AutoSeM

Automatic Task Selection and Mixing in Multi-Task Learning

Han Guo, Ramakanth Pasunuru, Mohit Bansal

UNC NLP | UNC DEPARTMENT OF COMPUTER SCIENCE

Overview

1. Introduction
2. Methods
3. Experiments

Overview

1.Introduction

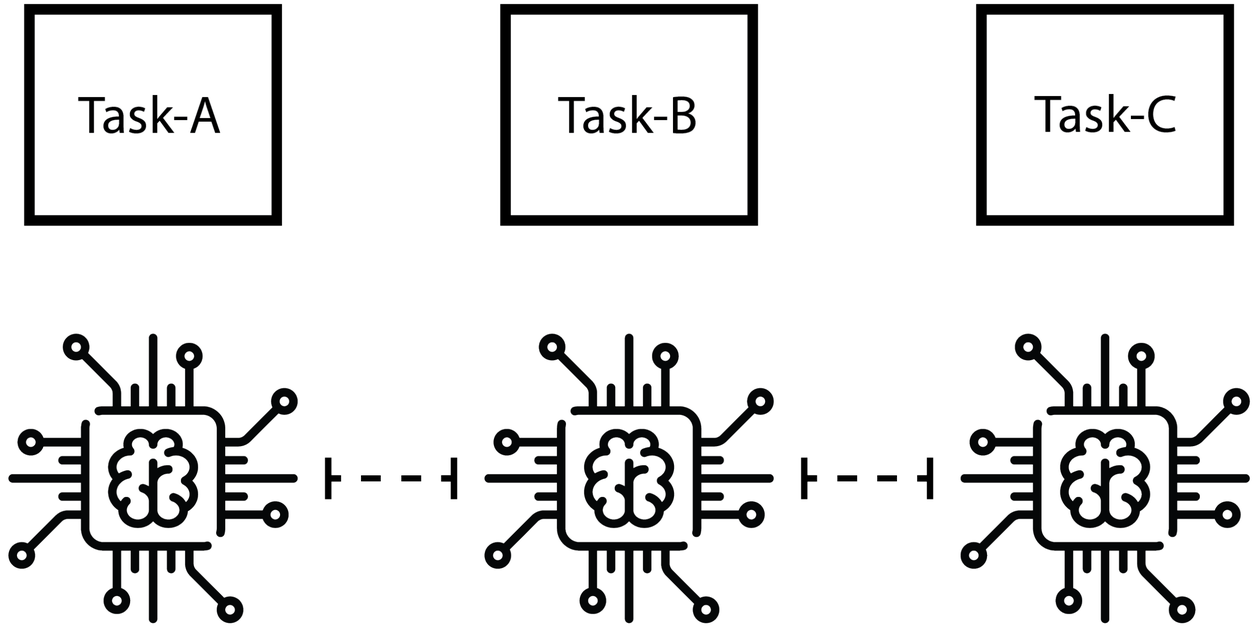
2.Methods

3.Experiments

Introduction

- Multi-task Learning (MTL) is an **inductive transfer mechanism** which **leverages information from related tasks** to improve the primary model's **generalization performance**.

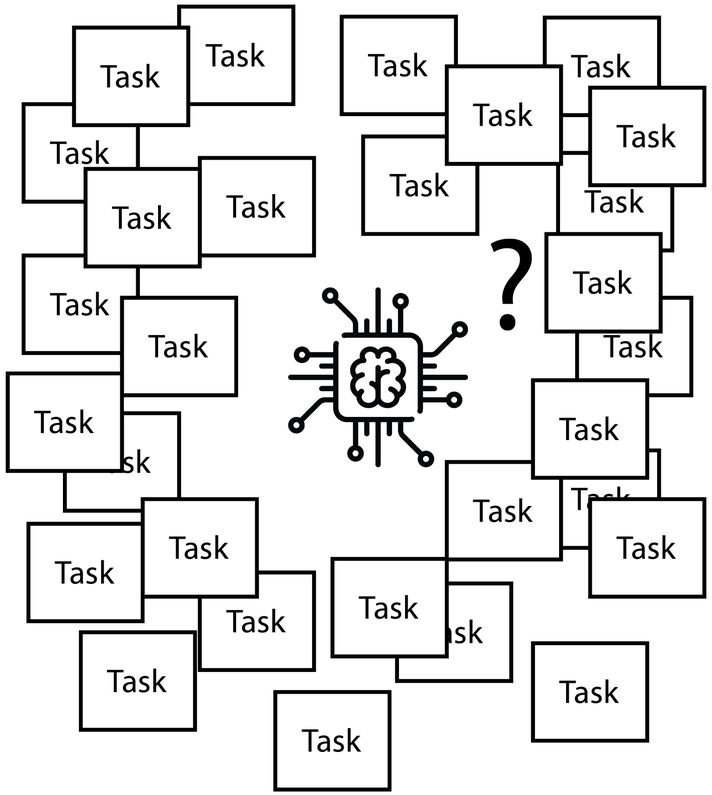
- It achieves this goal by training **multiple** tasks in **parallel** while **sharing representations**, where the training signals from the auxiliary tasks can help improve the performance of the primary task.



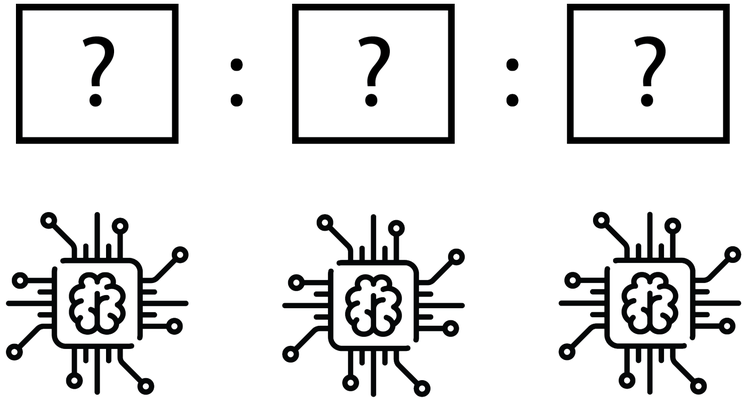
Caruana, 1997

Two Issues

$\mathcal{O}(2^k)$



+



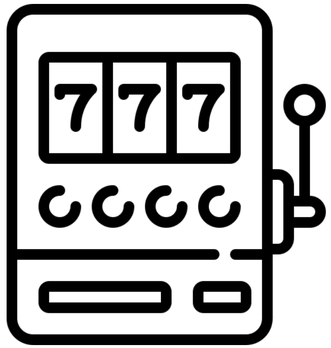
$\mathcal{O}(n^k)$

Auxiliary Task Selection

Mixing Ratio Learning

AutoSeM

Multi-Armed Bandit



+

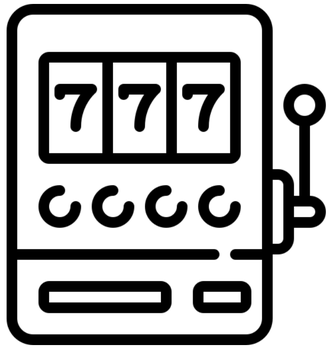
Gaussian Process

Auxiliary Task Selection

Mixing Ratio Learning

AutoSeM

Multi-Armed Bandit



+

Gaussian Process

Auxiliary Task Selection

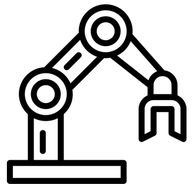
Mixing Ratio Learning

Related Works

Multi-Task Learning

Collobert and Weston, 2008; Girshick, 2015; Luong et al., 2015; Pasunuru and Bansal, 2017; Pasunuru et al., 2017

Misra et al.,2016; Kendall et al., 2017; Dai et al., 2016



Teh et al., 2017; Parisotto et al., 2015; Jaderberg et al., 2016

Shared Parameter Selection

Ruder et al. (2017); Xiao et al. (2018)

Identifying Task Relatedness

Ben-David and Schuller (2003), Bingel and Søgaard (2017)

Data Selection/Reordering

van der Wees et al., 2017; Moore and Lewis, 2010; Duh et al., 2013; Søgaard, 2011; Ruder and Plank (2017); Tsvetkov et al. (2016)

Multi-Armed Bandit

Graves et al. (2017); Sharma and Ravindran(2017)

Exploitation/Exploration

Kaelbling et al., 1996; Auer et al., 2002b; Auer et al., 2002a; Houthooft et al., 2016; Russo et al., 2018; Chapelle and Li, 2011

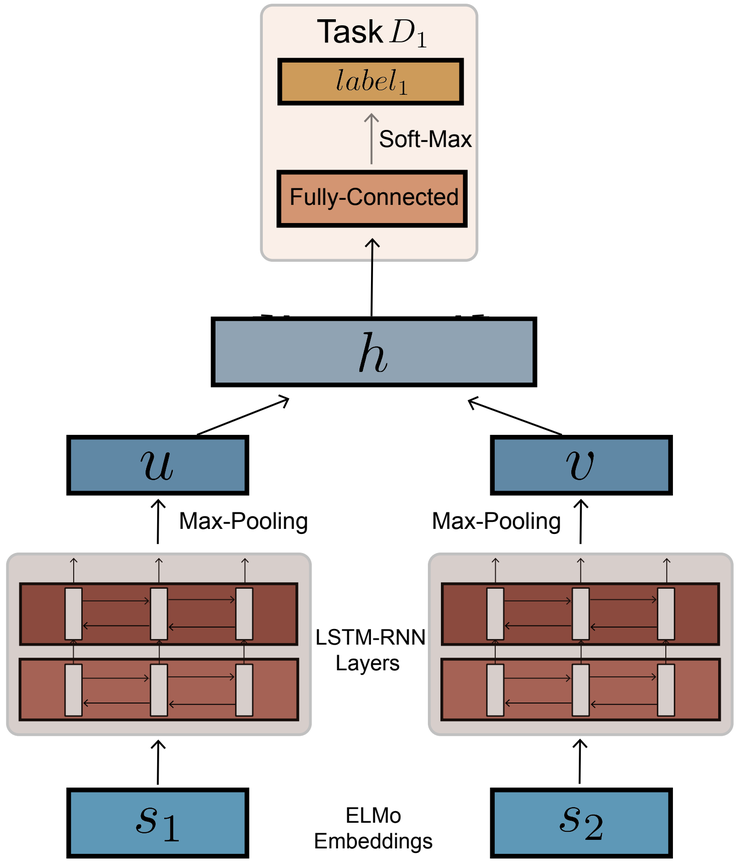
Gaussian Process

Rasmussen, 2004; Shahriari et al., 2016; Schulz et al.,2018; Snoek et al., 2012; Brochu et al., 2010; Knudde et al., 2017; Cully et al., 2018; Swersky et al., 2013; Golovin et al., 2017)

Overview

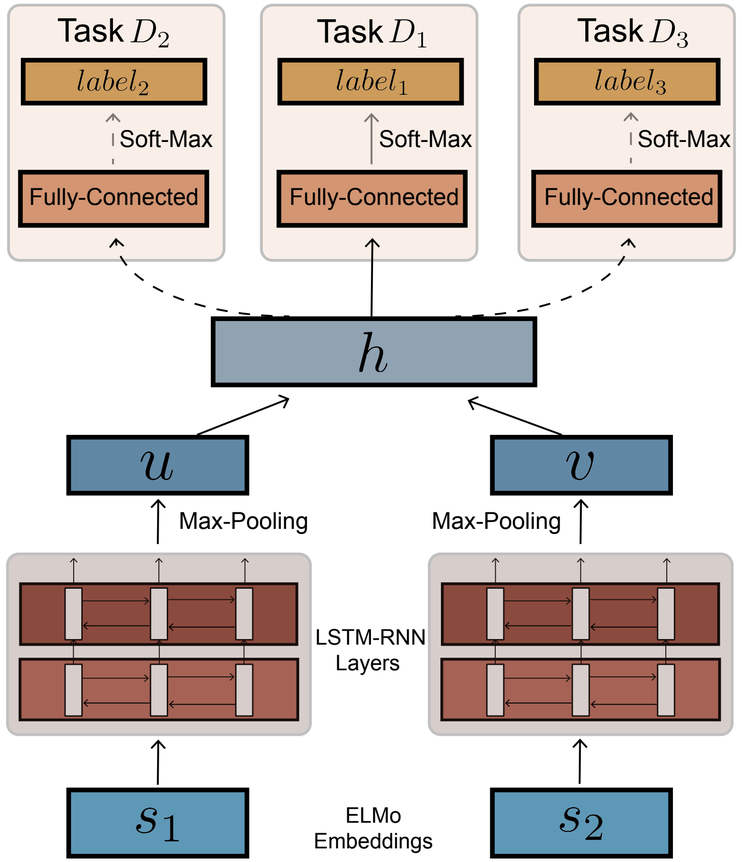
1. Introduction
2. **Methods**
3. Experiments

Base Model



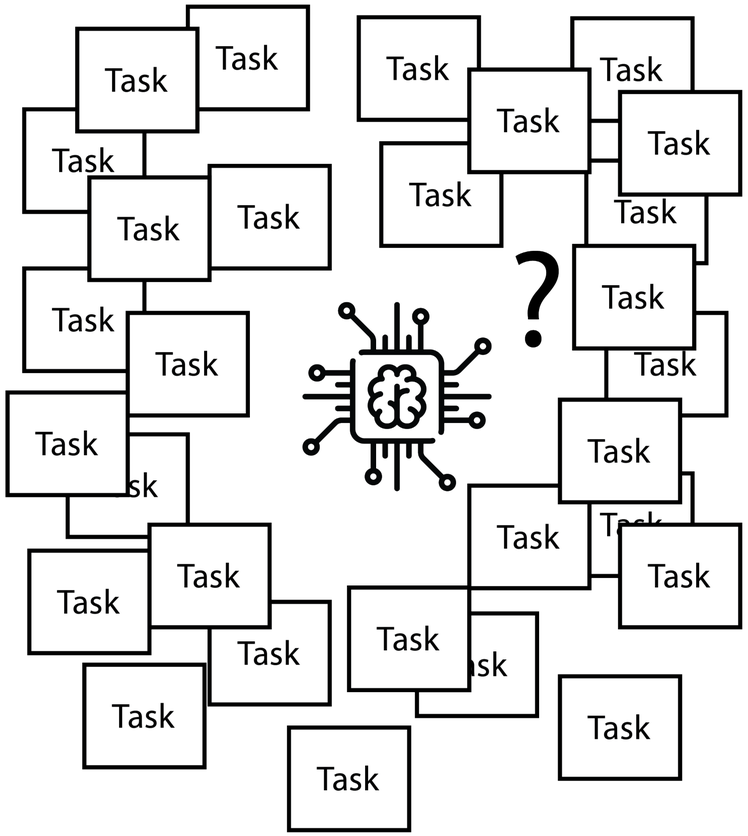
Peters et al., (2018); Conneau et al., (2017)

Multi-Task Model

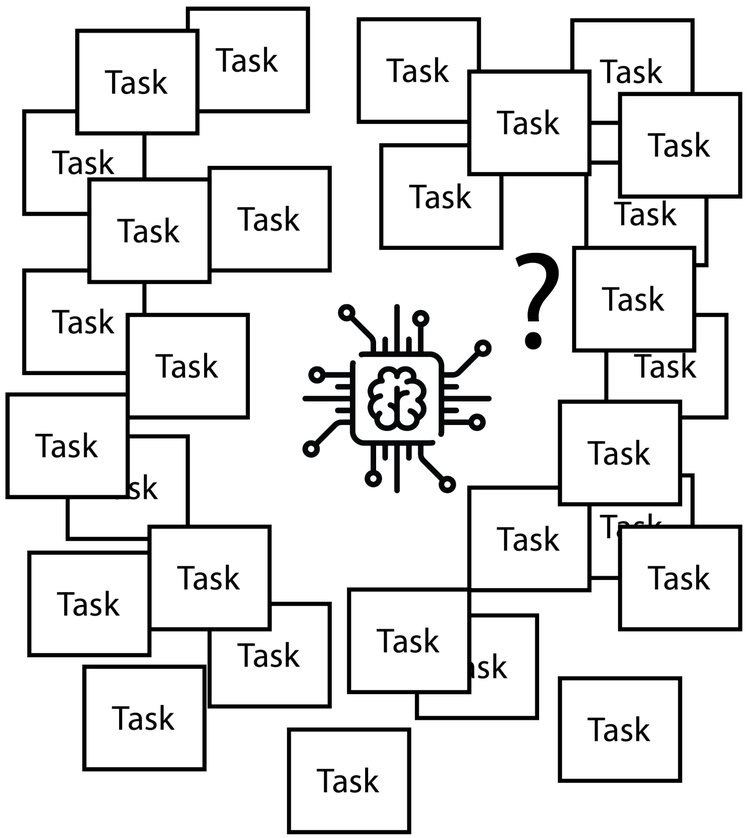


Peters et al., (2018); Conneau et al., (2017)

Auxiliary Task Selection



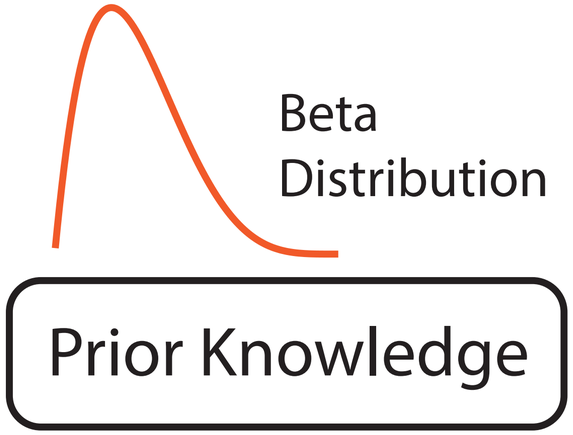
Auxiliary Task Selection



Multi-Armed Bandit

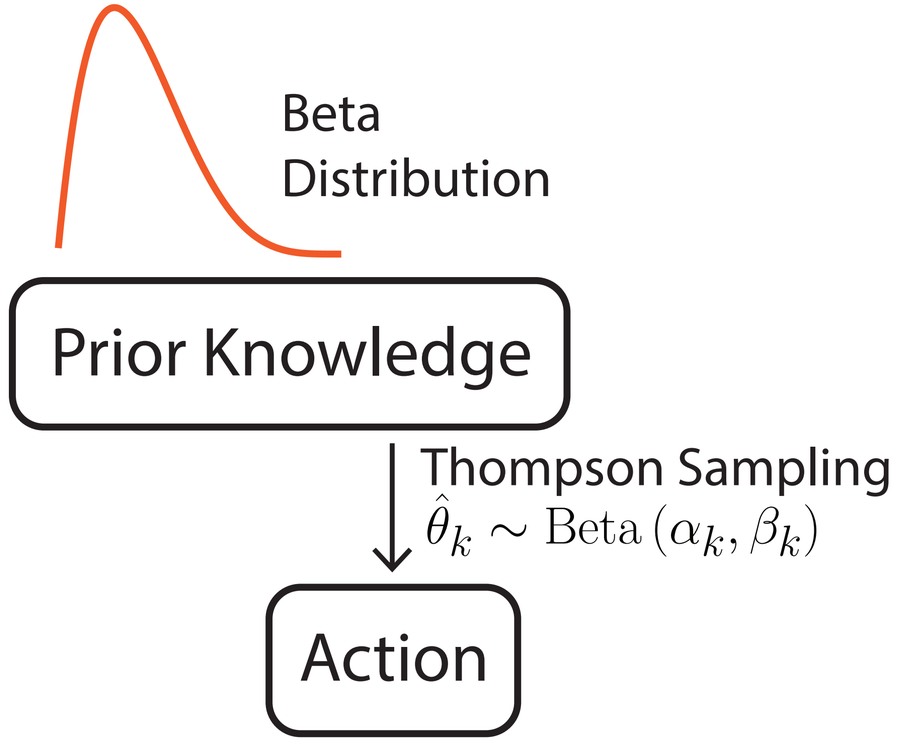
Image Source: https://www.spotx.tv/resources/blog/developer-blog/introduction-to-multi-armed-bandits-with-applications-in-digital-advertising/

Auxiliary Task Selection



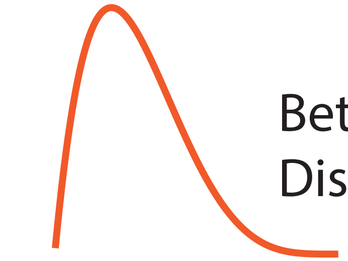
Russo et al., 2018; Chapelle and Li, 2011

Auxiliary Task Selection



Russo et al., 2018; Chapelle and Li, 2011

Auxiliary Task Selection



Beta Distribution

Prior Knowledge

↓ Thompson Sampling

$\hat{\theta}_k \sim \text{Beta}(\alpha_k, \beta_k)$

Action

↓ Train

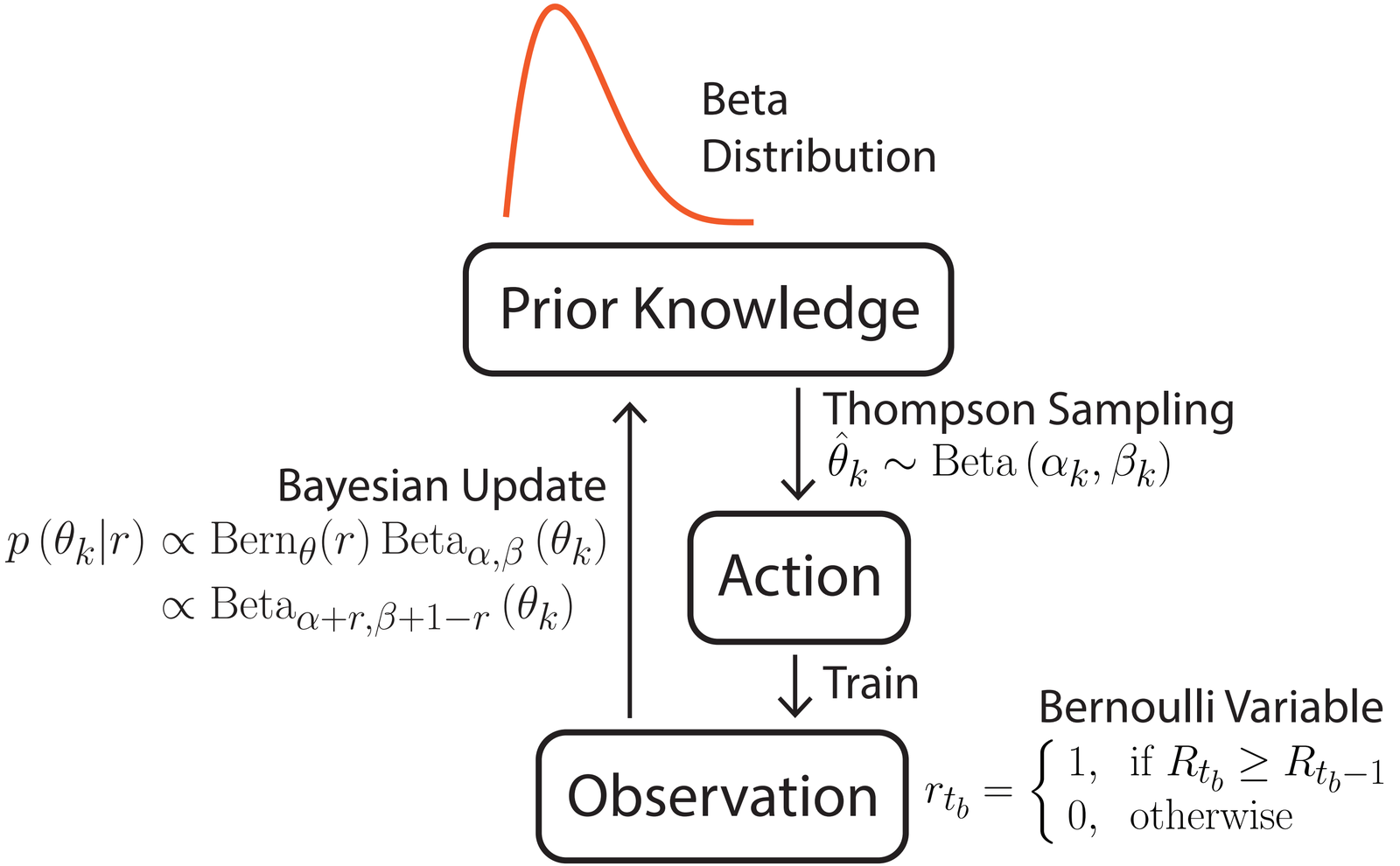
Observation

Bernoulli Variable

$$r_{t_b} = \begin{cases} 1, & \text{if } R_{t_b} \geq R_{t_b - 1} \\ 0, & \text{otherwise} \end{cases}$$

Russo et al., 2018; Chapelle and Li, 2011

Auxiliary Task Selection

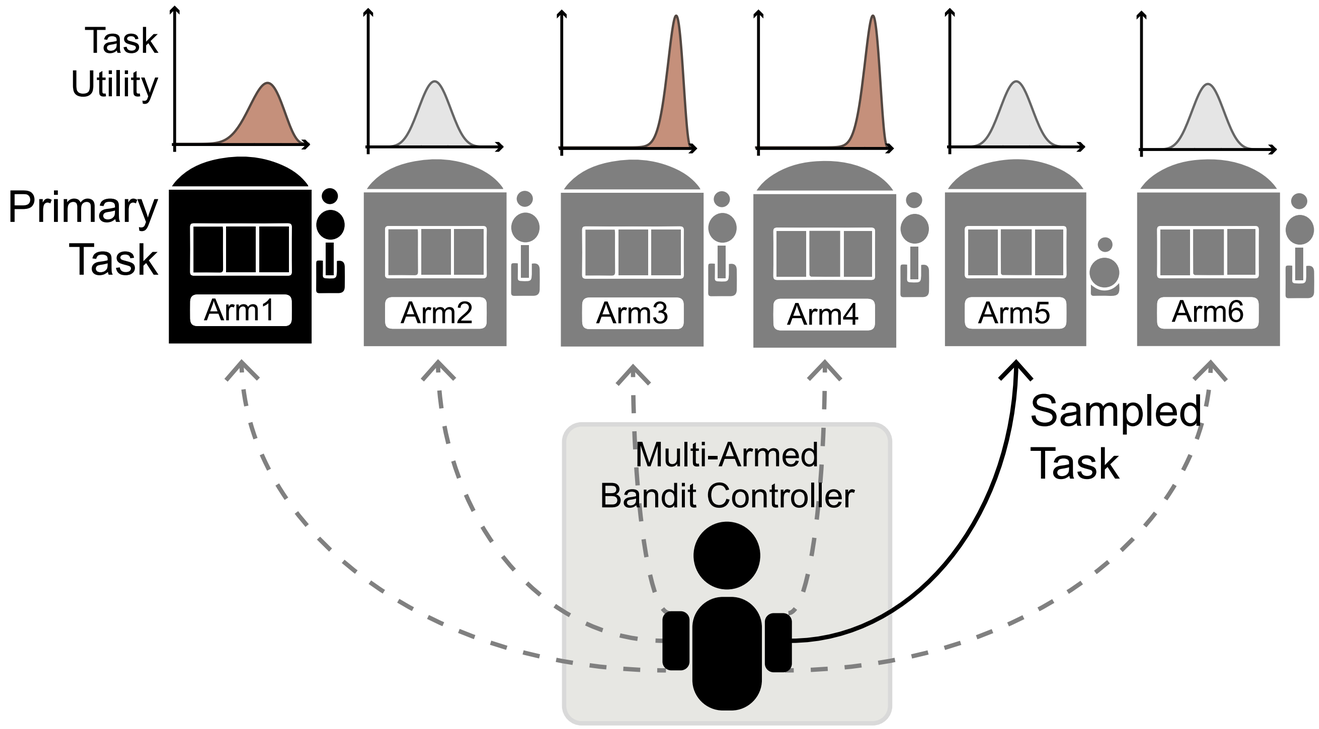


Russo et al., 2018; Chapelle and Li, 2011

Auxiliary Task Selection

Algorithm 1 BernThompson$(N, \alpha, \beta, \gamma, \alpha_0, \beta_0)$

for $t_b = 1, 2, \ldots$ **do**
  # sample model:
  for $k = 1, \ldots, N$ **do**
    Sample $\hat{\theta}_k \sim \text{Beta}(\alpha_k, \beta_k)$
  end for
  # select and apply action:
  $x^s_{t_b} \leftarrow \arg\max_k \hat{\theta}_k$
  Apply $x^s_{t_b}$ and observe r_{t_b}
  # non-stationarity
  for $k = 1, \ldots, N$ **do**
    $\hat{\alpha}_k = (1 - \gamma)\alpha_k + \gamma\alpha_0$
    $\hat{\beta}_k = (1 - \gamma)\beta_k + \gamma\beta_0$
    if $k \neq x^s_{t_b}$ **then**
      $(\alpha_k, \beta_k) \leftarrow (\hat{\alpha}_k, \hat{\beta}_k)$
    else
      $(\alpha_k, \beta_k) \leftarrow (\hat{\alpha}_k, \hat{\beta}_k) + (r_{t_b}, 1 - r_{t_b})$
    end if
  end for
end for

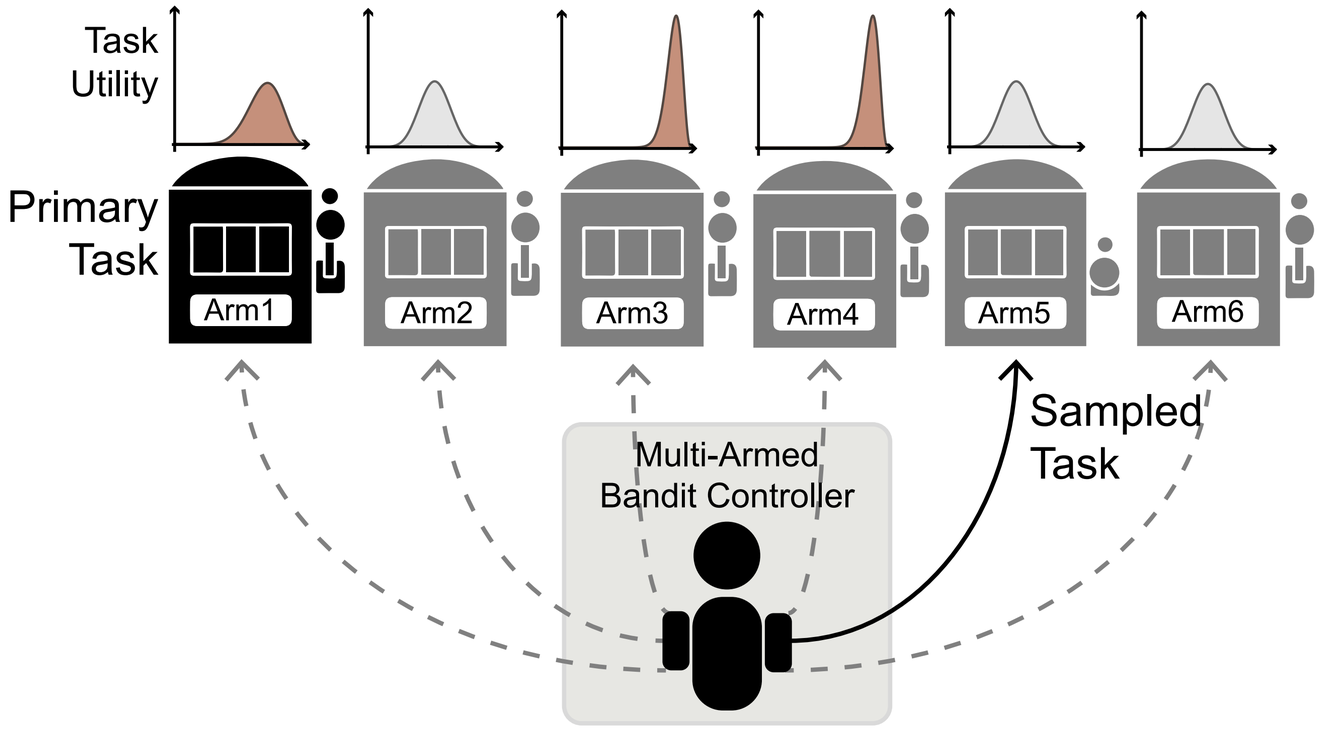


Russo et al., 2018; Chapelle and Li, 2011

Auxiliary Task Selection

Algorithm 1 BernThompson$(N, \alpha, \beta, \gamma, \alpha_0, \beta_0)$

for $t_b = 1, 2, \ldots$ **do**
  # sample model:
  for $k = 1, \ldots, N$ **do**
    Sample $\hat{\theta}_k \sim \text{Beta}(\alpha_k, \beta_k)$
  end for
  # select and apply action:
  $x^s_{t_b} \leftarrow \arg\max_k \hat{\theta}_k$
  Apply $x^s_{t_b}$ and observe r_{t_b}
  # non-stationarity
  for $k = 1, \ldots, N$ **do**
    $\hat{\alpha}_k = (1 - \gamma)\alpha_k + \gamma\alpha_0$
    $\hat{\beta}_k = (1 - \gamma)\beta_k + \gamma\beta_0$
    if $k \neq x^s_{t_b}$ **then**
      $(\alpha_k, \beta_k) \leftarrow (\hat{\alpha}_k, \hat{\beta}_k)$
    else
      $(\alpha_k, \beta_k) \leftarrow (\hat{\alpha}_k, \hat{\beta}_k) + (r_{t_b}, 1 - r_{t_b})$
    end if
  end for
end for



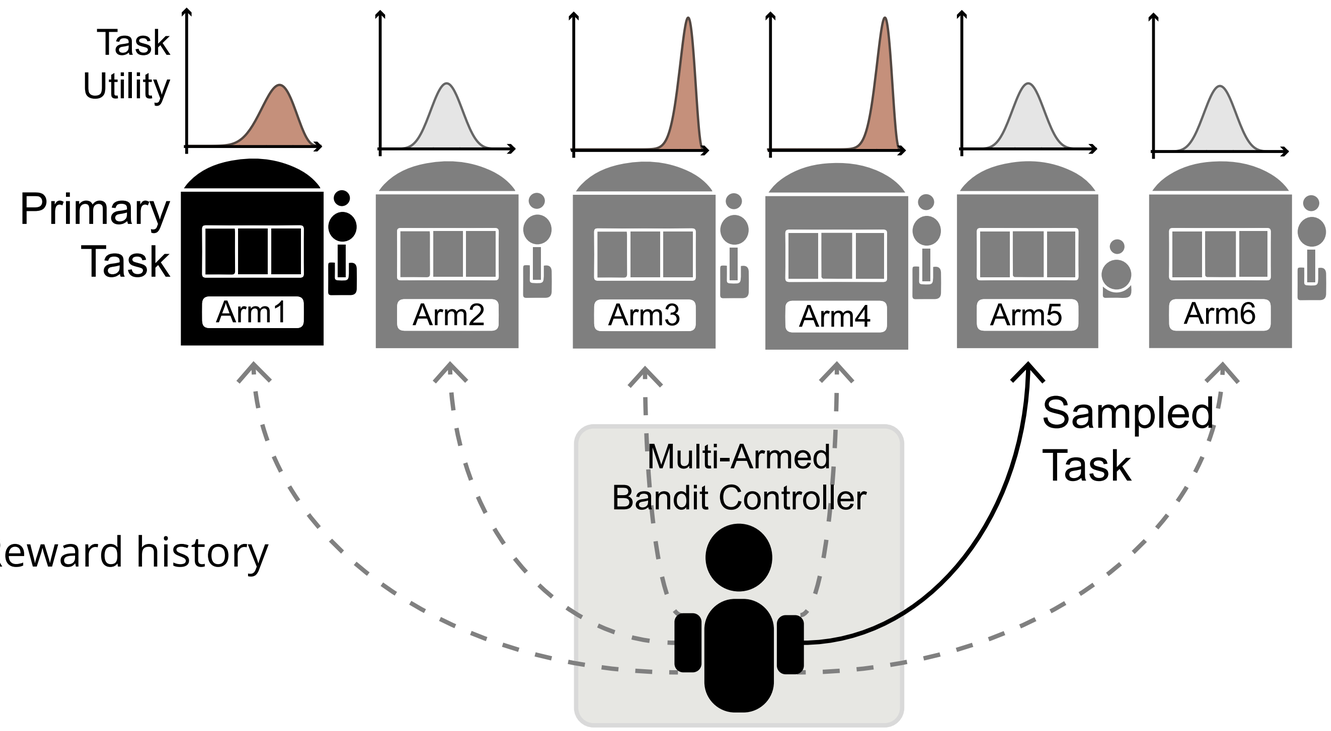
Russo et al., 2018; Chapelle and Li, 2011

Auxiliary Task Selection

Algorithm 1 BernThompson$(N, \alpha, \beta, \gamma, \alpha_0, \beta_0)$

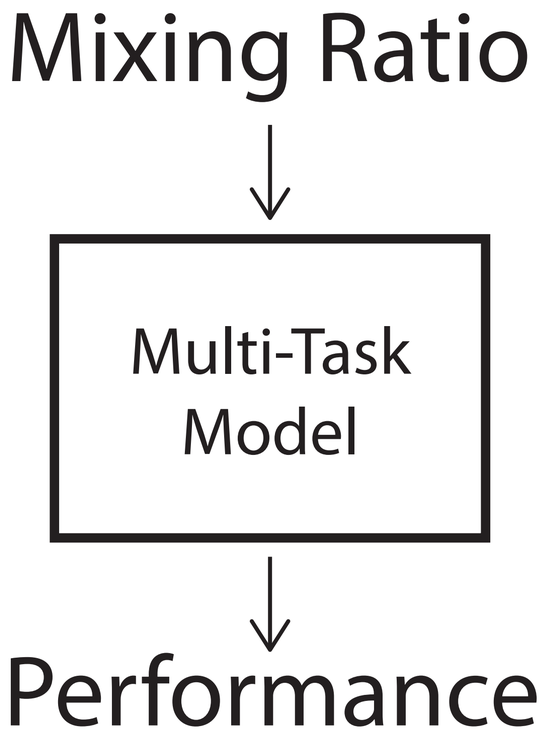
for $t_b = 1, 2, \ldots$ **do**
  # sample model:
  for $k = 1, \ldots, N$ **do**
    Sample $\hat{\theta}_k \sim \text{Beta}(\alpha_k, \beta_k)$
  end for
  # select and apply action:
  $x^s_{t_b} \leftarrow \arg\max_k \hat{\theta}_k$
  Apply $x^s_{t_b}$ and observe r_{t_b}
  # non-stationarity
  for $k = 1, \ldots, N$ **do**
    $\hat{\alpha}_k = (1-\gamma)\alpha_k + \gamma\alpha_0$
    $\hat{\beta}_k = (1-\gamma)\beta_k + \gamma\beta_0$
    if $k \neq x^s_{t_b}$ **then**
      $(\alpha_k, \beta_k) \leftarrow (\hat{\alpha}_k, \hat{\beta}_k)$
    else
      $(\alpha_k, \beta_k) \leftarrow (\hat{\alpha}_k, \hat{\beta}_k) + (r_{t_b}, 1 - r_{t_b})$
    end if
  end for
end for

Reward history



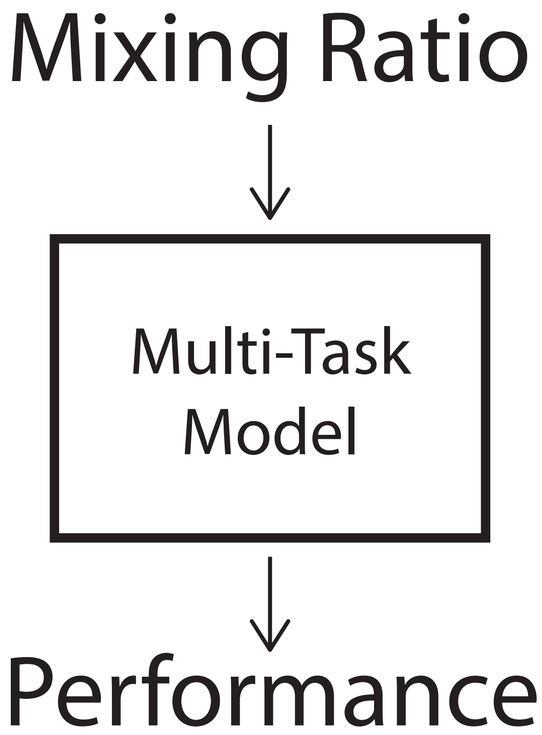
Russo et al., 2018; Chapelle and Li, 2011

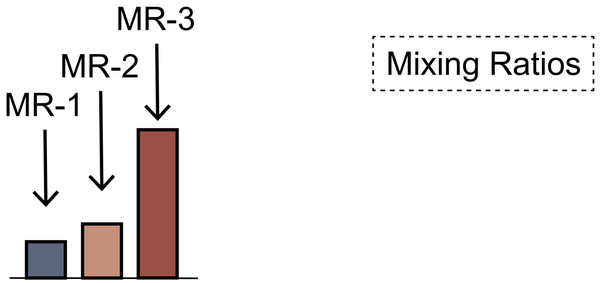
Mixing Ratio Learning



Rasmussen, 2004; Snoek et al., 2012; Shahriari et al., 2016

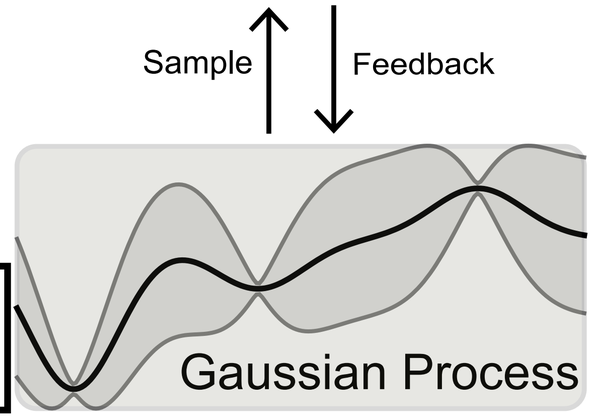
Mixing Ratio Learning





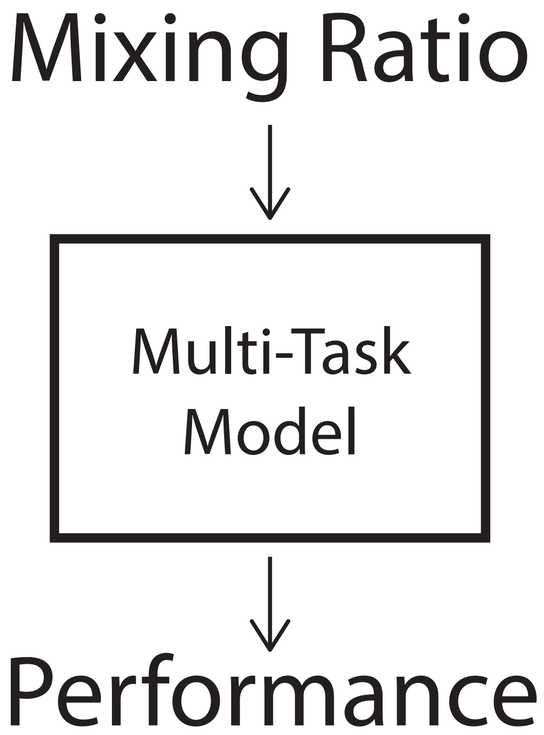
$$\boldsymbol{f}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K})$$
$$\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 \boldsymbol{I})$$

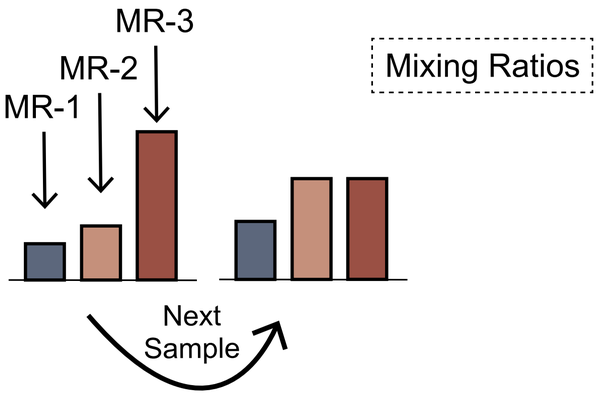
Kernel: Matern Kernel
Acquisition Function: Hedge

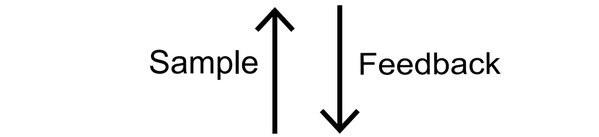


Rasmussen, 2004; Snoek et al., 2012; Shahriari et al., 2016

Mixing Ratio Learning





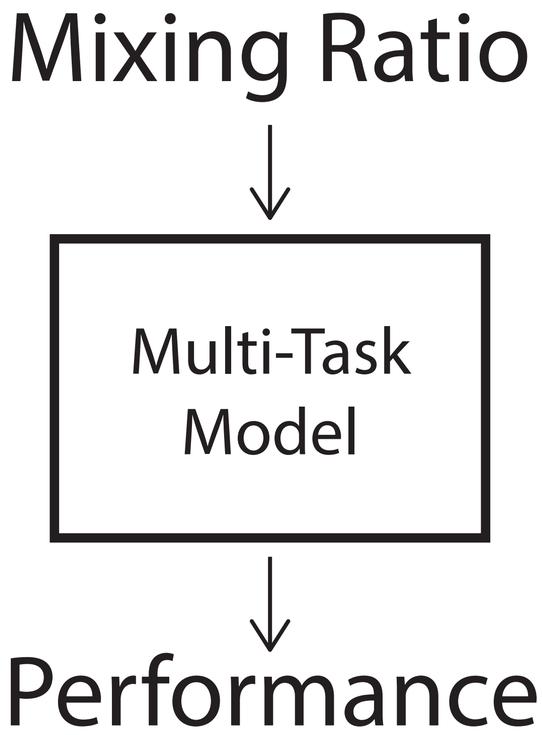


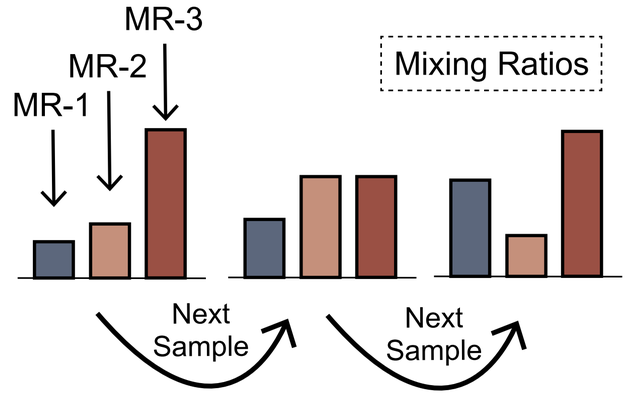
$$\boldsymbol{f}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K})$$
$$\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 \boldsymbol{I})$$

Kernel: Matern Kernel
Acquisition Function: Hedge

Rasmussen, 2004; Snoek et al., 2012; Shahriari et al., 2016

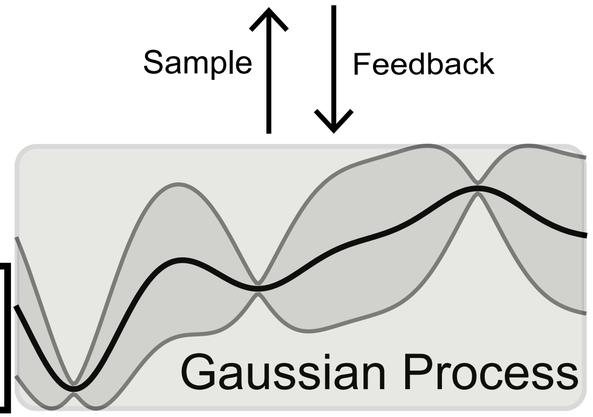
Mixing Ratio Learning





$$\boldsymbol{f}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K})$$
$$\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 \boldsymbol{I})$$

Kernel: Matern Kernel
Acquisition Function: Hedge



Rasmussen, 2004; Snoek et al., 2012; Shahriari et al., 2016

Overview

1. Introduction
2. Methods
3. **Experiments**

UNC NLP

Dataset

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Wang et al., 2018; Warstadt et al., 2018; Socher et al., 2013; Dolan & Brockett, 2005; Cer et al., 2017; Williams et al., 2018; Bowman et al., 2015; Rajpurkar et al. 2016; Dagan et al., 2006; Bar Haim etal.,2006; Giampiccoloetal.,2007; Bentivoglietal.,2009; Levesque et al., 2011

Dataset Example

Premise	Label	Hypothesis
Fiction		
The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.
Letters		
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.
Telephone Speech		
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.
9/11 Report		
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.

Experiments

Models	RTE	MRPC	QNLI	CoLA	SST-2
BiLSTM+ELMo (Single-Task) (Wang et al., 2018)	50.1	69.0/80.8	69.4	**35.0**	90.2
BiLSTM+ELMo (Multi-Task) (Wang et al., 2018)	55.7	76.2/83.5	66.7	27.5	89.6
Our Baseline	54.0	75.7/83.7	74.0	30.8	91.3
Our AUTOSEM	**58.7**	**78.5/84.5**	**79.2**	32.9	**91.8**

UNC NLP

Experiments

Models	RTE	MRPC	QNLI	CoLA	SST-2
BiLSTM+ELMo (Single-Task) (Wang et al., 2018)	50.1	69.0/80.8	69.4	**35.0**	90.2
BiLSTM+ELMo (Multi-Task) (Wang et al., 2018)	55.7	76.2/83.5	66.7	27.5	89.6
Our Baseline	54.0	75.7/83.7	74.0	30.8	91.3
Our AUTOSEM	**58.7**	**78.5/84.5**	**79.2**	32.9	**91.8**

Selected Auxiliary Tasks (Stage-1)

RTE: MRPC, **QQP**, **MultiNLI**, QNLI, WNLI
MRPC: QQP, **MultiNLI**, QNLI, **RTE**, WNLI
QNLI: MRPC, QQP, **MultiNLI**, RTE, **WNLI**
CoLA: MRPC, QQP, **MultiNLI**, QNLI, RTE, **WNLI**, SST-2
SST-2: **MRPC**, QQP, **MultiNLI**, QNLI, RTE, **WNLI**, CoLA

Experiments

Models	RTE	MRPC	QNLI	CoLA	SST-2
BiLSTM+ELMo (Single-Task) (Wang et al., 2018)	50.1	69.0/80.8	69.4	**35.0**	90.2
BiLSTM+ELMo (Multi-Task) (Wang et al., 2018)	55.7	76.2/83.5	66.7	27.5	89.6
Our Baseline	54.0	75.7/83.7	74.0	30.8	91.3
Our AUTOSEM	**58.7**	**78.5/84.5**	**79.2**	32.9	**91.8**

Selected Auxiliary Tasks (Stage-1)

RTE: MRPC, QQP, MultiNLI, QNLI, WNLI
MRPC: QQP, MultiNLI, QNLI, RTE, WNLI
QNLI: MRPC, QQP, MultiNLI, RTE, WNLI
CoLA: MRPC, QQP, MultiNLI, QNLI, RTE, WNLI, SST-2
SST-2: MRPC, QQP, MultiNLI, QNLI, RTE, WNLI, CoLA

MultiNLI is always chosen in Stage-1

Experiments

Models	RTE	MRPC	QNLI	CoLA	SST-2
BiLSTM+ELMo (Single-Task) (Wang et al., 2018)	50.1	69.0/80.8	69.4	**35.0**	90.2
BiLSTM+ELMo (Multi-Task) (Wang et al., 2018)	55.7	76.2/83.5	66.7	27.5	89.6
Our Baseline	54.0	75.7/83.7	74.0	30.8	91.3
Our AUTOSEM	**58.7**	**78.5/84.5**	**79.2**	32.9	**91.8**

Learned Mixing Ratios (Stage-2)

RTE: QQP, MultiNLI = 1:5:5
MRPC: RTE, MultiNLI = 9:1:4
QNLI: WNLI, MultiNLI = 20:0:5
CoLA: MultiNLI, WNLI = 20:5:0
SST-2: MultiNLI, MRPC, WNLI = 13:5:0:0

Removing Stage-1

Name	Validation	Test
Baseline	78.3	75.7/83.7
w/o Stage-1	80.3	76.3/83.8
w/o Stage-2	80.3	76.7/83.8
Final MTL	81.2	78.5/84.5

w/o Stage-1: Applying Gaussian process on all candidate auxiliary tasks.

(MRPC Dataset)

Removing Stage-2

Name	Validation	Test
Baseline	78.3	75.7/83.7
w/o Stage-1	80.3	76.3/83.8
w/o Stage-2	80.3	76.7/83.8
Final MTL	81.2	78.5/84.5

w/o Stage-2: Apply manual tuning of mixing ratios on selected auxiliary tasks.

(MRPC Dataset)

Stability of MTL Models

Name	RTE	MRPC	QNLI	CoLA	SST-2
BASELINES					
Mean	58.6	78.3	74.9	74.6	91.4
Std	0.94	0.31	0.30	0.44	0.36
MULTI-TASK MODELS					
Mean	62.0	81.1	76.0	75.7	91.8
Std	0.62	0.20	0.18	0.18	0.29

Educated-Guess Baselines

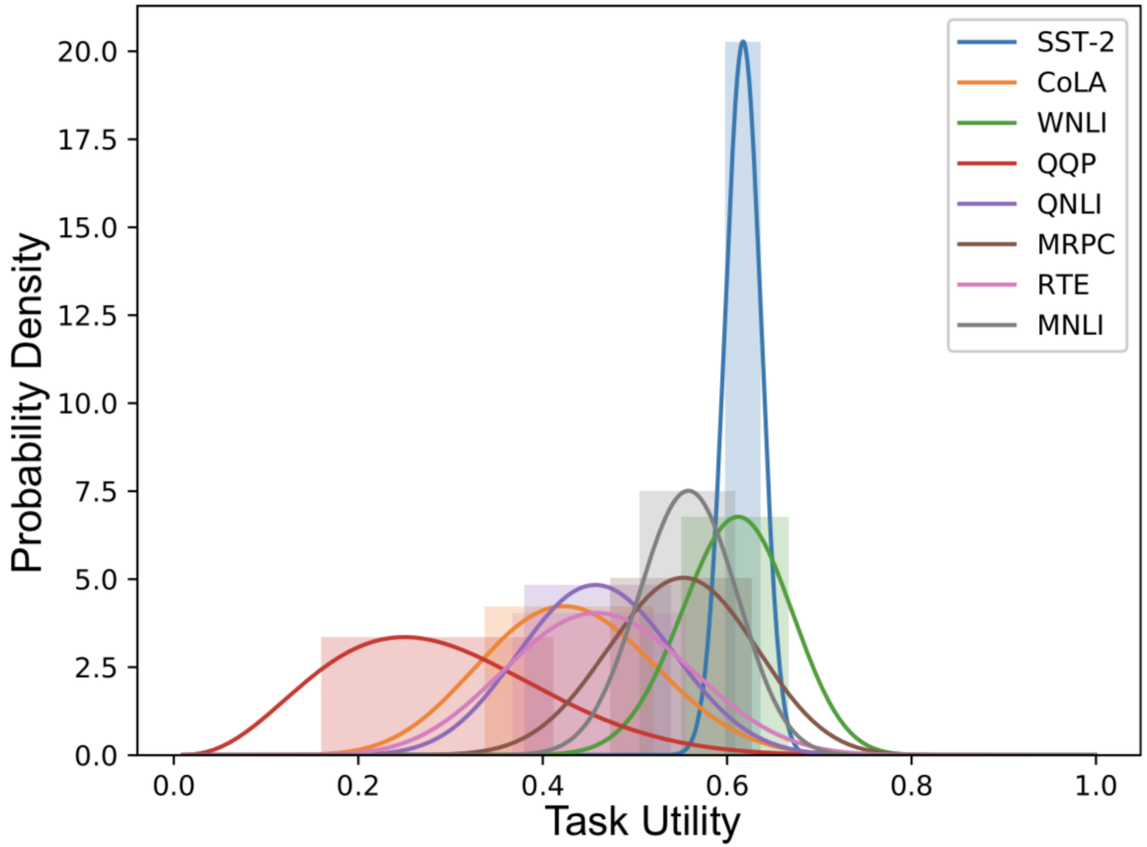
Our first educated-guess baseline is to choose other similar paraphrasing-based auxiliary tasks.

- MRPC+QQP

Our second educated-guess baseline added MultiNLI as an auxiliary task (in addition to QQP)

- MRPC+QQP+MultiNLI

Visualization of Stage-1



Visualization of task utility estimates from the multi-armed bandit controller on SST-2 (primary task). The x-axis represents the task utility, and the y- axis represents the corresponding probability density. Each curve corresponds to a task and the bar corresponds to their confidence interval.

Thanks!


Code: github.com/hanguo97/AutoSeM

Acknowledgement: DARPA-YFA, ONR, Google, Facebook, Baidu, Salesforce, and Nvidia.