文章编号：2095-2163(2021)12-0032-05 中图分类号：TP391 文献标志码：A

LDA 最大概率填充与 BiLSTM 模型的文本分类研究

袁丽莉，侯 磊，张正平
(贵州大学 大数据与信息工程学院，贵阳 550025)

摘 要：文本分类是自然语言处理(NLP)领域中的基础任务，双向长短时记忆网络(BiLSTM)具有遗忘细胞状态中的信息和记忆新的信息、在上下文中依赖能力较好的优势。为进一步增强文本的特征表达，本文提出一种基于 LDA 的最大概率填充模型。首先，运用 Word2Vec 词嵌入方式生成文本向量；其次，根据 LDA 模型对文本向量矩阵进行填充，丰富语义信息，采用 BiLSTM_Attention 模型训练填充后的向量矩阵；最后，采用 softmax 进行分类。实验结果表明，本文提出的方法在 IMDB 电影评论分析数据集中的分类准确率为 98.43%，相较于单向的 RNN 模型提高 1.63%，比双向的 BiLSTM_Attention 模型提高0.83%。

关键词：文本分类；LDA 模型；BiLSTM_Attention；Word2Vec

Research on text classification based on LDA maximum probability filling and BiLSTM model

YUAN Lili, HOU Lei, ZHANG Zhengping
(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 Text classification is a basic task in the field of natural language processing (NLP). Bi-directional long and short-term memory network (BiLSTM) has the advantages of forgetting the information in the cell state and memorizing new information and the ability to rely on the context. In order to further enhance the feature expression of the text, this paper proposes a maximum probability filling model based on LDA. First, Word2Vec word embedding method is used to generate text vector. Secondly, the text vector matrix is filled according to LDA model to enrich semantic information. The BiLSTM_Attention model is used to train the filled vector matrix, and finally softmax is used for classification. Experimental results show that the classification accuracy of the proposed method in the IMDB movie review analysis dataset is 98.43%, which is 1.63% higher than the one-way RNN model and 0.83% higher than the two-way BiLSTM_Attention model.

【Key words】 text classification; LDA model; BiLSTM_Attention; Word2Vec

0 引 言

随着大数据和“互联网+”的高速智能化发展，互联网已发展成为当今世界上最大的信息资源库，而大多数信息是以文本方式呈现出来的[1]。从这些海量且杂乱无章的数据文本中筛选出所需的有用信息是当前数据挖掘领域的研究热点。文本分类是自然语言处理(NLP)领域中的基础任务，能将复杂的文本信息有效的组织和管理，并已广泛用于信息检索、自动聊天系统、垃圾邮件过滤等领域。

目前，常用的文本分类方法可分为两类，一类是基于传统的机器学习，例如支持向量机(SVM)、决策树等；另一类是基于深度神经网络的文本分类方法，该种方法通常采用卷积神经网络(CNN)和循环神经网络(RNN)，其一般流程为文本预处理、结构化表示、特征选择和分类器构建，通过学习训练得到分类模型，最后利用分类模型对测试样本进行预测，从而达到对文本分类的目的[2]。通过对当前文本分类的研究不难发现，文本表示是文本分类中的难点之一，尤其是在文本的语义表达方面最为困难。文本在机器学习中表示大多采用向量空间模型，该模型容易造成向量维数过大，数据稀疏的问题，从而影响最终分类的结果。因此，如何将文本表示为机器可以理解的形式，又保留文本原有的潜在语义至关重要[3-4]。文献[5]提出一种基于 LDA 主题模型和 Word2Vec 词向量模型，完成对文本词向量的构建，结合神经网络对构建的词向量获取联合特征的方法，实现文本分类；文献[6]采用三层 CNN 模型，提取文本的局部特征，整合出全文的语义，利用长短时记忆网络(LSTM)存储历史信息的特征，以获取

基金项目：国家自然科学基金(61865002)。

作者简介：袁丽莉(1997-)，女，硕士研究生，主要研究方向：数据挖掘；侯 磊(1987-)，男，博士，讲师，主要研究方向：通信编码；张正平(1964-)，男，博士，教授，主要研究方向：通信与信息系统。

通讯作者：侯 磊 Email：llyuan_xb@126.com

收稿日期：2021-09-08

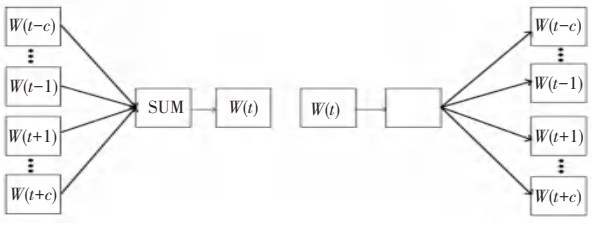
文本的上下文关联语义；文献[7]采用词向量完成原始文本的数字化，利用双向长短时记忆网络(BiLSTM)进行语义的提取，同时采用改进的注意力层(Attention)融合正向和反向特征，获得具有深层语义特征的短文本向量表示；文献[8]提出一种BiLSTM 和 CNN 混合神经网络文本分类方法，两种神经网络的结合充分发挥了 CNN 的特征提取能力和 BiLSTM 的上下文依赖能力，同时采用注意力机制提取信息的注意力分值，增强模型的特征表达能力。

基于上述的研究，经过 word2vec 形成的词向量矩阵，由于文档集里的各个文本长度不一，造成词向量矩阵的行数不一，在实验过程不能批量处理数据。本文提出一种基于 LDA 的最大概率填充模型，该模型对词向量矩阵进行填充，使构建的词向量矩阵行数等于最大文本长度，丰富了语义信息，采用 BiLSTM_Attention 模型对通过填充后具有上下文丰富语义信息的文本词向量矩阵进行训练，达到分类器可以根据给定的标签信息对输入信息进行分类的目的。

1　LDA_BiLSTM_Attention 文本分类模型

1.1　词向量矩阵生成模块

分词过后的文本需要转化为计算机能识别的形式，目前主要采用 Word2Vec 模型。Word2Vec 包含了 CBOW 和 Skip-gram 两种模型，CBOW 模型利用词的前后各 C 个词来预测当前词，如图 1(a)所示；Skip-gram 模型则是利用当前词预测其前后各 C 个词，如图 1(b)所示。在 CBOW 模型中，输入层是词 $W(t)$ 的前后各 C 个词向量，投影层将这些词向量累加求和，输出层是一棵以训练数据中所有词作为叶子节点，以各词在数据中出现的次数作为权重的树[9]。最后应用随机梯度上升法预测投影层的结果作为输出，Skip-gram 模型与之类似。当获得所有词的词向量后，可发现这样的规律："king" + "woman" = "queen"，可见词向量有效表达了词语的语义信息[10]。



(a) CBOW 模型　　(b) Skip-gram 模型

图 1　LDA 模型

Fig. 1　LDA model

1.2　LDA 主题填充模型

判断两个文本是否相似，传统的方法是找出两个文本中共有的词个数，这种方法的缺点在于忽略了文本的语义信息，使得两个原本相似的文档因为没有共有的词而造成判断错误。所以在进行文本分类时，语义也是需要考虑的一个重要的因素[11-12]。本文采用的 LDA 模型就能够很好地解决文本的语义问题，通过文本混合主题上的概率分布选择一种主题，从被抽取到的主题上所对应的单词概率分布中抽取一个词，然后重复上述过程，直至遍历文档中的每一个单词。图 2 为 LDA 生成模型，K 为主题个数，M 为文档总数，N_m 是第 m 个文档的词数，β 是每个主题下词的多项分布的 Dirichlet 先验参数，α 是每个文档下主题的多项分布的 Dirichlet 先验参数，$z_{m,n}$ 是第 m 个文档中第 n 个词的主题，$w_{m,n}$ 是第 m 个文档中的第 n 个词，隐含变量 θ_m 和 ϕ_k 分别表示第 m 个文档下的主题分布和第 k 个主题下词的分布，前者是 $\boldsymbol{k}$ 维向量，后者是 $\boldsymbol{v}$ 维向量。主题模型学习参数主要是基于 Gibbs 采样和基于推断 EM 算法求解，Gibbs 采样算法是一种特殊的马氏链的方法，是经过对词的主题采样生成马氏链。马氏链的生成过程是根据所有词的其它时刻的主题分布估计当前词分配于各个主题的分布概率，当算法重新选择了一个与原先不同的主题词时，反过来会影响文本-主题矩阵和主题-词矩阵，这样不断地进行循环迭代，就会收敛到 LDA 的误差范围内。当完成主题采样后，就可以学习模型的最终训练结果，生成两个矩阵分别为文本-主题分布矩阵 θ 及主题 - 词分布矩阵 $\boldsymbol{\Phi}$，公式(1)和公式(2)如下：

$$\boldsymbol{\theta}_{m,k} = \frac{n_m^{(k)} + \alpha}{\sum_{k=1}^{k}(n_m^{(t)} + \alpha)} \tag{1}$$

$$\boldsymbol{\Phi}_{k,t} = \frac{n_k^{(t)} + \beta}{\sum_{t=1}^{|V|}(n_k^{(t)} + \beta)} \tag{2}$$

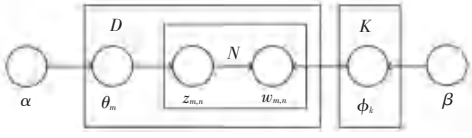


图 2　LDA 生成模型

Fig. 2　LDA generative model

1.3　最大概率的文本主题填充方式

运用词向量模型对输入文本进行词向量矩阵嵌入生成之后，因为文档里的文本长度长短不一样，导致文档集里的各个文本生成的词向量矩阵的大小各

不相等,在目前的处理方法中,通常采用填零法、循环法和随机法进行填充,导致构建的词向量矩阵存在稀疏性以及语义混乱等问题[5]。为了在实验中能够进行批处理数据和丰富文本特征信息,本文提出基于最大概率主题下的 LDA 填充方式,以文档集里的最大文本长度为基准,寻找文本对应文本-主题矩阵最大的概率主题,找到此主题下的词概率分布,依照概率大小将词映射为词向量,并依次对词向量矩阵进行填充,直至构建的词向量矩阵行数等于最大文本长度。填充流程如图 3 所示。

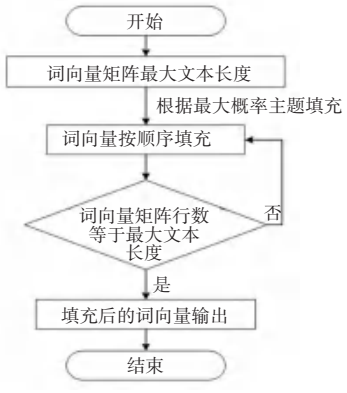


图 3 LDA 填充流程图

Fig. 3 LDA filling flow chart

1.4 BiLSTM 和注意力机制

例如:“The driver of this car was charged by speeding and hitting pedestrian”,若不联系后文则很难推断在此处 charge 是收费还是指控的意思,BiLSTM 双向捕捉能获得更细粒度的信息,提出了双向神经网络,结构如图 4 所示。

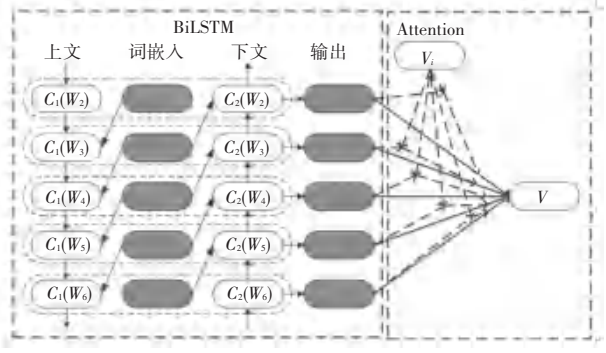


图 4 BiLSTM_Attention 结构图

Fig. 4 BiLSTM_Attention structure diagram

首先,使用 BiLSTM 学习当前词的上文词向量 $\boldsymbol{C}_1(\boldsymbol{W}_i)$ 和下文词向量 $\boldsymbol{C}_2(\boldsymbol{W}_i)$,再与当前的自身词向量 $\boldsymbol{C}(\boldsymbol{W}_i)$ 进行计算,公式(3) ~ 公式(5) 如下:

$$C_1(W_i) = f(W^{(1)}C_1(W_{i-1}) + W^{(s1)}e(W_{i-1})) \quad (3)$$

$$C_2(W_i) = f(W^{(2)}C_1(W_{i-1}) + W^{(s2)}e(W_{i+1})) \quad (4)$$

$$X_i = [C_1(W_i), e(W_i), C_2(W_i)] \quad (5)$$

将 X_i 作为 W_i 的语义特征,通过激活函数 tan h 得到的潜在语义信息 Y_i,公式(6) 和公式(7) 如下:

$$Y_i = \tan h(WX_i + b_i) \quad (6)$$

$$a_i = \frac{\exp(Y_i{}^T V_i)}{\sum_1^n \exp(Y_i{}^T V_i)} \quad (7)$$

注意力机制是一种权重分配的机制,通过模仿生物、观察行为,将内部经验和外部感觉对齐,进一步增强观察行为的精度,在数学模型上表示为通过计算注意力的概率分布来获得某个输入对输出的影响,该方法后来被引入到自然语言处理领域。图 4 中的 V_i 作为不同时刻的输出权重,公式(8) 对 BiLSTM 网络的输出进行加权求和。

$$V = \sum_1^n a_i Y_i \quad (8)$$

1.5 输出层

经典的全连接网络的输出层表示为公式(9):

$$y^{(2)} = W_2 y^{(1)} + b_1 \quad (9)$$

其中,W_2 为权重系数,b_1 为偏置项。

通过 *softmax* 函数分类,得到每个文本所在类别的概率分布,找出最大值的类别就是预测类别,计算公式(10)如下:

$$P_i = \frac{\exp(y_i^{(2)})}{\sum_{k=1}^n \exp(y_k^{(2)})} \quad (10)$$

2 性能评测与实验分析

2.1 实验数据

本次实验采用的数据集为 IMDB 电影评论分析数据集,共有 3 个部分:分别为带标签的训练集(labeledTrainData),不带标签的训练集(unlabeledTrainData)和测试集(testData),实验参数见表 1。

表 1 BiLSTM 网络参数

Tab. 1 BiLSTM network parameters

参数	值	参数	值
词向量维度	200	损失函数	交叉熵
隐藏层大小	256	参数 Epoch	10
学习率	0.001	Batch_size	128

2.2 实验评判标准

为客观评价本文提出的模型,将 IMDB 电影评论数据集按 8:2 的比例分为训练数据集和测试数据集。同时引入准确度(*accuracy*)、精确度

(*precision*)、召回率(*recall*)、综合评价指标(*F*1) 对实验结果定性分析,4 种指标计算公式如公式(11)~(14),各参数意义见表 2。

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (11)$$

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (12)$$

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (13)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \quad (14)$$

表 2　评价指标函数意义

Tab. 2　The meaning of evaluation index function

	分为 A 类	未分为 A 类
实际为 A 类	*TP*	*FN*
实际非 A 类	*FP*	*TN*

2.3　实验结果分析

本文实验涉及的开发工具与实验环境：

硬件环境：CPU：Inter(R)core(TM)i5-4210 M，内存：8 GB，硬盘：500 GB。

软件环境：Windows10(基于 X64 的处理器)，python3.6.8。

为验证本文模型的有效性，选取 RNN、BiLSTM 和 BiLSTM_Attention 3 种常规模型与本文提出的 LDA_BiLSTM_Attention 模型进行对比实验，实验结果见表 3。

表 3　评价指标函数意义

Tab. 3　The meaning of evaluation index function

模型	*accuracy*	*recall*	*F*1
RNN	96.8	96.7	95.9
BiLSTM	97.1	96.8	96.3
BiLSTM_Attention	97.6	97.4	97.1
本文模型	98.43	98.3	97.7

从表 3 中的数据可以看出，本文模型文本分类准确度达到 98.43%，比 BiLSTM_Attention 模型提高了0.83%，在召回率方面也表现很好，综合评价指标最佳。RNN 网络因不能双向捕捉特征值导致分类效果最差，加入了注意力机制的 BiLSTM 模型相对于单独的 BiLSTM 模型综合评价指标也有所提升。由此可知，在 BiLSTM_Attention 模型的基础上加入 LDA 算法填充词向量矩阵，能丰富语义，进一步捕获文本分类的信息，提高文本分类的准确率。

上述 4 种模型的损失值和准确率随迭代次数的变化曲线，如图 5 和图 6 所示。由图 5 和图 6 可以看出，采用 LDA 对词向量矩阵进行填充后训练的模型损失值最小，并且准确度最高，达到了 98.4%。采用 LDA 算法对词向量矩阵填充后进行分类的准确度相对于没有填充后进行分类的结果要高出 0.83%，因为本文提出的 LDA 模型对词向量矩阵进行填充，丰富了上下文的语义关系，使得分类的准确率更高，可解释性也更好。

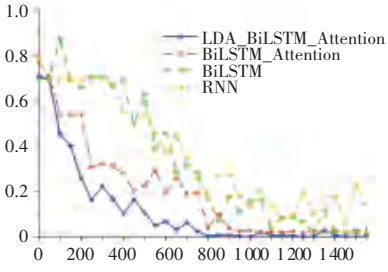


图 5　Loss 变化曲线

Fig. 5　Loss curve

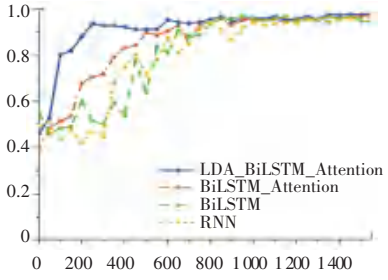


图 6　Accuracy 变化曲线

Fig. 6　Accuracy curve

上述 4 种模型的返回值和综合评价指标随迭代次数的变化曲线如图 7 和图 8 所示。可以看出本文提出的 LDA_BiLSTM_Attention 混合模型性能表现最佳，未加 LDA 填充算法的 BiLSTM_ Attention 模型性能表现次之，这说明 LDA 算法对词向量矩阵填充后使用混合 BiLSTM_Attention 结构作为模型主体的效果显著，充分发挥 BiLSTM 算法在长文本序列中获取历史信息的能力与 LDA 算法填充主题词向量矩阵的优势，从而提高了文本分类的综合指标。

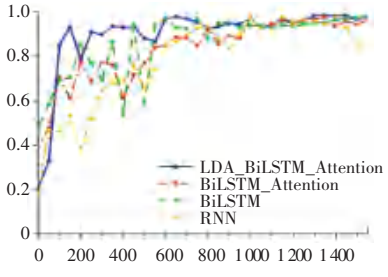


图 7　Recall 变化曲线

Fig. 7　Recall curve

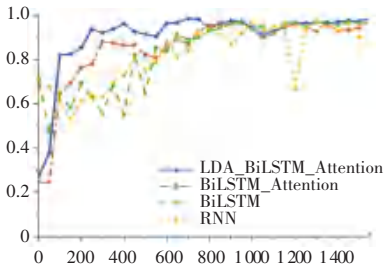


图 8 *F*1 变化曲线

Fig. 8 *F*1 curve

3 结束语

文本表示是文本分类的重要过程,针对文本复杂语义表达的问题,本文提出一种 LDA 最大概率主题填充模型来丰富文本词向量矩阵。首先,运用 word2vec 词嵌入方式生成文本向量;其次,根据 LDA 模型对文本向量进行填充,丰富语义信息,采用 BiLSTM_Attention 模型训练填充后的词向量矩阵;最后,采用 softmax 进行分类。通过与其他几种文本分类模型的对比可知,本文提出的 LDA 最大概率填充算法能有效提高文本分类的准确度。

参考文献

[1] 余本功,汲浩敏. 基于 DW-TCI 的半监督文本分类方法研究[J/OL]. 数据分析与知识发现 1-17[2020-09-13]. https://kns.cnki.net/kcms/detail/detail.aspx? FileName = XDTQ20200728000&DbName = CAPJ2020.

[2] 高云龙,吴川,朱明. 基于改进卷积神经网络的短文本分类模型[J]. 吉林大学学报(理学版),2020,58(4):923-930.

[3] 霍达,赵禹萌,张丽霞,等. 基于组合神经网络的舆情短文本表示模型降维方法研究[J/OL]. 内蒙古工业大学学报(自然科学版) 1-7[2020-09-13].

[4] 潘秋羽,王伟,王明明,等. 基于卷积特征建模的目标检测方法[J/OL]. 计算机应用研究 1-5[2020-09-13]. https://kns.cnki.net/kcms/detail/detail.aspx? FileName = JSYJ2020070700B&DbName = CAPJ2020.

[5] 郑飞,韦德壕,黄胜. 基于 LDA 和深度学习的文本分类方法[J]. 计算机工程与设计,2020,41(8):2184-2189.

[6] 王海涛,宋文,王辉. 一种基于 LSTM 和 CNN 混合模型的文本分类方法[J]. 小型微型计算机系统,2020,41(6):1163-1168.

[7] 陶志勇,李小兵,刘影,等. 基于双向长短时记忆网络的改进注意力短文本分类方法[J]. 数据分析与知识发现,2019,3(12):21-29.

[8] 万齐斌,董方敏,孙水发. 基于 BiLSTM-Attention-CNN 混合神经网络的文本分类方法[J]. 计算机应用与软件,2020,37(9):94-98,201.

[9] 马思丹,刘东苏. 基于加权 Word2vec 的文本分类方法研究[J]. 情报科学,2019,37(11):38-42.

[10] 彭俊利,谷雨,张震,等. 融合改进型 TC 与 word2vec 的文档表示方法[J/OL]. 计算机工程 1-7[2020-09-13]. https://kns.cnki.net/kcms/detail/detail.aspx? FileName = JSJC20200331001 & DbName = CAPJ2020.

[11] 陈欢,黄勃,朱翌民,等. 结合 LDA 与 Self-Attention 的短文本情感分类方法[J/OL]. 计算机工程与应用:1-8[2020-09-13]. https://kns.cnki.net/kcms/detail/detail.aspx? FileName = JSGG202018025&DbName = CJFQ2020.

[12] BUENAÑO-FERNANDEZ D, GONZÁLEZ M, GIL D, et al. Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach[J]. IEEE Access, 2020, 8: 35318-35330.

(上接第 31 页)

[4] 曹雷,陈洪斌,邱琪,等. 盲图像复原研究现状[J]. 中国光学,2014(1):68-78.

[5] KRISHNAN D, TAY T, FERGUS R. Blind deconvolution using a normalized sparsity measure[C]//CVPR 2011. IEEE, 2011: 233-240.

[6] LIN T C, HOU L, LIU H, et al. Reconstruction of single image from multiple blurry measured images[J]. IEEE Transactions on Image Processing, 2018, 27(6): 2762-2776.

[7] LI Y, CLARKE K C. Image deblurring for satellite imagery using small-support-regularized deconvolution[J]. ISPRS journal of photogrammetry and remote sensing, 2013, 85: 148-155.

[8] PAN J, SUN D, PFISTER H, et al. Deblurring images via dark channel prior[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(10): 2315-2328.

[9] YAN Y, REN W, GUO Y, et al. Image deblurring via extreme channels prior[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4003-4011.

[10] DONG W, TAO S, XU G, et al. Blind Deconvolution for Poissonian Blurred Image With Total Variation and L 0-Norm Gradient Regularizations[J]. IEEE Transactions on Image Processing, 2020, 30: 1030-1043.

[11] 徐宁珊,王琛,任国强,等. 混合梯度稀疏先验约束下的图像盲复原[J]. 光电工程,2021,48(6):58-69.

[12] 王允森,王勇,左晨,等. 基于维纳滤波和综合评价因子的遥感图像复原[J]. 空间电子技术,2021,18(3):13-20.

[13] PAN J, HU Z, SU Z, et al. Deblurring text images via L0-regularized intensity and gradient prior[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2901-2908.

[14] WEN F, YING R, LIU Y, et al. A simple local minimal intensity prior and an improved algorithm for blind image deblurring[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020.

[15] ZUO W, REN D, ZHANG D, et al. Learning iteration-wise generalized shrinkage-thresholding operators for blind deconvolution[J]. IEEE Transactions on Image Processing, 2016, 25(4): 1751-1764.

[16] KRISHNAN D, FERGUS R. Fast image deconvolution using hyper-Laplacian priors[J]. Advances in neural information processing systems, 2009, 22: 1033-1041.

[17] LIU Y, WANG J, CHO S, et al. A no-reference metric for evaluating the quality of motion deblurring[J]. ACM Trans. Graph., 2013, 32(6): 1-12.