Machine Learning
https://doi.org/10.1007/s10994-019-05815-0

Atenção ao padrão temporal para previsão de séries temporais multivariadas

Shun-Yao Shih[1] · **Fan-Keng Sun**[1] · **Hung-yi Lee**[1]

Recebido: 26 novembro 2018 / Revisado: 1 de maio de 2019 / Aceito: 29 may 2019

Abstrair
A previsão de dados de séries temporáticas multivariadas, por exemplo, a previsão do consumo de eletricidade, produção de energia solar e peças de piano polifônicos, tem inúmeras aplicações valiosas. No entanto, interdependências complexas e não lineares entre etapas do tempo e séries complicam essa tarefa. Para obter uma previsão precisa, é crucial modelar a dependência de longo prazo em dados de séries temporais, que podem ser alcançadas por redes neurais recorrentes (RNNs) com um mecanismo de atenção. O mecanismo de atenção típico revisa as informações a cada tempo anterior e seleciona informaçõesrelevantinformations para ajudar a regenerar os pontos de vista;no entanto, ele não consegue capturar padrões temporais em várias etapas do tempo. Neste artigo, propomos o uso de um conjunto de filtros para extrair padrões temporais invariantes, semelhante à transformação de dados de séries temporais em seu "domínio de frequência". Em seguida, propomos um novo mecanismo de atenção para selecionar séries temporizais relevantes e usamos suas informações de domínio de frequência para previsão multivariada. Aplicamos o modelo proposto em várias tarefas do mundo real e alcançamos a execução de última geração em quase todos os casos. Nosso código fonte está disponível em https://github.com/gantheory/ TPA-LSTM.

Palavras-chave Série temporal multivariada · Mecanismo de atenção · Rede neural recorrente ·

Rede neural convolucional · Geração de música polifônica

Shun-Yao Shih e Fan-Keng Sun contribuíram igualmente para este estudo.

[1] Universidade Nacional de Taiwan, Taipei, Taiwan

Published online: 11 June 2019

Editores: Karsten Borgwardt, Po-Ling Loh, Evimaria Terzi, Antti Ukkonen.

B Shun-Yao Shih shunyaoshih@gmail.com
B03901056@ntu.edu.tw Fan-Keng Sun
Hung-yi Lee
hungyilee@ntu.edu.tw

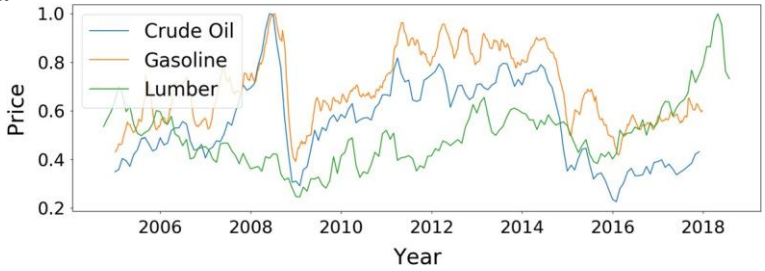

Fig. 1 Preços históricos de petróleo bruto, gasolina e madeira. Unidades são omitidas e balanças são normalizadas para a simplicidade

1 Introdução

Na vida cotidiana, os dados das séries tempordias estão por toda parte. Observamos variáveis em evolução geradas a partir de sensores em etapas de tempo discretas e as organizamos em dados de séries temporidas. Por exemplo, o consumo de eletricidade doméstica, taxa de ocupação rodoviária, taxa de câmbio, produção de energia solar e até notas de música podem ser vistos como dados de séries temporâneais. Na maioria dos casos, os dados coletados são, muitas vezes, dados de séries temporizais multivariadas (MTS), como o consumo de eletricidade de vários clientes, que são rastreados pela empresa de energia local. Podem existir interdependências dinâmicas complexas entre diferentes séries que são significativas, mas difíceis de capturar e analisar.

Analistas muitas vezes buscam prever o futuro com base em dados históricos. Quanto melhor as interdependências among diferentes séries são modeladas, mais precisa a previsão pode ser. Por exemplo, como mostrado em Fig. 1,[2] o preço do petróleo bruto influencia fortemente o preço da gasolina, mas tem uma influência menor no preço da madeira. Assim, dada a constatação de que a gasolina é produzida a partir de petróleo bruto e madeira não é, podemos usar o preço do petróleo bruto para prever o preço da gasolina.

No aprendizado de máquina, queremos que o modelo aprenda automaticamente essas interdependências a partir de dados. Machinelearninghasbeenappliedtotimeseriesanalysforbothclassification and forecasting (Zhang et al. 1998; Zhang 2003; Lai et al. 2018; Qin et al. 2017). Na classificação, a máquina aprende a atribuir um rótulo a uma série temporal, por exemplo, avaliando as categorias de diagnóstico de um paciente, reading valores de sensores médicos. Na previsão, a máquina prevê séries temporizouissas futuras com base em dados observados anteriormente. Por exemplo, a precipitação nos próximos dias, semanas ou meses pode ser prevista de acordo com medições históricas. Quanto mais à frente tentarmos prever, mais difícil será.

[2] Fonte: https://www.eia.gov e https://www.investing.com.

Quando se trata de previsão de MTS usando aprendizado profundo, redes neurais recorrentes (RNNs) (Rumelhart et al. 1986; J.Werbos 1990; Elman 1990) são frequentemente usados. No entanto, uma desvantagem no uso de RNNs em análises de séries temporizadas é sua fraqueza no gerenciamento de dependências de longo prazo, por exemplo, padrões anois em uma sequência registrada diária (Cho et al. 2014). O mecanismo de atenção (Luong et al. 2015; Bahdanau et al. 2015), originalmente utilizado em codificador-decodificador (Sutskever et al. 2014) as redes, um pouco alivia esse problema, e assim aumenta a eficácia da RNN (Lai et al. 2018).

Neste artigo, propomos a atenção do padrão temporal, um novo mecanismo de atenção para a previsão do MTS, onde usamos o termo "padrão temporal" para se referir a qualquer padrão invariante de tempo em várias etapas do tempo. O mecanismo de atenção típico identifica os passos de tempo relevantes para a previsão, e extrai as informações a partir desses passos de tempo, o que coloca limitações óbvias para a previsão de MTS. Considere o exemplo em Fig. 1. Para prever o valor da gasolina, a máquina deve aprender a se concentrar no "petróleo bruto" e ignorar a "madeira". Em nossa atenção ao padrão temporal, em vez de selecionar os passos de tempo relevantes como no mecanismo de atenção típico, a máquina the aprende a selecionar as séries temporais relevantes.

Além disso, os dados de séries temporais muitas vezes implicam padrões temporais periódicos perceptíveis, que são críticos para a prescrição. No entanto, operiodicpatternsspanningmultipletimestepsaredifficult para o mecanismo típico de attention para identificar, pois geralmente se concentra apenas em alguns passos do tempo. Na atenção do padrão temporal, introduzimos uma rede neural convolucional (CNN) (LeCun e Bengio 1995; Krizhevsky et al. 2012) extrair informações de padrão temporal de cada variável individual.

As principais contribuições deste artigo são resumidas da seguinte forma:

- Weintroduceanewattentionconceptinwhselecttherelevantvariablesasopposed to oposição às etapas de tempo relevantes. O método é simples e geral para aplicar na RNN.
- Usamos exemplos de brinquedos para verificar se nosso mecanismo de atenção permite que o modelo extraia padrões temporais e se concentre em diferentes etapas do tempo para diferentes séries temporais.
- Atestados por resultados experimentais em dados do mundo real que vão desde tarefas periódicas e parcialmente lineares até tarefas não periódicas e não lineares, mostramos que o mecanismo de atenção proposto alcança resultados de última geração em vários conjuntos de dados.
- Os filtros da CNN aprendidos em nosso mecanismo de atenção demonstram um comportamento interessante e inter-pretoble.

O restante deste paper é organizado da seguinte forma. Em Sect. 2 revisamos o trabalho relacionado e na Seita. 3 descrevemos o conhecimento de fundo. Então, em Seita. 4 descrevemos o mecanismo de atenção proposto. Em seguida, nós apresentamos eanalyzeourattention mechanismontoyexamples inSect. 5,andonMTSandpolyphonicmusicdatasetinSect. 6. Finalmente, nósconcluimos oEct. 7.

2 Trabalho relacionado

O modelo mais conhecido para previsão de séries de tempo univariadas lineares é a média móvel integrada autoregressiva (ARIMA) (Box et al. 2015), que engloba outros modelos

de séries temporais autoregressivas, incluindo autoregressão (AR), média móvel (MA) e média móvel autoregressiva (ARMA). Além disso, regressão vetorial de suporte linear (SVR) (Cao e Tay 2003; Kim 2003) trata o problema de previsão como um problema típico de regressão com parâmetros de variação de tempo. No entanto, esses modelos são principalmente limitados a séries temporizais lineares univariadas e não escalam bem para MTS. Para prever os dados do MTS, foi proposta a autoregressão vetorial (VAR), uma generalização dos modelos baseados em AR. VAR é provavelmente o modelo mais conhecido na previsão de MTS. No entanto, nem modelos baseados em AR nem baseados em VAR capturam não linearidade. Por essa razão, esforços substanciais foram colocados em modelos não lineares para previsão de séries temporais com base em métodos de kernel (Chen et al. 2008), conjuntos (Bouchachia e Bouchachia 2008), processos gaussianos (Frigola e Rasmussen 2014) ou mudança de regime (Tong and Lim 2009). Ainda assim, essas abordagens aplicam não linearidades pré-determinadas e podem não reconhecer diferentes formas de não linearidade para diferentes MTS.

Recentemente, deepneuralnetworkshavereceivedagreatamountofattentionduetheirability to capture interdependências não lineares. A memória de longo curto prazo (LSTM) (Hochreiter e Schmidhuber 1997), uma variante da rede neural recorrente, tem mostrado resultados promissores em várias tarefas de PNL e também foi empregada para a previsão de MTS. O trabalho nessa área começou com o uso de naive RNN (Connor et al. 1991), melhorado com modelos híbridos que combinavam ARIMA e perceptrons multicamadas (Zhang et al. 1998; Zhang 2003; Jain e Kumar 2007),

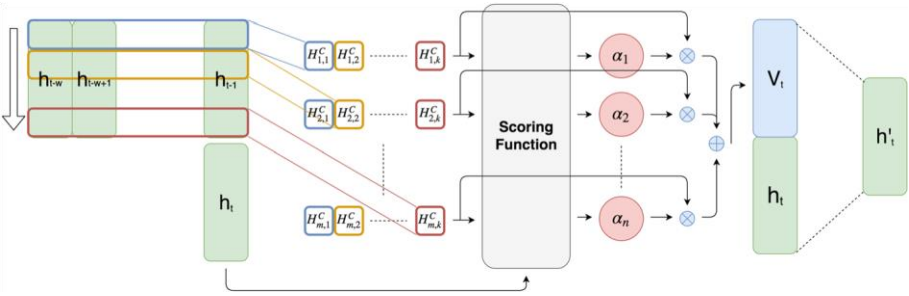

Fig. 2 Mecanismo de atenção proposto. *ht* representa o estado oculto do RNN no passo *t*. Existem filtros K 1-D CNN com comprimento w, mostrados como diferentes cores de retângulos. Em seguida, cada filtro se mistura sobre características m de estados ocultos e produz um *HC* de matriz com linhas *m* e colunas k. Em seguida, a função de pontuação calcula um peso para cada linha de *HC*, comparando-se com o estado oculto atual *ht*. Por último, mas não menos importante, os pesos são normalizados e as linhas de *HC* são ponderadas somadas por seus pesos correspondentes para gerar *Vt*. Finalmente, concatenamos *Vt,ht* e executamos multiplicação matricial para gerar *ht*, que é usado para criar o valor final de previsão (Figura de cor on-line)

ethemtrecentlyprogressedtodynamicBoltzmannmachineswithRNN(Dasguptaand Osogami 2017). Embora esses modelos possam ser aplicados ao MTS, eles visam principalmente séries temporizais univariadas ou bivarias.

Até onde sabemos, a rede de séries de tempo de longo e curto prazo (LSTNet) (Lai et al. 2018) é o primeiro modelo projetado especificamente para a previsão de MTS com até centenas de séries temporáticas. O LSTNet usa CNNs para capturar padrões de curto prazo e LSTM ou GRU para memorizar padrões relativamente de longo prazo. Na prática, porém, o LSTM e o GRU não podem memorizar interdependências de longo prazo devido à instabilidade de treinamento e ao problema de desaparecimento do gradiente. Para lidar com

isso, o LSTNet adiciona uma camada de salto recorrente ou um mecanismo de atenção típico. Também parte do modelo desmodelistraditionalautoregression, que ajuda a mitigar a insensibilidade de escala das redes neurais. No entanto, o LSTNet tem três grandes deficiências quando comparado ao nosso mecanismo de atenção proposto: (1) o comprimento do salto da camada de salto recorrente deve ser ajustado manualmente para corresponder ao período dos dados, enquanto nossa abordagem proposta aprende os padrões periódicos por si só; (2) o modelo LSTNet-Skip é especificamente projetado para dados MTS com padrões periódicos, enquanto nosso modelo proposto, como mostrado em nossos experimentos, é simples e adaptável a conjuntos de dados variosos, mesmo os não periódicos; e (3) a camada de atenção no modelo LSTNet-Attn seleciona um estado oculto relevante como no mecanismo de atenção típico, enquanto nosso mecanismo de atenção proposto seleciona séries de tempo relevantes que é uma mecan mais adequada ism para dados MTS.

3 Preliminares

Nesta seção, descrevemos brevemente dois módulos essenciais relacionados ao nosso modelo proposto: o módulo RNN e o mecanismo de atenção típico.

3.1 Redes neurais recorrentes

Dada uma sequência de informações $\{x_1, x_2, \ldots, x_t\}$, onde $x_i \in {}^{R_n}$, um RNN geralmente define uma função recorrente, F, e calcula $h_t \in \mathrm{Rm}$ para cada passo de tempo, t, como

$$h_t = F(h_{t-1}, x_t) \tag{1}$$

onde a implementação da função F depende de que tipo de célula RNN é usada.

Células de memória de longo prazo (LSTM) (Hochreiter e Schmidhuber 1997) são amplamente utilizadas, que têm uma função recorrente ligeiramente diferente:

$$\mathrm{ht,ct} = F(_{ht-1}, ct-1, xt), \tag{2}$$

que é definido pelas seguintes equações:

$$_{it} = sigmoid\ (Wxi\ xt + Whi\ ht_{-1}) \tag{3}\ _{ft} = sigmoid\ (Wx_f\ xt + Wh_f\ ht_{-1}) \tag{4}$$

$$o_t = sigmoid\ (Wxo\ xt + Whoht_{-1}) \tag{5}\ ct = {}_{ft}\ ct_{-1} + i_t \odot tanh(_{Wxg\ xt} +$$

$$_{Whg\ ht\ -1}) \tag{6}\ ht = ot\ tanh(ct) \tag{7}$$

onde it, ft, ot, e $ct \in {}_{Rm}$, Wxi , Wx_f, Wxo e $Wxg \in \mathrm{Rm}\times n$, Whi , Wh_f, $_{Who}$ e $_{Whg} \in \mathrm{Rm}\times {}^{m}$, e denota multiplicação em termos de elementos.

3.2 Mecanismo típico de atenção

No mecanismo de atenção típico (Luong et al. 2015; Bahdanau et al. 2015) em um RNN, dado os estados anteriores $H = \{h1,h2,...,ht-1\}$, um vetor de contexto vt é extraído dos estados anteriores. vt isaweightedsumofeachcolumnhi in H,que representa as informações relevantes para o passo atual. vt é ainda mais integrado com o estado present *ht* para produzir a previsão.

Assuma uma função de pontuação $f : Rm \times Rm \rightarrow R$ que computa a relevância entre seus vetores de entrada. Formalmente, temos a seguinte fórmula para calcular o vetor de contexto vt:

$$\alpha \quad \sum_{j=1}^{t-1} \exp(f(h_j \quad _i = \frac{\exp(f(\text{hi,ht}))}{(ht))} \tag{8}$$

$$_{vt} = \sum_{i=1}^{t-1} \alpha \, ihi. \tag{9}$$

4 Atenção ao padrão temporal

Embora o trabalho anterior se concentre principalmente em mudar a arquitetura de rede dos modelos baseados em atenção através de diferentes configurações para melhorar o desempenho em várias tarefas, acreditamos que há um defeito crítico na aplicação de mecanismos típicos de atenção na RNN para previsão de MTS. O mecanismo de atenção típico seleciona informações relevantes para a etapa de tempo atual, e o vetor de contexto vt é a soma ponderada dos vetores de coluna de estados ocultos RNN anteriores, $H = \{h1,h2,...,_{ht-1}\}$. Este design se presta a tarefas em que cada etapa de tempo contém uma única informação, por exemplo, uma tarefa NLP na qual cada passo de tempo corresponde a uma única palavra. Se houver múltiplas variáveis em cada etapa de tempo, ela não ignora variáveis que são ruivas em termos de utilidade de previsão. Além disso, uma vez que o mecanismo de atenção típico faz uma média das informações em várias etapas do tempo, ele não detecta padrões temporais úteis para a previsão.

A visão geral do model proposto é mostrada em Fig. 2. Na abordagem proposta, dado os estados ocultos anteriores da RNN $H \in Rm^{\times(t\_1)}$, o mecanismo de atenção proposto basicamente atende aos seus vetores de linha. Os pesos de atenção nas linhas selecionam as variáveis que são úteis antes de serem forecasting. Uma vez que o contextovectorvt é agora os teretores de sumô ponderados contendo a informação em várias etapas do tempo, ele captura a informação temporal.

4.1 Formulação de problemas

Na previsão de MTS, dado um MTS, $X = \{x1, x2,... , xt-1\}$, onde $xi \in Rn$ representa o valor observado no momento *i*, a tarefa é prever o valor de $xt-1+\Delta$, onde Δ é um horizonte fixo em

relação a diferentes tarefas. Denotamos a previsão correspondente como $y_{t-1+\Delta}$, e o valor da verdade terrestre como $y_{...t-1+\Delta} = x_{t-1+\Delta}$. Além disso, para cada tarefa, usamos apenas $\{x_{t-w}, x_{t-w+1}, \ldots, x_{t-1}\}$ para prever $x_{t-1+\Delta}$, onde w é o tamanho da janela. Esta é uma prática comum (Lai et al. 2018; Qin et al. 2017), porque a suposição é que não há informações úteis antes da janela e a entrada é, portanto, fixada.

4.2 Detecção de padrão temporal usando a CNN

O sucesso da CNN não está em grande parte em sua capacidade de capturar vários padrões de sinal importantes; como tal, usamos uma CNN para aumentar a capacidade de aprendizado do modelo, aplicando filtros DA CNN nos vetores de linha de H. Especificamente, temos k filtros $Ci \in {}^{R1\times T}$, onde T é a máxima atenção lengthwearepayingattentionto. Ifunspecified,weassume T = w. Operações convolucionais produzem $HC \in Rn\times^{k}$ onde $_{HiC,j}$ representa o valor convolucional do vetor da décima linha e do filtro *jth*. Formalmente, esta operação é dada por

$$= \sum_{l=1}^{w} H_{i,(t-w-1+l)HiC,j} \times C_{j,T-w+l}. \tag{10}$$

4.3 Mecanismo de atenção proposto

Calculamos vt como uma soma ponderada de vetores de linha de HC. Definido abaixo está a função de pontuação $f : Rk \times Rm \rightarrow R$ para avaliar a relevância:

$$(H_i^C, h_t) = (H_i^C)_{fWaht} \tag{11}$$

onde $_{HiC}$ é a décima linha do HC, e $Wa \in Rk\times^{m}$. O peso da atenção α_i é obtido como

$$\alpha_i = sigmoid(f(H_i^C, h_t)). \tag{12}$$

Note que usamos a função de ativação sigmoid em vez de softmax, pois esperamos que mais de uma variável seja útil para a previsão.

Completando o processo, os vetores de linha do HC são ponderados by α_i para obter o vetor de contexto vt ∈ Rk,

$$_{vt} = \sum_{i=1}^{n} \alpha_i H_i^C. \tag{13}$$

Então nós integramos vt e *ht* para produzir a previsão final

$$_{ht} = Whht + \text{Wvvt}, \tag{14}$$

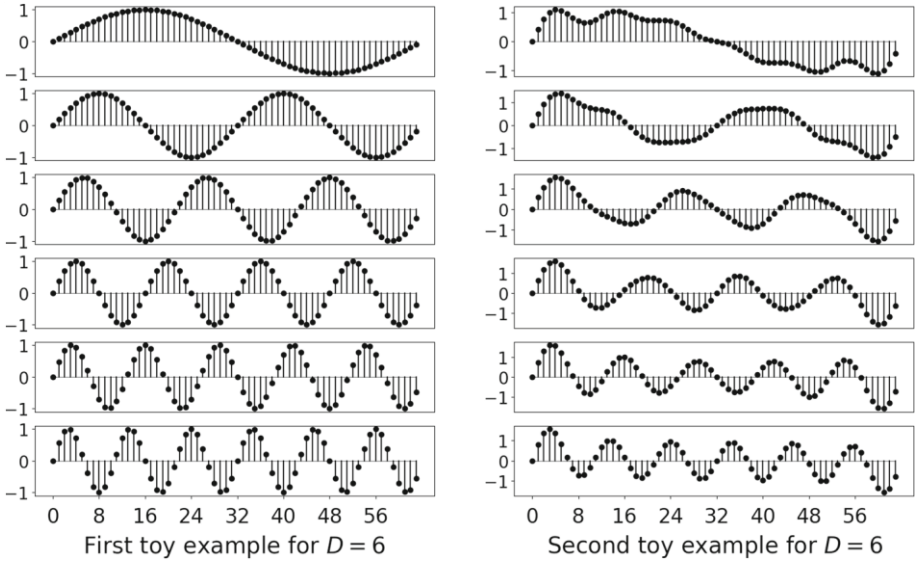

Fig. 3 Visualização do primeiro tipo de exemplos de brinquedos sem interdependências (esquerda) e o segundo tipo de exemplos de brinquedos com interdependências (direita) para $D = 6$, o que significa que há 6 séries de tempo em cada exemplo

$$_{yt-1+\Delta} = Whht, \tag{15}$$

onde ht, $ht \in {}^{Rm}$, $Wh \in \text{Rm}\times^{m}$, $Wv \in \text{Rm}\times^{k}$, e $Wh \in \text{Rn}\times^{m}$ e $yt-1+\Delta \in \text{R}\ ^{n}$.

5 Análise da atenção proposta em exemplos de brinquedos

Para elaborar a falha dos mecanismos tradicionais de atenção e a influência das interdependências, estudamos o desempenho de diferentes mecanismos de atenção em dois tipos de exemplos de brinquedos artificialmente construídos.

No primeiro tipo de exemplos de brinquedos, o primeiro passo da série do tempo é definido como $\sin(\frac{2\pi it}{64})$, ou seja, cada série de tempo é uma onda seno com períodos diferentes. Observe que qualquer série de duas vezes são mutuamente independentes no primeiro tipo, portanto não há interdependência.

O segundo tipo de exemplos de brinquedos adiciona interdependências ao primeiro tipo misturando séries temporizadas, e, portanto, o primeiro passo da série da décima vez é formulado como:

$$\sin\left(\frac{2\pi it}{64}\right)+\frac{1}{D-1}\sum_{j=1,j\neq i}^{D}\sin\left(\frac{2\pi jt}{64}\right) \quad (16)$$

onde D é o número de séries temporariais. Ambos os tipos de exemplos de brinquedos são visualizados em Fig. 3 para
$D = 6$.

Todos os modelos nas análises a seguir são treinados com tamanho de janela w = 64, horizonte Δ = 1, e quantidade semelhante de parâmetros. Nesta configuração, cada um dos nossos exemplos de brinquedos consiste em 64 amostras. Cada série de tempo na primeira amostra compreende valores de Eq. 16 de t = 0 a 63, e podemos mudar um passo de tempo para obter a segunda amostra com valores de t = 1 a 64. Pelo

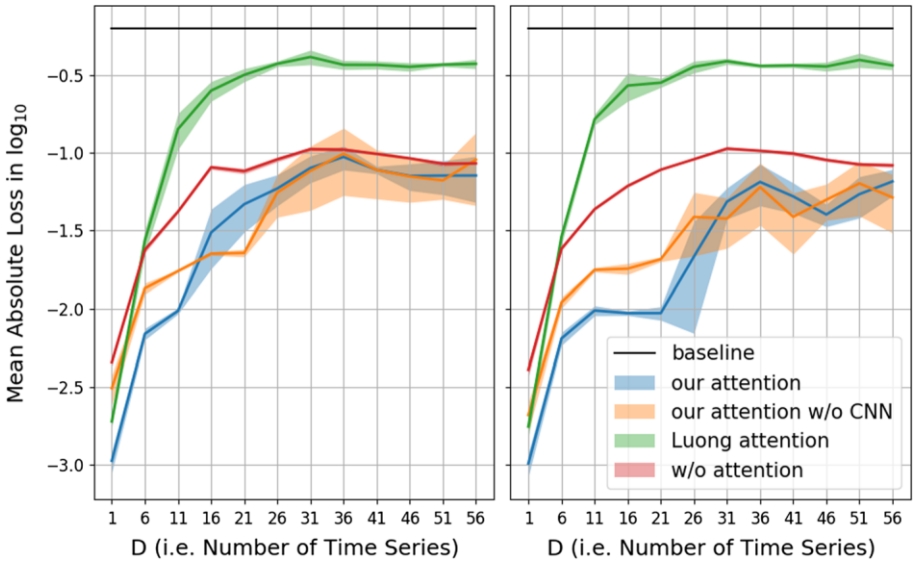

Fig. 4 Perda absoluta média e a faixa de desvio padrão no $\log_{10}$ do primeiro tipo de exemplos de brinquedos sem interdependências (esquerda) e o segundo tipo de exemplos de brinquedos com interdependências (direita), ambos em dez corridas. A linha de base indica a perda se todos os valores previstos forem zero

última amostra, utilizamos valores de t = 63 a 126 como a série de entrada correspondentemente. Observe que os valores de t = 64 a 127 são iguais aos de t = 0 a 63. Treinamos os modelos para 200 épocas em dois tipos de exemplos de brinquedos para D = {1,6,11,... ,56} e recorde significa perda absoluta no treinamento. Não há validação e testes de dados porque a intenção desta seção é demonstrar a maior capacidade de nossa atenção sobre a atenção típica para encaixar dados MTS e não a generalização de nossa atenção. Os resultados são mostrados em Fig. 4.

5.1 Falha of mecanismos de atenção tradicionais

Intuitivamente, para o primeiro exemplo de brinquedo, o modelo pode prever com precisão o próximo valor memorizando o valor que parece exactactlyoneperiod antes. No entanto, sabemos que diferentes séries temporais têm períodos diferentes, o que significa ter uma boa previsão, o modelo deve ser capaz de olhar para trás diferentes números de passos de tempo para diferentes séries. A partir deste ponto, fica claro que a falha dos mecanismos tradicionais de atenção vem da extração de apenas uma etapa de tempo anterior, enquanto retira as informações em outras etapas do tempo. Por outro lado, nosso mecanismo de atenção atende aos recursos extraídos dos vetores de linha de estados ocultos RNN pelos filtros CNN, o que permite que o modelo selecione informações relevantes em vários steps de tempo.

A explicação acima mencionada é verificada pelo enredo esquerdo em Fig. 4, onde observamos que o desempenho do LSTM com a atenção de Luong é ruim quando *D1* , em comparação com os outros. Observe que todos os modelos possuem quantidade semelhante de parâmetros, o que implifica que o LSTM sem atenção tem um tamanho oculto maior quando comparado ao LSTM com a atenção luong. Consequentemente, o LSTM sem atenção supera o LSTM com a atenção de Luong quando *D1* , porque o tamanho oculto maior ajuda o modelo a make previsão enquanto a atenção luong é quase inútil. Pelo contrário, nossa atenção é útil, então o LSTM com nossa atenção é melhor que o LSTM sem atenção em média, mesmo que seu tamanho oculto seja menor. Além disso, remover a ATENÇÃO da CNN, que resulta no mesmo modelo da célula "Sigmoid - W/o CNN" na Tabela 4, não afeta o desempenho, o que implica que nossa atenção em termos de características é indispensável.

5.2 Influência das interdependências

Quando há interdependências nos dados MTS, é desejável aproveitar as interdependências para melhorar ainda mais a precisão de previsão. O enredo certo em Fig. 4 mostra que tanto o LSTM com a atenção de Luong quanto o LSTM sem atenção não se beneficiam das interdependências adicionadas, os valores de perda permanecem os mesmos. Por outro lado, a perda do LSTM com a oferta de atenção é menor quando há interdependências, o que sugere que nossa atenção utilizou com sucesso as interdependências para facilitar a previsão do MTS. Mais uma vez, remover a CNN de nossa atenção não afeta o desempenho neste caso.

6 Experimentos e análises

Inthissection,wefirstdescribethedatasetsuponwhichweconductedourexperiments. Em seguida, apresentamos nossos resultados experimentais e uma visualização da prediction contra a LSTNet. Então, discutimos o estudo da ablação. Finalmente, analisamos em que sentido os filtros da CNN se assemelham às bases em DFT.

6.1 Conjuntos de dados

Para avaliar a eficácia e a capacidade de generalização do mecanismo de atenção proposto, weusedtwodissimilartypesofdatasets:typicalMTSdatasetsandpolyphonicmusicdatasets. Os conjuntos de dados típicos do MTS são publicados pela Lai et al. (2018); existem quatro conjuntos de dados:

- Energia Solar[3]: os dados de produção de energia solar de usinas fotovoltaicas no Estado do Alabama em 2006.
- Tráfego[4]: dois anos (2015-2016) de dados fornecidos pelo Departamento de Transportes da Califórnia que descreve a taxa de ocupação rodoviária (entre 0 e 1) nas rodovias da Baía de São Francisco.
- Eletricidade[5]: registro do consumo de energia elétrica de 321 clientes em kWh.
- Taxa de Câmbio: as taxas de câmbio de oito países estrangeiros (Austrália, Reino Unido, Canadá, China, Japão, Nova Zelândia, Cingapura e Suíça) de 1990 a 2016.

Esses conjuntos de dados são dados reais que contêm interdependências lineares e não lineares. Além disso, os conjuntos de dados de Energia Solar, Tráfego e Eletricidade apresentam fortes padrões periódicos indicando atividades humanas diárias ou semanais. De acordo com os autores do LSTNet, cada vez que as series em todos os conjuntos de dados foram divididas em treinamento (60%), validação (20%) e conjunto de testes (20%) em ordem cronológica.

Em contraste, os conjuntos de dados de música polifônica introduzidos abaixo são muito mais complicados, no sentido de que não existem linearidade aparente ou padrões repetitivos:

Tabela 1 Estatísticas de todos os conjuntos de dados, onde *L* é a duração da série temporal, *D* é o número de séries tempordicais, *S* é o espaçamento amostral, e *B* é o tamanho do conjunto de dados em bytes

Dataset	*L*	D	*S*	*B*
Energia solar	52,560	137	10 min.	172 M
Tráfego	17,544	862	1 hora	130 M
Eletricidade	26,304	321	1 hora	91 M
Taxa de câmbio	7,588	8	1 dia	534 K
MuseData	216–102,552	128	1 batida	4,9 M
LPD-5-Cleansed	1072–1,917,952	128	1 batida	1,7 G

MuseData e LPD-5-Cleansed têm séries de tempo de vários comprimentos, uma vez que o comprimento das peças musicais varia

- MuseData Boulanger-Lewandowski et al. (2012): uma coleção de peças musicais de vários compositores de música clássica em formato MIDI.
- LPD-5-Cleansed Dong et al. (2018); Raffel (2016): 21.425 rolos de piano multi-faixas que contêm bateria, piano, guitarra, baixo e cordas.

[3] http://www.nrel.gov/grid/solar-power-data.html.

[4] http://pems.dot.ca.gov.

[5] https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagramas20112014.

Para treinar modelos nesses conjuntos de dados, consideramos cada nota reproduzida como 1 e 0 de outra forma (ou seja, um descanso musical), e definimos uma batida como uma vez step como mostrado na Tabela 1. Dadas as notas tocadas de 4 barras compostas por 16 batidas, a tarefa é prever se cada arremesso na próxima etapa é jogado ou não. Para treinamento, validação e conjuntos de testes, seguimos a separação original do MuseData, que é dividida em 524 peças de treinamento, 135 peças de validação e 124 peças de teste. LPD-5-Cleansed, no entanto, não foi dividido em trabalhos anteriores (Dong et al. 2018; Raffel 2016); assim, dividimos-o aleatoriamente em conjuntos de treinamento (80%), validação (10%) e testes (10%) . O tamanho do conjunto de dados LPD-5-Cleansed é muito maior do que outros, por isso decidimos usar um conjunto de validação e teste menor.

A principal diferença entre conjuntos de dados típicos de MTS e conjuntos de dados de música polifônica é que os escalares nos conjuntos de dados típicos do MTS são contínuos, mas os escalars em conjuntos de dados de música polifônica são discretos (0 ou 1). As estatísticas dos conjuntos de dados típicos do MTS e dos conjuntos de dados de música polifônica são resumidas na Tabela 1.

6.2 Métodos para comparação

Comparamos o modelo proposto com os seguintes métodos nos conjuntos de dados típicos do MTS:

- AR: modelo de autorreressão padrão.
- LRidge: modelo VAR com regularização L2: o modelo mais popular para previsão de MTS.
- LSVR: modelo VAR com função objetiva SVR (Vapnik 1997).
- GP: Modelo de processo gaussiano (Frigola-Alcade 2015; Roberts et al. 2011).
- SETAR: Modelo de autorreressão de limiar auto-emocionante, um modelo clássico univariado não linear (Tong e Lim 2009).
- LSTNet-Skip: LSTNet com camada de salto recorrente.– LSTNet-Attn: LSTNet com camada de atenção.

AR, LRidge, LSVR, GP e SETAR são métodos tradicionais de linha de base, enquanto LSTNet-Skip e LSTNet-Attn são métodos de última geração baseados em redes neurais profundas.

No entanto, como ambos os métodos tradicionais de linha de base e LSTNet são inadequados para conjuntos de dados de música polifônica devido à sua não linearidade e à falta de periodicidade, usamos LSTM e LSTM com a atenção luong como modelos de linha de base para avaliar o modelo proposto em conjuntos de dados de música polifônica:

- LSTM: Células RNN introduzidas na Sect. 3.
- LSTM com atenção luong: LSTM com uma função de pontuação do mecanismo de atenção da qual $f(_{hi,ht}) = (hi)Wht$, onde $W \in \mathrm{Rm}\times^{m}$ (Luong et al. 2015).

Configuração do modelo 6.3 e configurações de parâmetros

Para todos os experimentos, usamos unidades LSTM em nossos modelos RNN, e fixamos o número de filtros CNN em 32. Além disso, inspirados no LSTNet, incluímos um

componente de autoregressão em nosso modelo ao treinar e testar em conjuntos de dados típicos do MTS.

Para conjuntos de dados típicos do MTS, realizamos uma pesquisa de grade sobre parâmetros tunable como feito com LSTNet. Especificamente, em Energia Solar, Tráfego e Eletricidade, o intervalo para o tamanho da janela w foi { 24,48.96.120.144.168}, o intervalo para o número de unidades ocultas *m* foi de {25,45,70}, e o ra nge para o passo da decadência da taxa de aprendizagem exponencial com uma taxa de 0,995 foi {200.300.500.1000}. Na Taxa de Câmbio, esses três parâmetros foram {30,60}, {6,12}e {120.200}, respectivamente. Dois tipos de normalização de dados também foram vistos como parte da pesquisa de grade: um normalizou cada série de tempo pelo valor máximo em si mesmo, e o outro normalizado cada vez que séries pelo valor máximo sobre todo o conjunto de dados. Por fim, usamos a função de perda absoluta e Adam com uma taxa de aprendizagem de 10-3 em Solar Energy, Tráfego e Eletricidade, e um 3 · Taxa de aprendizagem de 10-3 sobre taxa de câmbio. Para AR, LRidge, LSVR e GP, seguimos as configurações do parâmetro conforme relatado no papel LSTNet (Lai et al. 2018). Para a SETAR, pesquisamos a dimensão de incorporação sobre {24,48.96.120.144.168} para Energia Solar, Tráfego e Eletricidade, e fixamos a dimensão de incorporação para 30 para taxa de câmbio. As duas configurações diferentes entre nosso método e LSTNet são (1)temos dois métodos de normalização de dados para escolher, enquanto o LSTNet usou apenas o primeiro tipo de normalização de dados; e (2) a pesquisa de grade sobre o tamanho da janela w é diferente.

Para modelos utilizados para os conjuntos de dados de música polifônica, incluindo as linhas de base e modelos propostos nas seguintes subseções, utilizamos 3 camadas para todas as RNNs, como feito em Chuan e Herremans (2018), e fixamos os parâmetros treináveis para cerca de 5 · 106 ajustando o número de unidades LSTM para comparar de forma justa diferentes modelos. Além disso, usamos o otimizador Adam com uma taxa de aprendizado de 10-5 e uma função de perda de entropia cruzada.

6.4 Métricas de avaliação

Nos conjuntos de dados típicos do MTS, uma vez que comparamos o modelo proposto com o LSTNet, seguimos as mesmas métricas de avaliação: RAE, RSE e CORR. A primeira métrica é o erro absoluto relativo (RAE), que é definido como

$$\text{A RAE} = \frac{\sum_{t=t_0}^{t_1}\sum_{i=1}^{n} |(y_{t,i} - \hat{y}_{t,i})|}{\sum_{t=t_0}^{t_1}\sum_{i=1}^{n} |y_{t,i} - \bar{y}_{t_0:t_1,1:n}|}. \qquad (17)$$

A métrica next é o erro relativo de raiz ao quadrado (RSE):

$$\text{RSE} = \frac{\sqrt{\sum_{t=t_0}^{t_1} \sum_{i=1}^{n} (y_{t,i} - \hat{y}_{t,i})^2}}{\sqrt{\sum_{t=t_0}^{t_1} \sum_{i=1}^{n} (\hat{y}_{t,i} - \overline{\hat{y}_{t_0:t_1,1:n}})^2}} \quad (18)$$

Tabela2 ResultsontypicalMTSdatasetsusingRAE,RSEandCORRasmetrics

RAE	SolarEnergy			Traffic			
Horizon	3	6	12	24	3	6	12	24
AR	0.1846	0.3242	0.5637	0.9221	0.4491	0.4610	0.4700	0.4696
LRidge	0.1227	0.2098	0.4070	0.6977	0.4965	0.5115	0.5198	0.4846
LSVR	0.1082	0.2451	0.4362	0.6180	0.4629	0.5483	0.7454	0.4761
GP	0.1419	0.2189	0.4095	0.7599	0.5148	0.5759	0.5316	0.4829
SETAR	0.1285	0.1962	0.2611	0.3147	—	0.3226	0.3372	0.3368	—	—	0.3348	—
LSTNet-Skip	0.0985	0.1554	0.2018	—	0.3551	0.3287	0.3627	0.3518	0.3852	—	—	—
LSTNet-Attn	**0.0900**	0.1332	0.2202	0.4308	0.3196	—	0.3277	0.3557	0.3666	—	—	—
Nosso modelo	0.0918 ± 0.0005	**0.1296** ± **0.0008**	**0.1902** ± **0.0021**	**0.2727** ± **0.0045**	**0.2901** ± **0.0095**	**0.2999** ± **0.0022**	**0.3112** ± **0.0015**	**0.3118** ± **0.0034**

RaeElectricityExchangeRate								
AR	0.0579	0.0598	0.0603	0.0611	0.0181	0.0224	0.0291	0.0378	—	—	—	—
LRidge	0.0900	0.0933	0.1268	0.0779	0.0144	0.0225	0.0358	0.0602	—	—	—	—
LSVR	0.0858	0.0816	0.0762	0.0690	0.0148	0.0231	0.0360	0.0576	—	—	—	—
GP	0.0907	0.1137	0.1043	0.0776	0.0230	0.0239	0.0355	0.0547	—	—	—	—
SETAR	0.0475	—	0.0524	0.0545	0.0565	**0.0136**	0.0199	0.0288	0.0425	—	—	—
LSTNet-Skip	0.0509	0.0587	0.0598	0.0561	—	—	—	—	0.0180	0.0226	0.0296	0.0378	—	—	—	—
LSTNet-Attn	0.0515	0.0543	0.0561	0.0579	0.0229	0.0269	0.0384	0.0517	—	—	—	—
Nosso	**0.0463** ± **0.0007**	**0.0491** ± **0.0007**	**0.0541** ± **0.0006**	**0.0544** ± **0.0007**	0.0139 ± 0.0001	**0.0192** ± **0.0002**	**0.0280** ± **0.0006**	**0.0372** ± **0.0005**

RSE	SolarEnergy			Traffic			
AR	0.2435	0.3790	0.5911	0.8699	0.5991	0.6218	0.6252	0.6293
LRidge	0.2019	0.2954	0.4832	0.7287	0.5833	0.5920	0.6148	0.6025
LSVR	0.2021	0.2999	0.4846	0.7300	0.5740	0.6580	0.7714	0.5909
GP	0.2259	0.3286	0.5200	0.7973	0.6082	0.6772	0.6406	0.5995
SETAR	0.2374	0.3381	0.4394	0.5271	0.4611	—	0.4805	0.4846	0.4898

Tabela2contínuo

RSESolarEnergyTraffic																
LSTNet-Skip	0.1843		0.2559		0.3254				0.4643		0.4777		0.4893		0.4950	0.4973
LSTNet-Attn	0.1816		0.2538		0.3466		0.4403				0.4897		0.4973		0.5173	0.5300
Nosso	**0.1803**	±	**0.0008**	**0.2347**	±	**0.0017**	**0.3234**	±	**0.0044**	**0.4389**	±	**0.0084**	**0.4487**	±	**0.0180**	**0.4658**	±	**0.0053**	**0.4641**	±	**0.0034**	**0.4765**	±	**0.0068**
RSEElectricityExchangeRate																
AR	0.0995		0.1035		0.1050		0.1054		0.0228		0.0279		0.0353			
LRidge	0.1467		0.1419		0.2129		0.1280		0.0184		0.0274		0.0419		0.0675	
LSVR	0.1523		0.1372		0.1333		0.1180		0.0189		0.0284		0.0425		0.0662	
GP	0.1500		0.1907		0.1621		0.1273		0.0239		0.0272		0.0394		0.0580	
SETAR	0.0901		0.1020		0.1048		0.1009		0.0178							
LSTNet-Skip	0.0864		0.0931		0.1007		0.1007				0.0226		0.0280		0.0356	0.0449
LSTNet-Attn	0.0868		0.0953		0.0984				0.1059		0.0276		0.0321		0.0448	0.0590
Nosso	**0.0823**	±	**0.0012**	**0.0916**	±	**0.0018**	**0.0964**	±	**0.0015**	**0.1006**	±	**0.0015**	**0.0174**	±	**0.0001**	**0.0241**	±	**0.0004**	**0.0341**	±	**0.0011**	**0.0444**	±	**0.0006**
CORRSolarEnergyTraffic																
AR	0.9710		0.9263		0.8107		0.5314		0.7752		0.7568		0.7544		0.7519	
LRidge	0.9807		0.9568		0.8765		0.6803		0.8038		0.8051		0.7879		0.7862	
LSVR	0.9807		0.9562		0.8764		0.6789		0.7993		0.7267		0.6711		0.7850	
GP	0.9751		0.9448		0.8518		0.5971		0.7831		0.7406		0.7671		0.7909	
SETAR	0.9744		0.9436		0.8974		0.8420		0.8641		0.8506		0.8465		0.8443	
LSTNet-Skip	0.9843		0.9690		0.9467				0.8870		0.8721				0.8690	0.8614		0.8588
LSTNet-Attn	0.9848		0.9696		0.9397		0.8995				0.8704		0.8669		0.8540	0.8429
Nosso	**0.9850**	±	**0.0001**	**0.9742**	±	**0.0003**	**0.9487**	±	**0.0023**	**0.9081**	±	**0.0151**	**0.8812**	±	**0.0089**	**0.8717**	±	**0.0034**	**0.8717**	±	**0.0021**	**0.8629**	±	**0.0027**

Tabela2contínuo

CORRSolarEnergyTraffic								
CorrElectricityExchangeRate								
AR0.8845	0.8632	0.8591	0.8595	0.9734	0.9656	0.9526	0.9357	
LRidge0.8890	0.8594	0.8003	0.8806	0.9788	———	**0.9722**	0.9543	0.9305
LSVR0.8888	0.8861	0.8961	0.8891	0.9782	0.9697	0.9546	———	0.9370
GP0.8670	0.8334	0.8394	0.8818	0.8713	0.8193	0.8484	0.8278	
SETAR0.9402	———	0.9294	0.9202	**0.9171**	0.9759	0.9675	0.9518	0.9314
LSTNet-Skip0.9283	0.9135	0.9077	0.9119	0.9735	0.9658	0.9511	0.9354	
LSTNet-Attn0.9243	0.9095	0.9030	0.9025	0.9717	0.9656	0.9499	0.9339	
Nosso	**0.9429** ± **0.0004**	**0.9337** ± **0.0011**	**0.9250** ± **0.0013**	0.9133 ± 0.0008	**0.9790** ± **0.0003**	0.9709 ± 0.0003	**0.9564** ± **0.0005**	**0.9381** ± **0.0008**

Bestperformanceinboldface;secondbestperformanceisunderlined. Wereportthemeanandstandarddeviationofourmodelintenruns.Allnumbersbe sidestheresultsofour modelisreferencedfromthepaperofLSTNetLaietal. (2018)

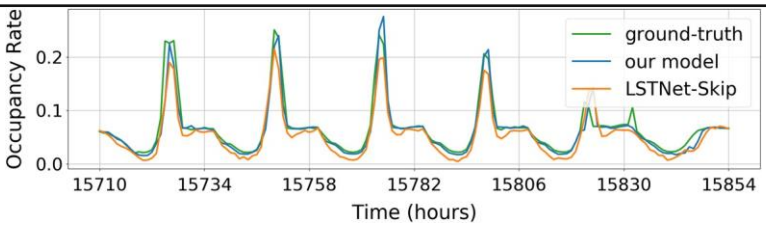

Fig. 5 Resultados de previsão para o modelo proposto e LSTNet-Skip no teste de tráfego definido com horizonte de 3 horas.
Modelo proposto claramente produz melhores previsões em torno da linha plana após o pico e no vale

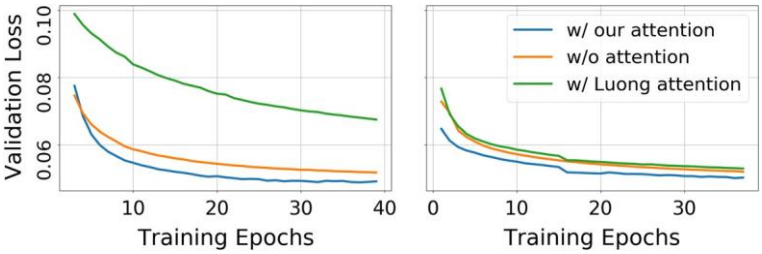

Fig. 6 Perda de validação em diferentes épocas de treinamento no MuseData (esquerda), e LPD-5-Cleansed

(direita) e, finalmente, a terceira métrica é o coeficiente de coeficiente de coecimento empírico de coerelação $\frac{\sum_{t=t_0}^{t_1}(y_{t,i} - \overline{y_{t_0:t_1,i}})(\hat{y}_{t,i} - \overline{\hat{y}_{t_0:t_1,i}})}{\sqrt{\sum_{t=t_0}^{t_1}(y_{t,i} - \overline{y_{t_0:t_1,i}})^2 \sum_{t=t_0}^{t_1}(\hat{y}_{t,i} - \overline{\hat{y}_{t_0:t_1,i}})^2}}$ (CORR):

$$\text{CORR} = \frac{1}{n}\sum_{i=1}^{n} \qquad (19)$$

onde y, y... é definido em Seita. 4.1, $y\forall_t \in$ [t0,t1] é o valor da verdade dos dados de teste, e y denota a média do conjunto y. RAE e RSE desconsideram a escala de dados e é uma versão normalizada do erro absoluto médio (MAE) e do erro quadrado médio raiz (RMSE), respectivamente. Para RAE e RSE, quanto menor, melhor, enquanto para CORR, melhor.

Para decidir qual modelo é melhor em conjuntos de dados de música polifônica, usamos perda de validação (probabilidade negativa de log), precisão, recall e pontuação de F1 como medidas amplamente utilizadas no trabalho na geração de música polifônica (Boulanger-Lewandowski et al. 2012; Chuan e Herremans 2018).

6.5 Resultados em conjuntos de dados típicos do MTS

OntypicalMTSdatasets,wechosethebestmodelonthevalidationsetusingRAE/RSE/CORR como a métrica para o conjunto de testes. Os resultados numéricos são tabulados na Tabela 2, onde a métrica das duas primeiras tabelas são RAE, seguida por duas tabelas de métrica RSE, e terminadas por outras duas tabelas usando métrica CORR. Ambas as tabelas

mostram que o modelo proposto supera quase todos os outros métodos em todos os conjuntos de dados, horizontes e métricas. Além disso, nossos modelos são capazes de deal com uma ampla gama de tamanho de conjunto de dados, desde o menor conjunto de dados de 534 KB Exchange Rate até o maior conjunto de dados de Energia Solar de 172 MB. Nesses resultados, o modelo proposto demonstra consistentemente sua superioridade para a previsão de MTS.

Na comparação com lSTNet-Skip e LSTNet-Attn, os métodos de última geração anteriores, o modelo proposto exibe desempenho superior, especialmente em Tráfego e Eletricidade,

Tabela 3 Precisão, recall e pontuação F1 de diferentes modelos em conjuntos de dados de música polifônica

Métrica	MuseData		
	Precisão	Lembrar	F1
Atenção de W/o	0.84009	0.67657	0.74952
Atenção de W/ Luong	0.75197	0.52839	0.62066
Atenção proposta por W/	**0.85581**	**0.68889**	**0.76333**
	LPD-5-Cleansed		
Atenção de W/o		0.837940.73041	0.78049
Atenção de W/ Luong		0,835480.72380	0.77564
Atenção proposta por W/		**0.839790.74517**	**0.78966**

Valores em negrito indicam melhor desempenho

que contêm a maior quantidade de séries tempores. Além disso, na Taxa de Câmbio, onde não existe um padrão repetitivo, o modelo proposto ainda é o melhor no geral; o desempenho do LSTNet-Skip e LSTNet-Attn está atrás dos métodos tradicionais, incluindo AR, LRidge, LSVR, GP e SETAR. Em Fig. 5 também visualizamos e comparamos a previsão do modelo proposto e LSTNet-Skip.

Em resumo, o modelo proposto alcança desempenho de última geração em conjuntos de dados MTS periódicos e não periódicos.

6.6 Resultados em conjuntos de dados de música polifônica

Nesta subseção, para verificar melhor a eficácia e a capacidade de generalização do modelo proposto para dados discretos, descrevemos experimentos realizados em conjuntos de dados de música polifônica; os resultados são mostrados em Fig. 6 e Tabela 3. Nós compared três modelos RNN: LSTM, LSTM com luong atenção, e LSTM com o mecanismo de atenção proposto. A Figura 6 mostra a perda de validação em épocas de treinamento, e na Tabela 3, usamos os modelos com a menor perda de validação para calcular precisão, recuo e pontuação de F1 no conjunto de testes.

A partir dos resultados, verificamos primeiro nossa alegação de que o mecanismo de atenção típico não funciona em tais tarefas, como em hiperparmetros semelhantes e pesos de formação, LSTM e o modelo proposto superam tais mecanismos de attention. Além disso, o modelo proposto também aprende de forma mais eficaz em comparação com o LSTM durante todo o processo de aprendizagem e produz melhor desempenho em termos de precisão, recall e pontuação de F1.

6.7 Análise dos filtros cnn

DFT é uma variante da transformação fourier (FT) que lida com amostras igualmente espaçadas de um sinal no tempo. No campo da análise de séries temporizadas, há um amplo conjunto de trabalhos que utiliza FT ou DFT para revelar características importantes em séries temporizadas (Huang et al. 1998; Bloomfield 1976). No nosso caso, uma vez que os dados do MTS também são igualmente espaçados e discretos, poderíamos aplicar DFT para analisá-los. No entanto, nos dados do MTS, há mais de uma série de tempo, por isso, naturalmente, nós naturalmente mediamos a variedade decomponentes de todos os tempos, representação de domínio de frequência andarriveatasingle. Denotamos isso como a média discreta da transformação fourier (avg-DFT). A representação de domínio de frequência única revela os componentes de frequência predominantes dos dados MTS. Por exemplo, é razoável assumir uma notável oscilação de 24 horas em Fig. 5, que é verificado pelo avg-DFT do conjunto de dados de tráfego mostrado na Fig. 7.

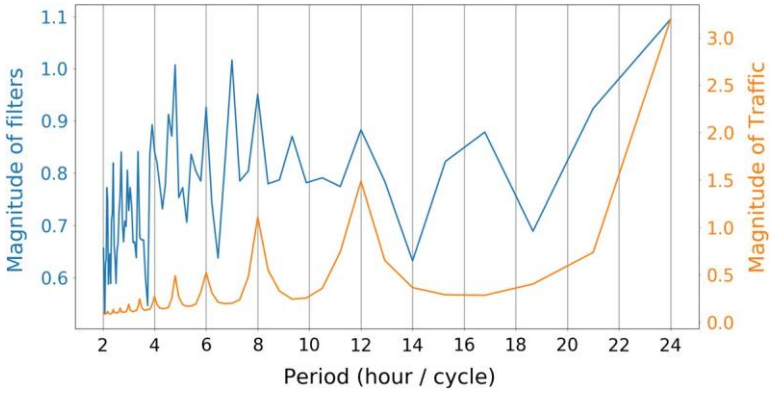

Fig. 7 Comparação de magnitude de (1) DFT de filtros CNN treinados no Tráfego com um horizonte de 3 horas, e (2) cada janela do conjunto de dados do Tráfego. Para tornar a figura mais intuitiva, a unidade do eixo horizontal é o período

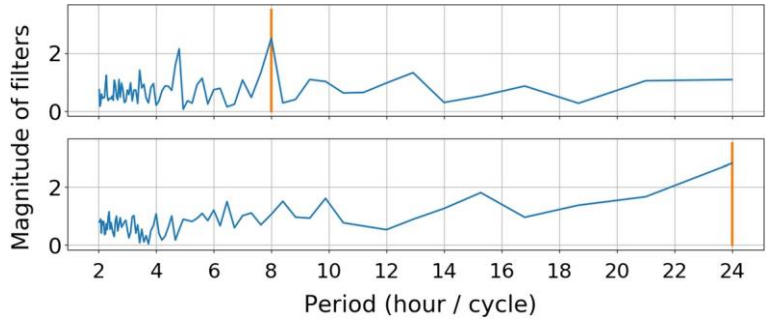

Fig. 8 Dois filtros diferentes da CNN treinaram no tráfego com um horizonte de 3 horas, que detectam diferentes períodos de padrões temporais

Uma vez que esperamos que nossos filtros CNN aprendam padrões de MTS temporais, os componentes de frequência predominantes nos filtros cnn médios devem ser semelhantes aos dos dados MTS de treinamento. Assim, também aplicamos avg-DFT no $k = 32$ filtros CNN que são treinados no Tráfego com um horizonte de 3 horas; em Fig. 7 nós plotamos o resultado junto com o avg-DFT de cada janela de conjunto de dados de tráfego. Impressionantemente, as duas curvas atingem picos nos mesmos períodos na maior parte do tempo, o que implica que os filtros cnn aprendidos se assemelham a bases em DFT. Nos períodos de 24, 12, 8 e 6 horas, não só a magnitude do conjunto de dados do Tráfego está no seu auge, mas a magnitude dos filtros da CNN também supera. Além disso, em Fig. 8, mostramos que diferentes filtros da CNN se comportam de forma diferente. Alguns se especializam em capturar padrões temporais de longo prazo (24 horas), enquanto outros são bons em reconhecer padrões temporais de curto prazo (8 horas). Como um todo, sugerimos que as filters propostas da CNN desempenham o papel de bases no DFT. Como demonstrado no trabalho de Rippel et al. (2015), esse "domínio de frequência" serve como uma representação poderosa para a CNN usar em treinamento e modelagem. Assim, o LSTM se baseia nas informações de domínio de frequência extraídas pelo mecanismo de atenção proposto para prever com precisão o futuro.

Tabela4 Estudo de

DatasetSolarEnergy (horizon=24)			Tráfego (horizon=24)		
	PositionFilterW/oCNN	PositionFilterW/oCNN			
Softmax0.4397	——— ± 0.0089	0.4414 ± 0.00930.4502	± 0.0099	**0.4696** ± **0.0062**	0.4832 ± 0.01090.4810	± 0.0083
Sigmóide	**0.4389** ± **0.0084**	0.4598 ± 0.00110.4639	± 0.01010.4765	——— ± 0.0068	0.4785 ± 0.00690.4803	± 0.0104
Concat0.4431	± 0.01000.4454	± 0.00930.4851	± 0.00490.4812	± 0.00820.4783	± 0.00770.4779	± 0.0073
Conjunto de dadosEletricidade (horizon=24)			MuseData			
Softmax	**0.0997** ± **0.0012**	0.1007 ± 0.00130.1010	± 0.00110.04923	——— ± 0.0037	0.04929 ± 0.00310.04951	± 0.0041
Sigmoid0.1006	——— ± 0.0015	0.1022 ± 0.00090.1013	± 0.0011	**0.04882** ± **0.0031**	0.04958 ± 0.00280.04979	± 0.0027
Concat0.1021	± 0.00170.1065	± 0.00290.1012	± 0.00080.05163	± 0.00400.05179	± 0.00360.05112	± 0.0027

AvaliaçãometricforSolarEnergy, Traffic andElectricityisRSE, andnegativelog-likelihoodforMuseData.WereportthemeanandstandarddeviationintenruIns. Oneach corpus,boldtextrepresentsthebestandunderlinedtextrepresentssecondbest

6.8 Estudo de ablação

A fim de verificar se a melhoria acima vem de cada componente adicionado em vez de um conjunto específico de hiperparmetros, realizamos um estudo de ablação sobre os conjuntos de dados Solar Energy, Traffic, Electricity e MuseData. Havia duas configurações principais: uma controlando como atendemos a estados ocultos, H, de RNN e a outra controlando como integramos a função de pontuação f no modelo proposto, ou mesmo desativar a function. Primeiro, no método proposto, deixamos o modelo atender aos valores de vários filtros em cada posição ($_{HiC}$); também podemos considerar o atendimento aos valores dos mesmos filtros em várias posições ((HC)$_i$) ou vetores de linha de H (Hi). Essas três approaches diferentes correspondem aos cabeçalhos da coluna na Tabela 4: "Posição", "Filtro" e "Sem CNN". Em segundo lugar, enquanto no mecanismo de atenção típico, o softmax é geralmente usado no valor de saída da função de pontuação f para extrair as informações mais relevantes, usamos sigmoid como nossa função de ativação. Portanto, comparamos essas duas funções diferentes. Outra possível estrutura para previsão é concatenar todos os estados ocultos anteriores e deixar o modelo automaticamente aprender quais valores são importantes. Levando em consideração esses dois grupos de configurações, treinamos modelos com todas as combinações de estruturas possíveis nesses quatro conjuntos de dados.

Os resultados do MuseData mostram que o modelo com ativação sigmoid e atenção no $_{HiC}$ (posição) é claramente o melhor, which sugere que o modelo proposto é razoavelmente eficaz para a previsão. Não importa qual componente proposto seja removido do modelo, gotas de desempenho. Forexample, usando ooftmaxinsteadofsigmoidraisesthenegativelog-probabilidade de 0. 04882 a 0. 04923; obtemos um modelo ainda pior com probabilidade negativa de log de 0. 4979 se não usarmos filtros CNN. Além disso, não observamos nenhuma melhora significativa entre o modelo proposto e esse modelo usando softmax nos três primeiros conjuntos de dados na Tabela 4: Energia Solar, Tráfego e Eletricidade. Isso não é surpreendente, dada a nossa motivação para o uso de sigmoid, como explicado na Seita. 4.3. Originalmente, esperávamos que os filtros da CNN encontrassem padrões básicos e esperávamos que a função sigmoid ajudasse o modelo a combinar esses padrões em um que ajuda. No entanto, devido à natureza fortemente periódica desses três conjuntos de dados, é possível que o uso de um pequeno número de padrões básicos seja suficiente para uma boa previsão. No geral, no entanto, o modelo proposto é mais geral e produz resultados competitivos de ând estável em diferentes conjuntos de dados.

7 Conclusões

Neste artigo, focamos na previsão do MTS e propomos um novo mecanismo de atenção ao padrão temporal que remove a limitação dos mecanismos típicos de atenção nessas tarefas. Permitimos que a dimensão de attention seja em termos de características para que o modelo aprenda interdependências entre múltiplas variáveis não apenas dentro do mesmo passo de tempo, mas também em todos os tempos e séries anteriores. Nossos experimentos em exemplos de brinquedos e conjuntos de dados do mundo real apoiam fortemente essa ideia e mostram que o modelo proposto alcança resultados de última geração. Além disso, a visualização de filtros também verifica nossa motivação de forma mais compreensível para os seres humanos.

Reconhecimentos Este trabalho foi apoiado financeiramente pelo Ministério da Ciência e Tecnologia de Taiwan.

Referências

Bahdanau, D., Cho, K., & Bengio, Y. (2015). *Tradução de máquina neural aprendendo conjuntamente a alinhar e traduzir*. O ICLR.

Bloomfield, P. (1976). *Análise de fourier da série temporal: Uma introdução*. Nova Iorque, NY: Wiley.

Bouchachia, A., & Bouchachia, S. (2008). Ensemble aprendendo para a previsão da série temporal. *Tramitação do 1º workshop internacional sobre dinâmica não linear e sincronização*.

Box, G.E., Reinsel, G.C., Jenkins, G.M., & Ljung, G.M. (2015). *Análise da série temporal: Previsão e controle*. Hoboken, NJ: Wiley.

Cao, L. J., & Tay, F. E. H. (2003). Suporte a máquina vetorial com parâmetros adaptativos na previsão de séries de tempo financeiro. *IEEE Transactions on Neural Networks*, pp. 1506-1518.

Chen, S., Wang, X. X., & Harris, C. J. (2008). Identificação do sistema não linear com base narx usando a caça da base ortogonal menos squares. *IEEE Transactions on Control Systems*, pp. 78-84.

Cho, K., Bahdanau, D., Van Merrienboer, B., & Bengio, Y. (2014). Nas propriedades da tradução da máquina neural: Abordagens codificadores-decodificadores. arXiv pré-impressão arXiv:14091259.

Chuan, C. H., & Herremans, D. (2018). Modelando relações tonais temporais na música polifônica através de redes profundas com representação baseada em imagens. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/ papel/vista/16679.

Connor, J., Atlas, L.E., & Martin, D. R. (1991). Redes recorrentes e modelagem NARMA. *Avanços nos Sistemas de Processamento de Informações Neurais*, pp. 301-308.

Dasgupta, S., & Osogami, T. (2017). Machines boltzmann dinâmicos não lineares para previsão de séries temporalhais.

Dong, H.-W., Yang, L.C., Hsiao, W.-Y., & Yang, Y. H. (2018). MuseGAN: Redes de contraditórios sequenciais multi-faixas para geração e acompanhamento simbólicos da música.

Elman, J. L. (1990). Encontrar estrutura a tempo. *Ciência Cognitiva*, pp. 179-211.

Frigola, R., & Rasmussen, C. E. (2014). Pré-processamento integrado para identificação do sistema não linear bayesiano com processos gaussianos. *Conferência IEEE sobre Decisão e Controle*, pp. 552-560.

Frigola-Alcade, R. (2015). *Séries temporais bayesianas aprendendo com processos gaussianos*. Tese de Doutorado, Universidade de Cambridge.

Hochreiter, S., & Schmidhuber, J. (1997). Memória de longo prazo. *Computação Neural*, *9*(8), 1735-1780. ht tps://doi.org/10.1162/neco.1997.9.8.1735.

Huang, N. E., Shen, Z., Long, S.R., Wu, M.C., Shih, H. H., Zheng, Q., et al. (1998). A decomposição do modo empírico e o espectro Hilbert para análise de séries temporâneas não lineares e não estacionárias. *Procedimentos da Sociedade Real de Londres. Série A*, *454*, 903-995.

Jain,A.,&Kumar, A.M.(2007). Modelos híbridos de redes para aquessagem de tempohidológico. *Computação macia aplicada*, *7*(2), 585-592.

Kim, K. J. (2003). Séries de tempo financeiro prevendo usando máquinas vetoriais de suporte. *Neurocomução*, *55*(1), 307-319.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Classificação imagenet com redes neurais convolucionais profundas. *Avanços nos Sistemas de Processamento de Informação Neural*, pp. 1097-1105.

Lai, G., Chang, W.C., Yang, Y., & Liu, H. (2018). Modelando padrões temporais de longo e curto prazo com redes neurais profundas. *SIGIR*, pp. 95-104.

LeCun, Y., & Bengio, Y. (1995). Redes convolucionais para imagens, fala e séries temporâneas. *O manual da teoria cerebral e das redes neurais*.

Luong, T., Pham, H., & Manning, C. D. (2015). Abordagens eficazes para a tradução da máquina neural baseada em atenção. Em *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1412-1421.

Nicolas Boulanger-Lewandowski, Y.B., & Vincent, P. (2012). Modelando dependências temporais em sequências highdimensionais : Aplicação à geração e transcrição da música polifônica.

Qin, Y.,Song, D., Cheng, H.,Cheng, W. ,Jiang,G.,&Cottrell, G.W.(2017). Rede neural de corrente de atenção à fase adual para previsão de séries temporais. Em *IJCAI'17*, pp. 2627-2633. http://dl.acm.org/citation. cfm?id=3172077.3172254.

Raffel, C. (2016). *Métodos baseados em aprendizagem para comparar sequências, com aplicativos ao alinhamento e correspondência áudio-MIDI*. Tese de doutorado.

Rippel, O., Snoek, J., & Adams, R. P. (2015). Representações espectrais para redes neurais convolucionais. *NIPS*, pp. 2449-2457.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., & Aigrain, S. (2011). Processos gaussianos para modelagem de séries temporicientes. *Transações Filosóficas da Sociedade Real A.*

Rumelhart, D.E., Hinton, G. E., & Williams, R. J. (1986). Aprendendo representações by erros de retropropagação. *Natureza*, pp. 533-536.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequência para sequência de aprendizado com redes neurais. *Avanços nos Sistemas de Processamento de Informações Neurais*, pp. 3104-3112.

Tong, H., & Lim, K. S. (2009). Autoregressão de limiar, ciclos de limite e dados cíclicos. Em *Exploração de um mundo não linear: Uma apreciação das contribuições de Howell Tong para as estatísticas*, World Scientific, pp. 9-56.

Vapnik, V., Golowich, S.E.,&Smola,A.(1997). O método de suporte para a aproximação da função, a estimativa de regressão e o processamento de sinais. *Avanços nos Sistemas de Processamento de Informações Neurais*, pp. 281-287.

Werbos, P. J. (1990). Backpropagation através do tempo: O que ele faz e como fazê-lo. *Tramitação do IEEE*, pp. 1550-1560.

Zhang, G. P. (2003). Previsão da série Time usando um modelo híbrido de ARIMA e rede neural. *Neurocomução*, pp. 159-175.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Previsão com redes neurais artificiais: O estado da arte. *International Journal of Forecasting*, pp. 35-62.