研究生学习工作周报

院　　系	人工智能学院
专　　业	电子信息
姓　　名	佘依函
学　　号	231226006052
导　　师	周静 张俊驰
周报日期	2023 年 9 月 16 日

摘要

1. 机器学习知识复习：逻辑回归
2. 论文阅读：Robust Referring Video Object Segmentation with Cyclic Structural Consensus。End-to-End Referring Video Object Segmentation with Multimodal Transformers。
3. 工作内容：爬取领导留言板、爬取政法机构法律知识。
4. 下周安排及上周总结

目录

第一章 学习工作总结

1.1 机器学习知识复习

复习机器学习 week3 的知识内容，主要涉及分类问题、假说表示、判定边界、代价函数。

1.1.1 分类问题

如果我们要用线性回归算法来解决一个分类问题，对于分类， y 取值为 0 或者 1，但如果你使用的是线性回归，那么假设函数的输出值可能远大于 1，或者远小于 0，即使所有训练样本的标签 y 都等于 0 或 1。尽管我们知道标签应该取值 0 或者 1，但是如果算法得到的值远大于 1 或者远小于 0 的话，就会感觉很奇怪。所以我们在接下来的要研究的算法就叫做逻辑回归算法，这个算法的性质是：它的输出值永远在 0 到 1 之间。

1.1.2 逻辑回归

逻辑回归模型的假设：

$$h_\theta(x) = g(\theta^T X) \tag{1-1}$$

其中 X 表示特征向量，g 代表逻辑函数

Python 代码：

```
import numpy as np

def sigmoid(z):

   return 1 / (1 + np.exp(-z))
```

1.1.3 判定边界

决策边界(decision boundary)，需要运用函数来判定复杂的边界。

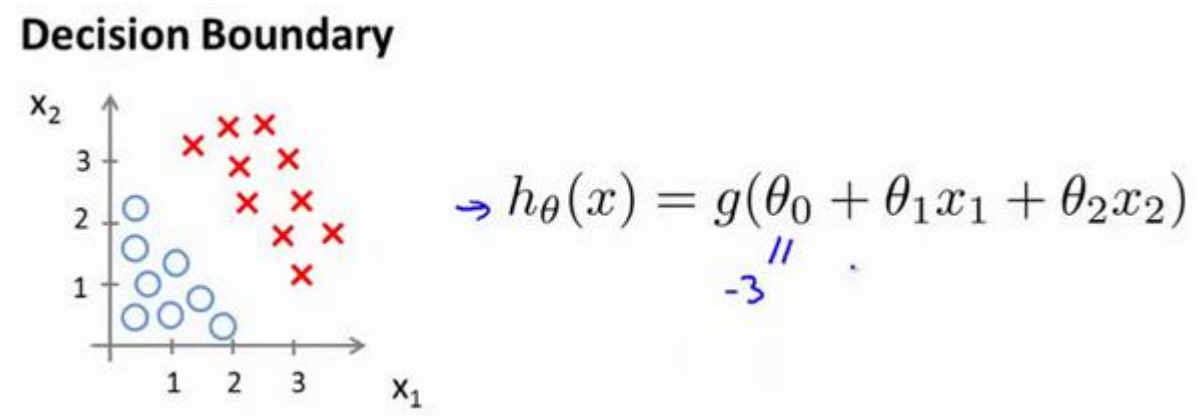


图 1.1 复杂边界举例

1.1.4 代价函数

与线性回归不同的是逻辑回归的代价函数是一个非凸函数如图 1.2 可知。

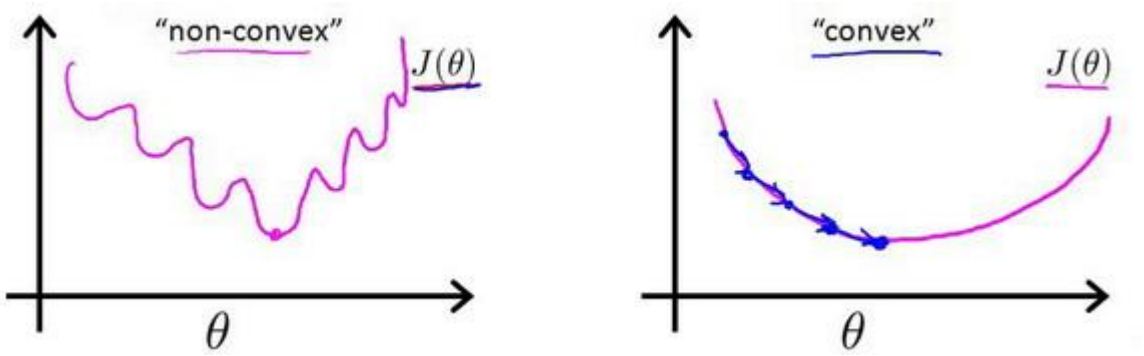


图 1.2 非凸代价函数

逻辑回归的代价函数为：

$$J(\theta)=\sum_{i=1}^{m} Cost\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) \tag{1-2}$$

Python 代码实现：

```
import numpy as np

def cost(theta, X, y):

      theta = np.matrix(theta)

       X = np.matrix(X)

       y = np.matrix(y)

       first = np.multiply(-y, np.log(sigmoid(X* theta.T)))

       second = np.multiply((1 - y), np.log(1 - sigmoid(X* theta.T)))

       return np.sum(first - second) / (len(X))
```

一些梯度下降算法之外的选择：除了梯度下降算法以外，还有一些常被用来令代价函数最小的算法，这些算法更加复杂和优越，而且通常不需要人工选择学习率，通常比梯度下降算法要更加快速。这些算法有：共轭梯度（Conjugate Gradient），局部优化法(Broyden fletcher goldfarb shann,BFGS)和有限内存局部优化法(LBFGS)。

1.2 论文阅读

对 Robust Referring Video Object Segmentation with Cyclic Structural Consensus 和 End-to-End Referring Video Object Segmentation with Multimodal Transformers 进行了通读

1.2.1 Robust Referring Video Object Segmentation with Cyclic Structural Consensus

如图 1.2 所示为 Robust Referring Video Object Segmentation with Cyclic Structural Consensus 阅读笔记。

Robust Referring Video Object Segmentation with Cyclic Structural Consensus

- **概述**
  - "semantic consensus" 语义共识
  - In this work, we highlight the need for a robust R- VOS model that can handle semantic mismatche本文目的
  - R^2-VOS
  - We tackle this problem by jointly modeling the primary R-VOS problem and its dual (text reconstruction).本文方法
  - overcoming the limitations of previous meth- ods that relied on the point-wise constraint.**本文解决的问题**
  - R2-Youtube-VOS**本文采用的数据集**
  - CSC：循环**一**致性
  - MTTR：现有的多模态方法
  - **本文贡献**
    - 解决R-VOS在输入未配对video-language 输入后引发的虚假报警的问题。
    - 引入了循环结构区分正负面视频，提高分割质量。
    - 提出了R^2-VOS 网络，引入本地模块，实现端到端的训练。
    - 在Ref- Youtube-VOS、Ref-DAVIS、new R^2-Youtube- VOS数据集中有较好的表现。
  - Robust R-VOS
    - Problem Definition
      - mask sequences {Mo}
      - unconstrained video set {V}
      - expression Eo
      - object o
    - Primary Problem
      - ◉ 最大后验估计+贝叶斯规则
    - Dual problem
      - 在给定的视频和对象mask的情况下重构文本
    - Joint optimization
    - we introduce a cross-modal proxy feature fp**与前人的区别**
    - An additional module Πcls for discriminating between positive and negative pairs.
    -
    - As mentioned, we utilize the text-to-text cycle consis- tency to measure the text-video semantic consensus.**衡量text-video一致性**
    -
    - Point-wise cycle consistency.
      - sub-optimal
- **Network**
  - feature extraction
    - We first extract the video feature f, word-level text feature g, and sentence- level text feature e.
    - fproxy
  - object localizing module (OLM)
    - fproxy
      - 粗略定位
    - p(Mo|V, Eo)
  - grounded
  - LN and FFN denote layer normalization and feed-forward network re- spectively.
  - Cq is the dimension of instance query and N is the instance query number.
- **Experiment**
  - {category, action, ap- pearance, position}
  - **具体实施**
    - 1.在COCO上进行预训练在Ref-Youtube-VOS上进行微调。

图 1.3 论文阅读笔记

1.2.2 End-to-End Referring Video Object Segmentation with Multimodal Transformers

本周对 End-to-End Referring Video Object Segmentation with Multimodal Transformers 进行阅读，将此文章和 Robust R-VOS 进行比较。MTTR 关注的是目标在

视频帧中的一个动作行为的识别，而 Robust R-VOS 对于其的改进是在正例和反例的分类和识别的准确率的改进。如图 1.4 是本周对 MTTR 阅读笔记内容。

End-to-End Referring Video Object Segmentation with Multimodal Transformers

Meta Data

Title	End-to-End Referring Video Object Segmentation with Multimodal Transformers
Journal	
Authors	Adam Botach; Evgenii Zheltonozhskii; Chaim Baskin
Pub. date	2022-04-03
期刊标签	
DOI	10.48550/arXiv.2111.14821
附件	Botach et al_2022_End-to-End Referring Video Object Segmentation with Multimodal Transformers.pdf

研究背景 & 基础 & 目的

“Referring video object segmentation.” (Botach 等, 2022, p. 1) 参考视频对象分割

“Transformers.” (Botach 等, 2022, p. 2) 介绍 Transformer

研究贡献

MTTR

“We present a Transformer-based RVOS framework, dubbed Multimodal Tracking Transformer (MTTR), which models the task as a parallel sequence prediction problem and outputs predictions for all objects in the video prior to selecting the one referred to by the text.” (Botach 等, 2022, p. 2) 我们提出了一个基于 Transformer 的 RVOS 框架，称为多模态跟踪 Transformer (MTTR)，它将任务建模为并行序列预测问题，并在选择文本所指的对象之前输出视频中所有对象的预测结果。

Temporal segment voting scheme

“Our sequence selection strategy is based on a temporal segment voting scheme, a novel reasoning scheme that allows our model to focus on more relevant parts of the video with regards to the text.” (Botach 等, 2022, p. 2) 我们的序列选择策略基于时间片段投票方案，这是一种新颖的推理方案，可让我们的模型关注视频中与文本更相关的部分。

End-to-end trainable

“The proposed method is end-to-end trainable, free of text-related inductive bias modules, and requires no additional mask refinement.” (Botach 等, 2022, p. 2) 所提出的方法是端到端的训练，不存在与文本相关的归纳偏差模块，也不需要额外的掩码改进。

研究内容

Tsak definition

1. “The input of RVOS consists of a frame sequence V = {vi}iT=1, where vi ∈ RC×H0×W0 , and a text query T = {ti}iL=1, where ti is the ith word in the text.” (Botach 等, 2022, p. 3) V 为单帧视频信息；T 为文本信息，其中 ti 为第 i 个单词。
2. “Then, for a subset of frames of interest VI ⊆ V of size TI, the goal is to segment the object referred by T in each frame in VI.” (Botach 等, 2022, p. 3) 然后，对于大小为 T_I 的感兴趣帧子集 V_I ⊆ V，目标是分割 V_I 中每个帧中 T 所指的对象。

Feature extraction

“deep spatio-temporal encoder.” (Botach 等, 2022, p. 3) 首先使用 deep spation-temproal encoder 提取 V 中的每一帧图像的特征

“Transformer-based [42] text encoder.” (Botach 等, 2022, p. 3) 同时使用 Transformer based encoder 提取文本特征

“inearly projected to a shared dimension D” (Botach 等, 2022, p. 3) 将两个提取的特征线性投影到 D

Instnce prediction

“TI” (Botach 等, 2022, p. 3) 对每个相关帧的特征进行扁平化处理，并分别与文本嵌入进行连接，生成一组 T_I 多模态序列

“exchange information” (Botach 等, 2022, p. 3) 互通信息

“Then, the decoder layers, which are fed with Nq object queries per input frame, query the multimodal sequences for entity-related information and store it in the object queries.” (Botach 等, 2022, p. 3) 解码器层在每个输入帧中输入 Nq 个对象查询，查询多模态序列中与实体相关的信息，并将其存储在对象查询中。

视频中的每一帧共享训练权重，以查询到相同的 instance sequence

Output generation

使用 FPN 和动态生成的条件卷积核生成相应的 mask

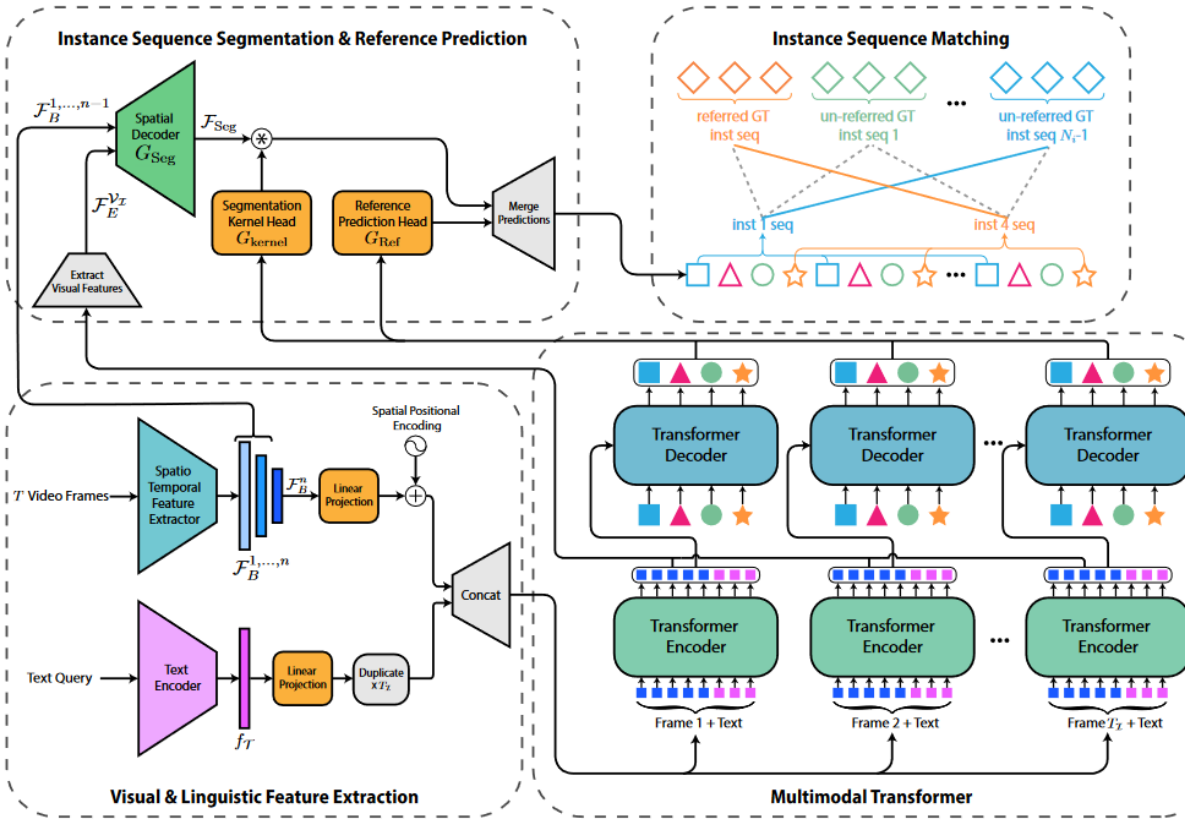
“novel text-reference score function” (Botach 等, 2022, p. 4) 新颖的文本参考评分函数用此函数来确定对象查询序列与描述对象是否具有强关联性。

Multimodal Transformer

对于每个感兴趣的帧，时间编码器生成一个特征图

文本编码器输出语言嵌入向量

理论推导



“The Instance Segmentation Process” (Botach 等, 2022, p. 4)
实例分割过程。

给定最后一个 Transformer 输出的 F_E 提取每个序列中和视频相关的部分并将其重塑为 $\mathcal{F}_E^{\mathcal{V}_T}$ 将 temporal encoder 的前 n-1 个输出为 $\mathcal{F}_B^{1,\dots,n-1}$ 通过类似于 FPN 的 spatial decoder 将 G_{Seg} 进行分层融合。产生语义丰富的高分辨率视频帧特征图 $\mathcal{F}_{Seg}$

$$\mathcal{F}_{\text{Seg}} = \left\{ f_{\text{Seg}}^t \right\}_{t=1}^{T_{\mathcal{I}}}, f_{\text{Seg}}^t \in R^{D_s \times \frac{H_0}{4} \times \frac{W_0}{4}}$$

使用双层感知器生成条件分割序列

$$G_{\text{kernel}}(\mathcal{Q}) = \{k_t\}_{t=1}^{T_\mathcal{I}}, k_t \in \mathbb{R}^I$$

将每个分割核与其对应的帧特征进行卷积，生成 mask ，双线性上采样，将 mask 调整为 grund-truth 分辨率。

“Instance Sequence Matching” (Botach 等, 2022, p. 4) 实例序列匹配

首先寻找搜索成本最低的排序

$$\hat{\sigma} = \arg\min_{\sigma \in S_{N_q}} \sum_{i=1}^{N_q} \mathcal{C}_{\text{Match}}\left(\hat{y}_{\sigma(i)}\right.$$

其中，CMatch 是成对匹配成本。使用匈牙利算法可以高效计算。每个地面实况序列的形式为

$$y_i = (m_i, r_i) = \left(\{m_i^t\}_{t=1}^{T_\mathcal{I}}, \{\right.$$

使用一个参考预测头（用$ G_{Ref}$表示），它由一个形状为 D × 2 的线性层和一个 softmax 层组成。给定预测对象查询 q∈$\mathbb{R}^D$ 后，该预测头将 q 作为输入，并输出参考预测结果 $\hat{r} \equiv G_{\mathrm{Ref}}(q)$ 。

$$\hat{y}_j = (\hat{m}_j, \hat{r}_j) = \left(\{\hat{m}_j^t\}_{t=1}^{T_\mathcal{I}}, \{\right.$$

匹配函数成本为以下函数总和

$$\mathcal{C}_{\text{Match}}(\hat{y}_j, y_i) = \mathbb{1}_{\{m_i \neq \varnothing\}} \left[\lambda_d \mathcal{C}_{\text{Dice}}(\right.$$

CRef 利用相应的地面实况序列对参考预测进行监督，具体如下

$$\mathcal{C}_{\text{Ref}}(\hat{r}_j, r_i) = -\frac{1}{T_{\mathcal{I}}} \sum_{t=1}^{T_{\mathcal{I}}} \hat{r}_j^t \cdot r_i^t$$

“Loss Functions” (Botach 等, 2022, p. 5) 损失函数

$$\mathcal{L}(\hat{y}, y) = \sum_{i=1}^{N_q} \mathbb{1}_{\{m_i \neq \varnothing\}} \mathcal{L}_{\text{Mask}}($$

L_mask 被定义为 Dice 和每个像素 Focal 损失函数的组合

$$\mathcal{L}_{\text{Mask}}(\hat{m}_i, m_i) = \lambda_d \mathcal{L}_{\text{Dice}}(\hat{m}_i, m_i$$

L_Dice 和 L_Focal 在每个时间步长都会应用于相应的掩码，并根据训练批次中的实例数量进行归一化处理。

L_Ref 为交叉熵用于监督序列参考预测

$$\mathcal{L}_{\text{Ref}}(\hat{r}_i, r_i) = -\lambda_r \frac{1}{T_{\mathcal{I}}} \sum_{t=1}^{T_{\mathcal{I}}} r_i^t \,.$$

“Inference” (Botach 等, 2022, p. 5)

输出 R

给定参考预测值的 positive 类别概率

返回分段掩码序列和其得分

$$\hat{r}_{\text{pred}} = \arg\max_{\hat{r}_i \in \mathcal{R}} \sum_{t=1}^{T_{\mathcal{I}}} p_{\text{ref}}(\hat{r}_i^t).$$

将这种序列选择方案称为 "时间片段投票方案"（TSVS），它根据每个预测序列的术语与文本所指对象的总关联度对其进行分级。

实验结果

数据集

“A2D-Sentences and JHMDB-Sentences” (Botach 等, 2022, p. 5)在数据集上添加文本注释

“ReferYouTube-VOS dataset” (Botach 等, 2022, p. 6)每段视频每五帧都有像素级实例分割注释。

精度估计方法

“We adopt Overall IoU, Mean IoU, and precision@K to evaluate our method on these datasets.” (Botach 等, 2022, p. 6)我们在这些数据集上采用总体 IoU、平均 IoU 和精度@K 来评估我们的方法。

“Overall IoU computes the ratio between the total intersection and the total union area over all the test samples.” (Botach 等, 2022, p. 6)总体 IoU 计算的是所有测试样本的总交

又面积与总结合面积之间的比率。

“Mean IoU is the averaged IoU over all the test samples.” (Botach 等, 2022, p. 6)

“Precision@K considers the percentage of test samples whose IoU scores are above a threshold K, where K ∈ [0.5, 0.6, 0.7, 0.8, 0.9].” (Botach 等, 2022, p. 6)

“The primary evaluation metrics for this dataset are the average of the region similarity (J) and the contour accuracy (F) [35].” (Botach 等, 2022, p. 6)该数据集的主要评估指标是区域相似度（J）和轮廓精度（F）的平均值 [35]。

具体实施

预训练

我们使用最小的（"微小"）视频 Swin 变换器 [28] 作为时态编码器，并在 Kinetics-400 [17] 上进行了预训练。只用 swin transformer 的前三个区块 第三个区块作为多模态 transformer

参数

w=8/w=12

320×576/360×640

比较

Method	Precision					IoU		mAP
	50%	60%	70%	80%	90%	Overall	Mean	
Hu et al. [13]	34.8	23.6	13.3	3.3	0.1	47.4	35.0	13.2
Gavrilyuk et al. [12] (RGB)	47.5	34.7	21.1	8.0	0.2	53.6	42.1	19.8
RefVOS [1]	57.8	–	–	–	9.3	67.2	49.7	–
AAMN [51]	68.1	62.9	52.3	29.6	2.9	61.7	55.2	39.6
CMSA+CFSA [53]	48.7	43.1	35.8	23.1	5.2	61.8	43.2	–
CSTM [14]	65.4	58.9	49.7	33.3	9.1	66.2	56.1	39.9
CMPC-V (I3D) [25]	65.5	59.2	50.6	34.2	9.8	65.3	57.3	40.4
MTTR ($w = 8$, ours)	72.1	68.4	60.7	45.6	16.4	70.2	61.8	44.7
MTTR ($w = 10$, ours)	**75.4**	**71.2**	**63.8**	**48.5**	**16.9**	**72.0**	**64.0**	**46.1**

Table 1. Comparison with state-of-the-art methods on A2D-Sentences [12].

Method	Precision					IoU		mAP
	50%	60%	70%	80%	90%	Overall	Mean	
Hu et al. [13]	63.3	35.0	8.5	0.2	0.0	54.6	52.8	17.8
Gavrilyuk et al. [12] (RGB)	69.9	46.0	17.3	1.4	0.0	54.1	54.2	23.3
AAMN [51]	77.3	62.7	36.0	4.4	0.0	58.3	57.6	32.1
CMSA+CFSA [53]	76.4	62.5	38.9	9.0	**0.1**	62.8	58.1	–
CSTM [14]	78.3	63.9	37.8	7.6	0.0	59.8	60.4	33.5
CMPC-V (I3D) [25]	81.3	65.7	37.1	7.0	0.0	61.6	61.7	34.2
MTTR ($w = 8$, ours)	91.0	81.5	57.0	14.4	**0.1**	67.4	67.9	36.6
MTTR ($w = 10$, ours)	**93.9**	**85.2**	**61.6**	**16.6**	**0.1**	**70.1**	**69.8**	**39.2**

Table 2. Comparison with state-of-the-art methods on JHMDB-Sentences [12].

图 1.4 论文阅读笔记

1.3 工作内容

主要从事爬取河北省领导留言板和政法信息的内容

第二章 第四周学习计划及上周总结

序号	上周任务	完成情况及备注	本周任务
1	复习机器学习 week3、week4、week5	完成了 week3	复习 week4
2	Robust Referring Video Object Segmentation with Cyclic Structural Consensus 的阅读和整理	完成	阅读 Position-guided Text Prompt for Vision-Language Pre-training
3	跑通 R2VOS 代码	可以运行 demo	阅读复现 R2VOS
4	跑通 centerpoint 的代码	更改为 MTTR	
5	阅读 Exploring Simple 3D Multi-Object Tracking for Autonomous Driving	更改为阅读 MTTR，MTTR 阅读整理已完成	

参考

1. https://github.com/fengdu78/Coursera-ML-AndrewNg-Notes.git
2. https://github.com/lxa9867/R2VOS.git

3. https://github.com/mttr2021/MTTR.git

4.