *K×d*). Мы настраиваем эти параметры на проверочном наборе и наблюдаем наилучшую производительность на этом параметре

*L* = 3, *K* = 8 и *d* = 8 (*D* = 64).

Мы сравниваем GMAN со следующими базовыми методами: (1) Авторегрессионная интегрированная скользящая средняя (ARIMA) (Makridakis and Hibon 1997); (2) регрессия опорных векторов (УВО) (Wu, Ho, and Lee 2004); (3) Нейронная сеть с прямой связью (FNN); (4) FC-LSTM (Sutskever, Vinyals and Le 2014), который представляет собой модель последовательности с полностью соединенными слоями LSTM как в энкодере, так и в декодере; (5) Пространственно-временная сверточная сеть графа (STGCN) (Yu, Yin, and Zhu 2018), которая сочетает в себе сверточные слои графа и слои обучения сверточных последовательностей; (6) диффузионная сверточная рекуррентная нейронная сеть (DCRNN) (Li et al. 2018b), которая интегрирует диффузионную свертку с архитектурой последовательности-последовательность; (7) Граф WaveNet (Wu et al. 2019b), который сочетает в себе свертку графа с расширенной случайной сверткой.

Для моделей ARIMA, SVR, FNN и FC-LSTM мы используем настройки, предложенные (Li et al. 2018b). Для моделей STGCN, DCRNN и Graph WaveNet мы используем настройки по умолчанию из их исходных предложений.

## Экспериментальные результаты

В таблице 1 приведено сравнение различных методов прогнозирования на 15 минут (3 шага), 30 минут (6 шагов) и на 1 час (12 шагов) вперед в двух наборах данных. Мы наблюдаем, что: (1) подходы к глубокому обучению превосходят традиционные методы временных рядов и модели машинного обучения, демонстрируя способность глубоких нейронных сетей моделировать нелинейные данные о дорожном движении; (2) среди методов глубокого обучения модели на основе графов, включая STGCN, DCRNN, Graph WaveNet и GMAN, как правило, работают лучше, чем FC-LSTM, указывая на дорожную сеть.

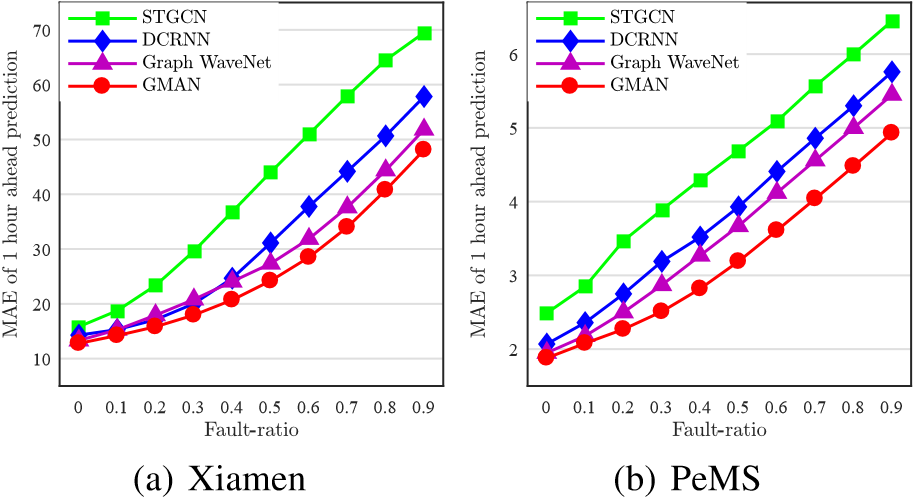


Рисунок 8: Сравнение отказоустойчивости.

информация о работе имеет важное значение для прогнозирования трафика; и (3) GMAN обеспечивает современную производительность прогнозирования, и преимущества более очевидны в долгосрочной перспективе (например, на 1 час вперед). Мы утверждаем, что долгосрочное прогнозирование трафика более полезно для практических приложений, например, оно позволяет транспортным агентствам иметь больше времени для принятия мер по оптимизации трафика в соответствии с прогнозом.

Мы также используем *T-тест* для проверки значимости GMAN в прогнозировании на 1 час вперед по сравнению с Graph WaveNet. p-значение меньше 0,01, что демонстрирует, что GMAN статистически превосходит Graph WaveNet.

Сравнение отказоустойчивости Значения условий трафика в реальном времени могут частично отсутствовать из-за неисправности датчика, потери пакетов во время передачи данных и т. д. Чтобы оценить отказоустойчивость, мы случайным образом отбрасываем долю *η* (коэффициент отказов, в диапазоне от 10% до 90%) исторических наблюдений (т.е. случайным образом заменяем  *входные значения η×N×P×C* нулями), чтобы сделать прогнозы на 1 час вперед. Как показано на рисунке 8, GMAN более отказоустойчив, чем современные методы. Это показывает, что GMAN может захватывать сложные пространственно-временные корреляции из «загрязненных» данных о трафике и корректировать зависимости из наблюдений на будущие временные шаги.

Чтобы исследовать влияние каждого компонента в нашей модели, мы оцениваем четыре варианта, удаляя пространственное внимание, временное внимание, стробированное слияние и трансформируя внимание из GMAN по отдельности, которые называются GMAN-NS, GMAN-NT, GMAN-NG и GMAN-NTr соответственно. На рисунке 9 представлен MAE на каждом шаге прогнозирования GMAN и четырех вариантах. Мы наблюдаем, что GMAN последовательно превосходит GMAN-NS, GMAN-NT и GMAN-NG, что указывает на эффективность пространственного внимания, временного внимания и стробируемого слияния при моделировании сложных пространственно-временных корреляций. Более того, GMAN работает лучше, чем GMAN-NTr, особенно в долгосрочной перспективе, демонстрируя, что механизм преобразования внимания эффективно смягчает эффект распространения ошибок.

Время вычисления Мы представляем время обучения и время вывода STGCN, DCRNN, Graph WaveNet и GMAN на наборе данных PeMS в таблице 2. На этапе обучения GMAN имеет схожую скорость с Graph WaveNet.

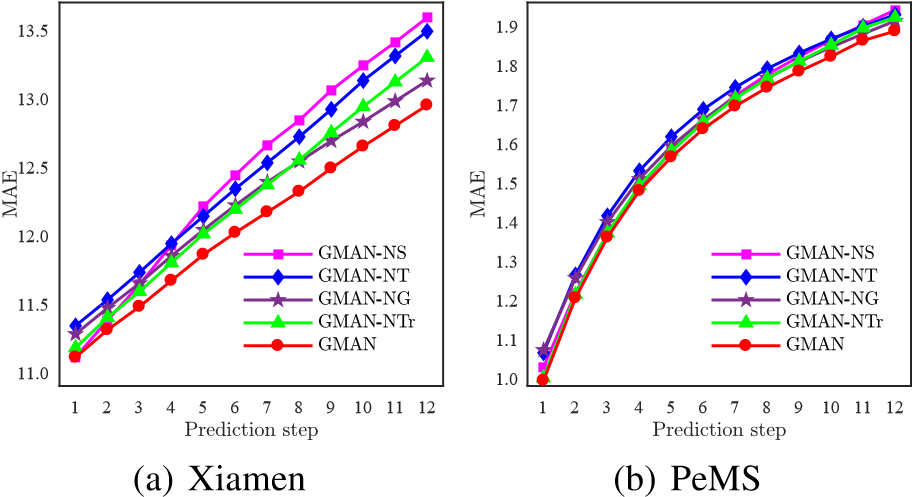


Рисунок 9: MAE каждого шага прогнозирования.

|  |  |  |
| --- | --- | --- |
| Метод | Обучение вычислениям Ti (с/эпоха) | меня  Вывод(ы) |
| СТГКН | 51.35 | 94.56 |
| DCRNN | 650.64 | 110.52 |
| График WaveNet | 182.21 | 6.55 |
| ГМАН | 217.62 | 9.34 |

Таблица 2: Время вычислений на наборе данных PeMS.

DCRNN работает намного медленнее, чем другие методы, из-за трудоемкого обучения последовательности в рекуррентных сетях. STGCN является наиболее эффективным, но показывает плохую производительность прогнозирования (Таблица 1). На этапе вывода мы сообщаем об общих временных затратах на данные проверки. STGCN и DCRNN менее эффективны, так как им требуются итеративные вычисления для получения 12 результатов прогнозирования. GMAN и Graph WaveNet могут создавать прогнозы на 12 шагов вперед за один прогон и, таким образом, тратить меньше времени на выводы.

Что касается второй наилучшей модели Graph WaveNet, как показано в Таблице 1, GMAN выгодно отличается от Graph WaveNet долгосрочными (например, на 1 час вперед) прогнозами трафика (Таблица 1) с аналогичными вычислительными затратами как для обучения, так и для вывода (Таблица 2).

# Заключение

Мы предложили графовую сеть множественного внимания (GMAN) для прогнозирования условий дорожного движения на временные шаги вперед на графе дорожной сети. В частности, мы предложили механизмы пространственного и временного внимания с помощью стробируемого слияния для моделирования сложных пространственно-временных корреляций. Кроме того, мы разработали механизм преобразования внимания, чтобы уменьшить эффект распространения ошибок и повысить производительность долгосрочного прогнозирования. Эксперименты на двух реальных наборах данных показывают, что GMAN достигает самых современных результатов, и преимущества становятся более очевидными, поскольку прогнозы делаются в далеком будущем. В будущем мы будем применять GMAN для других задач пространственно-временного прогнозирования, таких как прогнозирование потребления воды.

# Подтверждения

Эта работа была поддержана грантами Фонда естественных наук Китая (61872306 и U1605254) и Сямыньского научно-технического бюро (3502Z20193017).

# Ссылки

Этвуд, Д., и Таусли, Д. 2016. Диффузионно-сверточные нейронные сети. В *NeurIPS*, 1993–2001.

Бронштейн, М. М.; Бруна, Дж.; Лекун, И.; Слам, А.; и Вандергейнст,. 2017. Геометрическое глубокое обучение: выход за рамки евклидовых данных. *Журнал IEEE по обработке сигналов* 34(4):18–42.

Чен, Дж.; Ма, Т.; и Сяо, К. 2018. Fastgcn: Быстрое обучение с помощью графовых сверточных сетей с помощью выборки важности. В *ICLR*.

Ченг, В.; Шен, Ю.; Чжу, Ю.; и Хуанг, Л. 2018. Нейронная модель внимания для вывода о качестве воздуха в городах: изучение весов станций мониторинга. В *AAAI,* 2151–2158.

Куи,.; Ванг, Х.; Пей, Дж.; и Чжу, В. 2019. Опрос по встраиванию в сеть. *IEEE Transactions on Knowledge and Data Engineering* 31(5):833–852.

Дефферрар, М.; Брессон, Х.; и Вандергейнст,. 2016. Сверточные нейронные сети на графах с быстрой локализованной спектральной фильтрацией. В *NeurIPS*, 3844–3852.

Ду, С.; Ли, Т.; Гонг, Х.; и Horng, S.-J. 2018. Гибридный метод прогнозирования транспортных потоков с использованием мультимодального глубокого обучения. *Препринт arXiv arXiv:1803.02099*.

Гровер, А., и Лесковец, Й. 2016. Node2vec: масштабируемое обучение функций для сетей. В *KDD,* 855–864.

Гамильтон, У. Л.; Ин, Р.; и Лесковец, Й. 2017. Обучение индуктивным представлениям на больших графах. В *NeurIPS*, 1024– 1034.

Хе, К.; Чжан, Х.; Рен, С.; и Сан, Д. 2016. Глубокое остаточное обучение для распознавания изображений. В *CVPR,* 770–778.

Хоу, З., и Ли, Х. 2016. Повторяемость и схожесть транспортных потоков на автомагистралях и долгосрочное прогнозирование в условиях больших данных. *Труды IEEE по интеллектуальным транспортным системам* 17(6):1786–1796.

Кингма, Д.., и Ба, Д. Л. 2015. Adam: метод для стохастической оптимизации. В *ICLR*.

Кипф, Т. Н., и Веллинг, М. 2017. Полуконтролируемая классификация с графовыми сверточными сетями. В *ICLR*.

Ли, Р.; Ванг, С.; Чжу, Ф.; и Хуанг, Д. 2018а. Сверточные нейронные сети с адаптивным графом. В *AAAI.*

Ли, Ю.; Ю, Р.; Шахаби, К.; и Лю, Ю. 2018b. Диффузионная сверточная рекуррентная нейронная сеть: прогнозирование трафика на основе данных. В *ICLR*.

Lv, Z.; Сюй, Дж.; Чжэн, К.; Инь, Х.; Чжао,.; и Чжоу, X. 2018. Lc-rnn: модель глубокого обучения для прогнозирования скорости движения. В *IJCAI*, 3470–3476.

Ма, Х.; Тао, З.; Ванг, И.; Ю, Х.; и Ванг, Ю. 2015. Нейронная сеть с длительной кратковременной памятью для прогнозирования скорости движения по данным удаленных микроволновых датчиков. *Транспортные исследования Часть C: Новые технологии* 54:187–197.

Макридакис, С., и Хибон, М. 1997. Arma и методология boxjenkins. *Журнал прогнозирования* 16(3):147– 163.

Наир, В., и Хинтон, Г. Э. 2010. Ректифицированные линейные узлы улучшают станки Больцмана с ограниченным доступом. В *ICML,* 807–814.

Шен, Т.; Цзян, Дж.; Чжоу, Т.; Пан, С.; Лонг, Г.; и Чжан, К. 2018. Disan: Направленная сеть самовнимания для понимания языка без rnn/cnn. В *AAAI,* 5446–5455.

Сонг, Х.; Канасуги, Х.; и Сибасаки, Р. 2016. Deeptransport: Прогнозирование и моделирование мобильности человека и вида транспорта на общегородском уровне. В *IJCAI*, 2618– 2624.

Суцкевер, И.; Виньялс, О.; и Le, Q. V. 2014. Последовательное обучение с помощью нейронных сетей. В *NeurIPS*, 3104– 3112.

ван ден Оорд, А.; Дилеман, С.; Дзен, Х.; Симоньян, К.; Виньялс, О.; Грейвс, А.; Kalchbrenner, N.; Сениор, А.; и Кавукчуоглу, К. 2016. Wavenet: генеративная модель для необработанного звука. *Препринт arXiv arXiv:1609.03499*.

Васвани, А.; Шазир, Н.; Пармар, Н.; Uszkoreit, J.; Джонс, Л.; Гомес, А. Н.; ukasz Kaiser; и Полосухин, И. 2017. Внимание – это все, что вам нужно. В *NeurIPS*, 5998–6008.

Великовый,.; Cucurull, G.; Казанова, А.; Ромеро, А.; Ли,.; и Bengio, Y. 2018. Построение графиков сетей внимания. В *ICLR*.

Ванг, И.; Фан, Х.; Liu, X.; Чжэн, К.; Чен, Л.; Ванг, К.; и Li, J. 2017. Нелицензированный сервис обнаружения такси на основе данных о мобильности крупных транспортных средств. В *ICWS*, 857–861.

Ву, З.; Пан, С.; Чен, Ф.; Лонг, Г.; Чжан, К.; и Yu, P. S. 2019a. Всесторонний обзор графовых нейронных сетей.

*Препринт arXiv arXiv:1901.00596*.

Ву, З.; Пан, С.; Лонг, Г.; Цзян, Дж.; и Чжан, К. 2019b. Графовый вейвнет для глубокого пространственно-временного моделирования графов. В *IJCAI.*

Wu, C.-H.; Ho, J.-M.; и Ли, Д. Т. 2004. Прогнозирование времени в пути с помощью регрессии опорных векторов. *Труды IEEE по интеллектуальным транспортным системам* 5(4):276–281.

Яо, Х.; Ву, Ф.; Ке, Дж.; Tang, X.; Цзя, Ю.; Лу, С.; Гонг,.; Йе, Дж.; и Li, Z. 2018. Глубокая многовидовая пространственно-временная сеть для прогнозирования спроса на такси. В *AAAI,* 2588–2595.

Яо, Х.; Tang, X.; Вэй, Х.; Чжэн, Г.; и Li, Z. 2019. Возвращаясь к пространственно-временному сходству: фреймворк глубокого обучения для прогнозирования трафика. В *AAAI.*

Ю, Б.; Инь, Х.; и Чжу, З. 2018. Сверточные сети пространственно-временного графа: среда глубокого обучения для прогнозирования трафика. В *IJCAI*, 3634–3640.

Чжан, Дж.; Чжэн, Ю.; и Qi, D. 2017. Глубокие пространственно-временные остаточные сети для прогнозирования потоков толпы в масштабах города. В *AAAI,* 1655–1661.

Чжэн, З., и Су, Д. 2014. Краткосрочное прогнозирование объема трафика: подход k-ближайшего соседа, усиленный ограниченным линейным принципом компонентного алгоритма. *Транспортные исследования Часть C: Новые технологии* 43:143–157.

Чжэн, К.; Фан, Х.; Вэнь, К.; Чен, Л.; Ванг, К.; и Li, J. 2019. Deepstd: Извлечение пространственно-временных возмущений из множества контекстных факторов для прогнозирования транспортных потоков в масштабах города. *IEEE Transactions on Intelligent Transportation Systems*. будет опубликован.