深层神经网络多任务学习综述（中文翻译）

Sebastian Ruder†
Insight Centre for Data Analytics, NUI Galway†
孙天祥 ‡
Xidian University‡

Abstract

多任务学习在自然语言处理、语音识别、计算机视觉和药物设计等多个机器学习应用上取得了巨大成功。本文旨在对多任务学习，特别是深层神经网络中的多任务学习，进行一般性的综述。本文介绍了深度学习中多任务学习的两种最常见的框架，概括和讨论了相关文献和最新进展。最终，本文希望能够通过阐释多任务学习的工作机理来帮助机器学习研究人员在实践中应用多任务学习，并为选择合适的辅助任务提供了指导性建议。

1. 引言

在机器学习中，我们一般关心如何优化一个特定指标，这个指标可能是某个基准测试的分数，也可能是商业 KPI。为此，我们一般训练单个模型或者一组集成模型来完成这个任务。然后，我们会微调这些模型直到它们的性能不再提升。虽然我们通常可以通过这种方式获得可接受的效果，但是由于局限于我们关注的单个任务，常常忽略了可能帮助我们获得更好性能的那些信息。特别地，这些信息来自于相关任务的训练信号。通过让多个相关任务共享表示，我们能够让模型在原始任务上达到更好的泛化能力。这种方法被称作多任务学习（Multi-Task Learning, MTL）。

多任务学习已经在机器学习的各个应用领域中被广泛使用，包括自然语言处理 [14]、语音识别 [15]、计算机视觉 [19]和药物设计 [34]。多任务学习有很多形式：联合学习、学习如何学习、通过辅助任务学习等，这些概念只不过是多任务学习的不同名称。一般地，只要你在优化多个损失函数，你就已经在进行多任务学习了。在这些场景中，显式地考虑多任务学习的概念和机理对于提升性能会有所帮助。

即使你只是像通常设定的那样在优化一个损失函数，也有可能存在能够帮助提升主任务性能的辅助任务。Caruana 等人 [11]对多任务学习给出了一个简洁的总结：“多任务学习通过利用包含在相关任务训练信号中的领域特定信息来提升泛化能力”。

在本文中，我们将介绍多任务学习的研究现状，特别是当其应用在深度神经网络中的情况。首先，我们将在第 2 节中从不同的视角来解释多任务学习，然后我们将在第 3 节中介绍深度学习中的两种最常见的多任务学习方法，在第 4 节，我们将描述使得多任务学习在实践中奏效的机制。在介绍最近的基于神经网络的多任务学习方法之前，我们会先在第 5 节讨论非神经网络的多任务学习。在第 6 节中将会介绍近期提出的强大的基于神经网络的多任务学习方法。最后，我们将在第 7 节中介绍常被用到的辅助任务并讨论如何构造好的辅助任务。

2. 动机

我们可以从多个角度来解释多任务学习的动机。在生物学上，我们可以把多任务学习看作是受到人类学习的启发。在学习新任务时，我们经常运用我们在相关任务中学到的知识。例如，婴儿首先学会识别人脸，然后可以将这一知识应用于识别其他物体。

† 文章原作者及单位
‡ 文章译者及单位

从教育学的角度来看，我们经常首先学习那些帮助我们掌握更复杂的技术的必要技能。我们在学习编程、武术等专业技能时都是如此。

举一个更为具体的例子，在电影空手道小子 (1984) 中，宫城老师教这个空手道的孩子一些看似无关的任务，比如擦地板和给汽车打蜡。然而，事后看来，这些都为他提供了与学习空手道相关的宝贵技能。

最后，我们可以从机器学习的角度来解释多任务学习：多任务学习可以被看作是一种归纳转移（Inductive Transfer）。归纳转移可以通过在模型中引入偏向某些假设的归纳偏差来帮助改进模型。例如，归纳偏差的一种常见形式是 ℓ_1 正则化，它导致对稀疏解的偏好。在 MTL 的情况下，归纳偏差由辅助任务提供，这使得模型更倾向于选择能够同时解释多个任务的假设，这通常会导致更好的泛化能力。

3. 深度学习中的两种多任务学习方法

至此，我们已经在理论上讨论了多任务学习的机理，下面我们将介绍两种具体的多任务学习方法，它们在深层神经网络中有广泛的应用。在深度学习中，多任务学习一般通过隐层参数*硬共享*和*软共享*两种方式来实现。

3.1. 硬参数共享

硬参数共享是神经网络中最常用的 MTL 方法，可以追溯到 1993 年 [10]。它通常通过在所有任务之间共享隐藏层来应用，同时保留几个特定于任务的输出层，如图 1 所示。

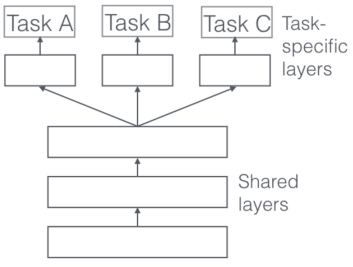

图 1. 硬参数共享

硬参数共享大大降低了过拟合的风险。事实上，Baxter 等人 [7]表明，过拟合共享参数的风险比过拟合任务特定参数（即输出层）要小 N 倍（这里 N 是任务数量）。这在直觉上是有道理的：我们同时学习的任务越多，我们的模型就越能找到捕获所有任务的表示，而我们对原始任务过拟合的可能性就越小。

3.2. 软参数共享

在软参数共享中，每个模型都有自己的模型和参数。为了使得不同模型之间的参数尽可能得相似，软共享对模型参数之间的距离施加正则化约束，如图 2 所示。比如，Duong 等人 [16]使用 ℓ_2 距离进行正则化，而 Yang and Hospedales 等人 [40]使用迹范数（trace norm）。

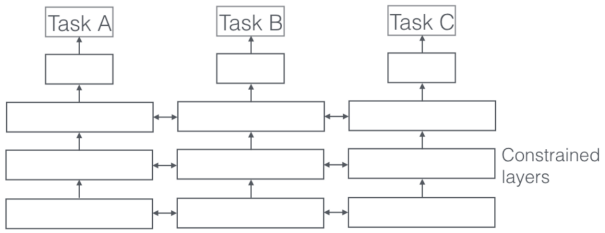

图 2. 软参数共享

深层神经网络中用于软参数共享的约束受到了为其他模型开发的多任务学习正则化技术的极大启发，我们将很快讨论这些技术。

4. 为什么多任务学习有效

即使通过多任务学习获得的归纳偏置似乎直观可信，但为了更好地理解多任务学习，我们需要研究其背后的机制。其中大部分首先由 Caruana [11]提出。对于所有示例，我们将假设我们有两个相关的任务 A 和 B，它们依赖于共享隐藏层表示 F.

4.1. 隐式数据增强

MTL 有效地增加了我们用于训练模型的样本量。由于所有任务都存在噪声，当在某个任务 A 上训练模型时，我们的目标是学习任务 A 的良好表示，理想情况下这个表示能够忽略数据相关的噪声并取得良好的泛化能力。由于不同的任务具有不同的噪声模式，同时学习两个任务的模型能够学习更一般的表示。学习任务 A 会有过拟合任务 A 的风险，而同时学习 A 和 B 共同使模型能够通过平均噪声模式获得更好的表示 F.

4.2. 注意力聚集

如果任务噪声较大，或者数据高维而数据量有限，模型将很难分辨相关特征和无关特征。多任务学习可以帮助模型将注意力集中在那些真正重要的特征上，因为其他任务将为这些特征的相关性或不相关性提供额外的证据。

4.3. 窃听

对于某些任务 B，某些特征 G 易于学习，而对于另一个任务 A 则难以学习。这可能是因为 A 以更复杂的方式与特征交互，或者因为其他特征阻碍了模型学习 G 的能力。通过多任务学习，我们可以允许模型被*窃听*，即通过任务 B 学习 G. 最简单的方法是通过*提示* [1]，即直接训练模型来预测最重要的特征。

4.4. 表示偏置

多任务学习使得模型倾向于选择能够同时解释多个任务的表示。这也将有助于模型在未来推广到新任务，因为对于足够大量的训练任务表现良好的假设空间也可以很好地学习新任务，只要它们来自同一环境 [8]。

4.5. 正则化

最终，多任务学习可以被当做一种正则化手段，因为它引入了归纳偏置。因此，它降低了过拟合的风险以及模型的 Rademacher 复杂性（即其适应随机噪声的能力）。

5. 非神经网络中的多任务学习

为了更好地理解深度神经网络中的多任务学习，我们现在将查看有关线性模型，核方法和贝叶斯方法中多任务学习的现有文献。特别是，我们将讨论在多任务学习历史中普遍存在的两个主要思想：1）通过正规化强化跨任务的稀疏性；2）建模任务之间的关系。

注意，文献中许多多任务学习方法涉及同质设置：它们假设所有任务都与单个输出相关联，例如，多类 MNIST 数据集通常被转换为 10 个二进制分类任务。最近的方法涉及更实际的异质设置，其中每个任务对应于一组唯一的输出。

5.1. 块稀疏的正则化

符号说明. 为了更好的解释下面的方法，我们首先介绍用到的符号的意义。假设我们有 T 个任务，对于每一个任务 t，有一个对应的模型 m_t，该模型有 d 维参数 $\boldsymbol{a}_t$. 我们可以用一个列向量表示模型参数：

$$\boldsymbol{a}_t = \begin{bmatrix} a_{1,t} \\ \vdots \\ a_{d,t} \end{bmatrix}^\top$$

现在我们把这些列向量 $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T$ 按列组成一矩阵 $\boldsymbol{A} \in \mathbb{R}^{d \times T}$. 矩阵 $\boldsymbol{A}$ 的第 i 行包含的参数 $a_{i,\cdot}$ 对应模型在每个任务上的第 i 个特征，而矩阵的第 j 列包含的参数 $a_{\cdot,j}$ 对应第 j 个模型。

很多现有的工作斗鱼模型的参数作了稀疏性假设。Argyriou and Pontil 等人 [4]假设所有的模型都共享少量特征。这意味着在我们的矩阵 $\boldsymbol{A}$ 中除了少数行元素都为 0，也就是说只有少量特征在所有任务中都被使用。为此，他们将 ℓ_1 范数推广到多任务设定中。由于 ℓ_1 范数是对参数数量进行约束，这使得除了少量参数之外的大部分参数都为 0，这种方法也被称作 LASSO.

在单任务中，ℓ_1 范数是在任务 t 的参数 $\boldsymbol{a}_t$ 上计算。而在多任务中，我们是在矩阵 $\boldsymbol{A}$ 上计算。为此，我们首先计算对应于所有任务的第 i 个特征的参数，即矩阵的每一行 $\boldsymbol{a}_i$ 的 ℓ_q 范数，得到一向量 $\boldsymbol{b} = [\|\boldsymbol{a}_1\|_q \ldots \|\boldsymbol{a}_d\|_q] \in \mathbb{R}^d$. 然后我们计算该向量的 ℓ_1 范数，这使得向量 $\boldsymbol{b}$ 除少数元素外都为 0，也即 $\boldsymbol{A}$ 的除少数行外元素为 0.

我们可以看到，根据我们想要在每一行上施加的约束，我们可以使用不同的 ℓ_q 范数。通常，我们将这些混合范数约束称为 ℓ_1/ℓ_p 范数。它们也被称为块稀疏正则化，因为它们导致 $\boldsymbol{A}$ 的整行被设置为 0. Zhang and Huang 等人 [44]使用 ℓ_1/ℓ_∞ 正则化，而 Argyriou and Pontil 等 [4]使用混合 ℓ_1/ℓ_2 范数。后者也称为组 LASSO，最初由 Yuan and Lin 等人 [43]提出。

在 [4]中，通过惩罚矩阵 $\boldsymbol{A}$ 的迹范数将非凸组 LASSO 优化问题转化为凸优问题，使得 $\boldsymbol{A}$ 称为低秩矩阵，因而将列向量约束在一低纬子空间中。Lounici 等人 [30]进一步证明了多任务学习中组 LASSO 的上界。

尽管这种块稀疏正则化在直觉上是合理的，但它非常依赖于跨任务共享特征的程度。Negahban and Wain-

wright 等人的工作 [33]表明，如果特征不重叠，那么 ℓ_1/ℓ_q 正则化可能实际上比元素级的 ℓ_1 正则化的效果更差。

出于这个原因，Jalali 等人 [23]通过提出一种结合了块稀疏和逐元素稀疏正则化的方法来改进块稀疏模型。他们将任务参数矩阵 $\boldsymbol{A}$ 分解为两个矩阵 $\boldsymbol{B}$ 和 $\boldsymbol{S}$，其中 $\boldsymbol{B}=\boldsymbol{A}+\boldsymbol{S}$. 然后使用 ℓ_1/ℓ_∞ 正则化强制 $\boldsymbol{B}$ 为块稀疏，而使用 LASSO 使 $\boldsymbol{S}$ 成为元素稀疏。最近，Liu 等人 [27]提出了分组稀疏正则化的分布式版本。

5.2. 学习任务之间的关系

虽然群组稀疏性约束迫使我们的模型仅考虑一些特征，但这些特征主要用于所有任务。因此，所有先前的方法都假设在多任务学习中使用的任务密切相关。但是，每个任务可能与所有可用任务都不密切相关。在这些情况下，与不相关的任务共享信息实际上可能会损害性能，这种现象称为负转移。

因此，我们希望利用某些先验知识来知道哪些任务是相关的而其他任务则不相关，而不是通过稀疏性。在这种情况下，使用任务簇可能更合适。Evgeniou 等 [17]提出通过惩罚我们的任务列向量 $a_{\cdot,1},\ldots,a_{\cdot,T}$ 的范数来对任务簇和它们的方差进行约束：

$$\Omega=\|\overline{a}\|^2+\frac{\lambda}{T}\sum_{t=1}^{T}\|a_{\cdot,t}-\overline{a}\|^2$$

这里 $\overline{a}=(\sum_{t=1}^{T}a_{\cdot,t})/T$ 为平均参数向量。该惩罚项通过 λ 来控制不同任务的参数到它们的平均值的距离。该方法被应用到核方法中，但也可以被应用到线性模型中去。

Evgeniou and Pontil 等人 [17]也提出了类似的 SVM 约束。他们的约束受贝叶斯方法的启发，并试图使所有模型接近某种平均模型。在 SVM 中，损失函数权衡每个 SVM 模型与平均参数的距离。

Jacob 等人 [22]在假设预先知道聚类个数 C 的情况下对 $\boldsymbol{A}$ 进行聚类约束来使得聚类正则化的假设更加明确。他们将正则项分解为三个单独的部分：

- 用于衡量列参数向量的平均值的全局惩罚项：$\Omega_{mean}(\boldsymbol{A})=\|\overline{a}\|^2$.
- 用于衡量类间距离的指标：$\Omega_{between}(\boldsymbol{A})=\sum_{c=1}^{C}T_c\|\overline{a}_c-\overline{a}\|^2$，这里 T_c 是第 c 个任务类的任务数量，$\overline{a_c}$ 是第 c 个任务类的参数向量的平均。
- 用于衡量类内距离的指标：$\Omega_{within}(\boldsymbol{A})=\sum_{c=1}^{C}\sum_{t\in J(c)}\|a_{\cdot,t}-\overline{a}_c\|$，其中 $J(c)$ 是第 c 个任务类中任务的集合。

最终的约束由上述三项的加权和组成：

$$\Omega(\boldsymbol{A})=\lambda_1\Omega_{mean}(\boldsymbol{A})+\lambda_2\Omega_{between}(\boldsymbol{A})+\lambda_3\Omega_{within}(\boldsymbol{A})$$

这里假设提前已知任务聚类的个数，他们还在上面的约束中引进了一种凸松弛来运行模型能够同时学习聚类的个数。

上面的方法使用范数正则化的方式来建模任务之间的关系，还有一些方法没有使用正则化，例如使用半监督学习方法来从多个相关任务中学习一个共有结构 [3]。

此外，还有一些工作使用了贝叶斯方法来学习多任务之间的关系：Heskes 等人 [21]提出了一种贝叶斯网络，通过在模型参数上施加先验来使得不同任务的模型之间具有相似的参数。Lawrence and Platt 等人 [26]通过推断共享协方差矩阵的参数将高斯过程（GP）扩展到了多任务学习中。由于这样将导致非常高的计算复杂度，他们使用了稀疏近似机制来贪心地选择信息量最大的样本。类似的，Yu 等人 [42]基于所有模型都是从某个公有的先验中采样得到的假设将高斯过程应用到多任务学习中。

Bakker and Heskes 等人 [6] 将高斯作为先验分布放在每个任务的特定层上。为了鼓励不同任务之间的相似性，他们提出平均任务依赖并将混合分布引入任务的聚类。然而，他们的方法需要任务特征来定义任务簇并且需要预先指定的聚类个数。

Kang 等人 [24]假设任务形成不相交的簇，并且每个簇内的任务位于低维子空间中。在每个簇内，任务共享相同的特征表示，其参数与簇分配矩阵一起使用交替最小化方案联合学习。但是，簇之间的完全不相交可能不是理想的方式，因为这些任务可能仍然共享一些有助于预测的特征。

6. 深度多任务学习的研究进展

虽然许多最近的深度学习方法显式或隐式地使用多任务学习作为其模型的一部分，但它们都采用了我们之前介绍的两种方法，硬参数共享和软参数共享。相比之下，只有少数论文研究了在深层神经网络中如何开发更好的多任务学习机制。

6.1. 深层关系网络

在计算机视觉的多任务学习方法中，通常共享卷积层，同时学习特定于任务的全连接层。Long and Wang 等人 [29] 通过提出深层关系网络来改进这些模型。除了共享和任务特定层的结构（如图 3 所示），它们将矩阵先验放置在全连接层上，这允许模型学习任务之间的关系。然而，这种方法仍然依赖于预先定义的共享结构，虽然可能足以解决论文中研究的计算机视觉问题，但对于新任务而言非常容易出错。

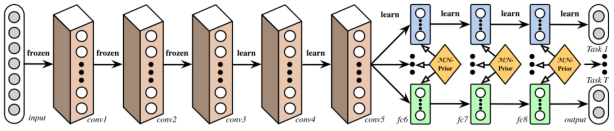

图 3. 深层关系网络

6.2. 完全自适应的特征共享

在另一方面，Lu 等人 [31] 提出了一种自下而上的方法，该方法从瘦网络开始，在训练期间使用促进类似任务分组的标准贪心地扩展此网络。可以在图 4 中看到动态创建分支的网络扩展过程。但是，贪心的方法可能无法找到全局最优的模型，此外将每个分支分配给一个特定的任务也导致模型无法学习任务之间更复杂的交互。

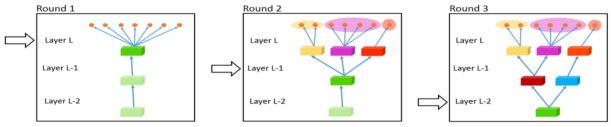

图 4. 完全自适应特征共享网络的扩展过程

6.3. 十字绣网络

Misra 等人 [32]从两个独立的模型架构开始（就像在软参数共享中一样），使用他们的十字绣单元，以允许模型通过学习先前层的输出的线性组合来确定特定于任务的网络以何种方式利用其他任务的知识。它们的结构可以在图 5 中看到，其中它们仅在池化层和全连接层之后放置十字绣单元。

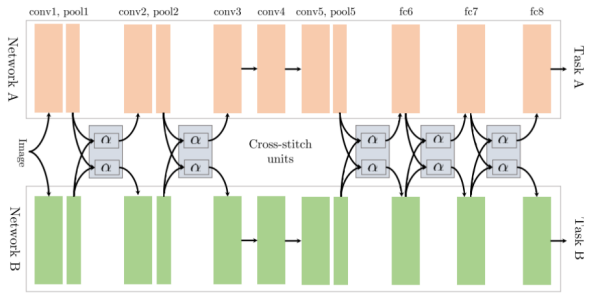

图 5. 十字绣网络

6.4. 分层监督

相比之下，在自然语言处理（NLP）中，近期的工作更多地关注如何为多任务学习找到更好的任务结构层次：Søgaard 和 Goldberg 等人 [38] 通过实验表明当低级别的 NLP 任务（即通常用于预处理的 NLP 任务，如词性标注和命名实体识别）用作辅助任务时，应在较低层进行监督。将监督信号分层施加的模式也被称为分层共享。

6.5. 联合多任务学习模型

基于上面的发现，Hashimoto 等人 [20] 为多个 NLP 任务预先定义了层次结构，如图 6 所示。

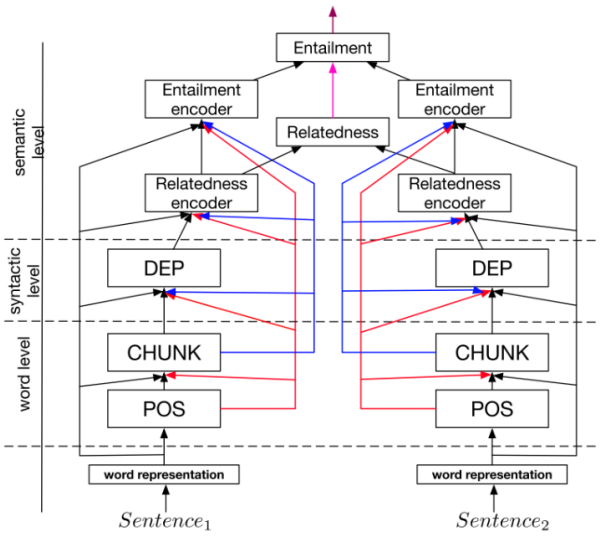

图 6. 联合多任务学习模型

6.6. 使用不确定性加权损失

不同于之前的学习共享结构的工作，Kendall 等人 [25] 通过考虑每个任务的不确定性提出了一个与之前方法正交的方法。他们通过最大化依赖于任务的不确定性的高斯似然来确定不同任务的损失函数的相对权重。他们的框架如图 7 所示，在该框架中联合学习像素深度回归、语义分割和实例分割三个计算机视觉任务。

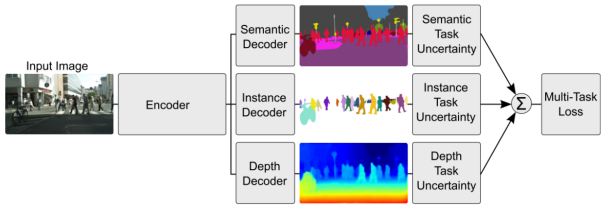

图 7. 基于不确定性的损失函数加权

6.7. 张量分解

在最近的研究中，已有的多任务学习方法被扩展到深度学习中：Yang and Hospedales 等人 [40] 使用张量因子分解来推广一些先前讨论的矩阵因子分解方法，以将模型参数分成每个层的共享和任务特定参数。

6.8. 水闸网络

Ruder 等人 [36] 提出了水闸网络，推广了基于深度学习的多任务学习方法，诸如硬参数共享、十字绣网络、块稀疏正则化方法以及最近 NLP 领域中的构造任务层次。该模型（如图 8 所示）能够学习哪些层以及这些层的子空间应当被共享，以及在神经网络的哪些层学到了输入句子的最好的表示。

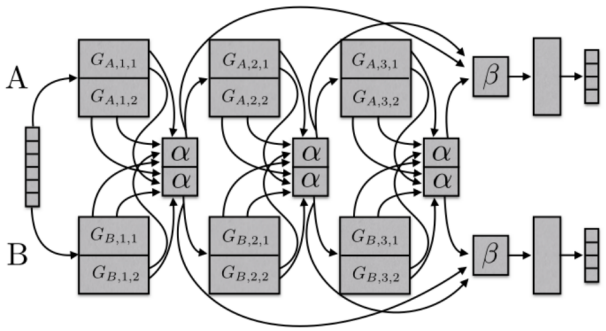

图 8. 水闸网络

6.9. 模型应当共享什么？

在对这些最近的方法进行了讨论后，现在我们简要总结一下模型应当共享的内容。多任务学习中的大多数方法都集中在从同一分布中抽取任务的情景 [7]。虽然这种情况有利于共享，但实际上并不总是如此。为了开发更强大的 MTL 模型，我们必须能够处理不相关或弱相关的任务。

深度学习中的多任务学习的一些早期工作预先指定了为每个任务配对分配哪些层，但此方式难以扩展并且严重依赖 MTL 架构。最初由 Caruana 等 [10] 提出的硬参数共享模式在 20 年后仍然是常态。虽然在许多情况下都很有用，但如果任务不紧密相关或需要在不同级别进行推理决策，那么硬参数共享模式将很快失效。因此，最近的方法着眼于学习分享什么并且通常获得了优于硬参数共享的效果。因此，赋予模型学习任务层次的能力是有必要的，特别是在需要不同粒度的情况下。

正如最开始所说的，只要我们在优化多于一个损失函数，就属于多任务学习的范畴，而不应狭隘地约束模型来将所有任务的知识压缩到相同的参数空间中。这样有助于利用我们讨论过的多任务学习中的进展并使我们的模型能够学习任务之间应该如何相互作用。

7. 辅助任务

多任务学习非常适合当我们希望同时获得多个任务的预测时的场景。这种情况在金融或经济预测中是常见的，我们可能希望预测许多可能相关指标的价值，或者在我们可能希望同时预测多种疾病症状的生物信息学中。在诸如药物发现的情景中，应该预测数十或数百种活性化合物，MTL 可以使得准确度随着任务的数量的增多不断增加。

但是，在大多数情况下，我们只关心一项任务的表现。在本节中，我们将研究如何找到合适的辅助任务，以期仍然能获得多任务学习的收益。

7.1. 相关任务

事实上，使用相关任务作为 MTL 的辅助任务是一个常见的设定。为了了解相关任务的内容，这里将提供一些典型的例子：Caruana 等 [11] 使用预测道路不同特征的任务作为预测自动驾驶汽车转向方向的辅助任

务；Zhang 等 [45] 使用头部姿势估计和面部关键点推断作为面部特征检测的辅助任务；Liu 等人 [28]联合学习查询分类和网络搜索；Girshick 等人 [19]同时预测图像中目标的类别和坐标。最后，Arık 等 [5]联合预测文本到语音的音素持续时间和频率分布。

7.2. 对抗

通常情况下，我们很难获得一个相关任务的标注数据。但是，在某些情况下，我们可以使用与我们想要实现的任务相反的任务。可以使用对抗损失来利用该数据，该对抗性损失不寻求最小化但是使用梯度反转层来最大化训练误差。这种设置最近在域适应方面取得了成功 [18]。在这种情况下，对抗性任务是预测输入的域；通过逆转对抗性任务的梯度，使得对抗性任务损失最大化，因为它迫使模型学习到一个无法区分域的表示，从而提高主任务的性能。

7.3. 提示

如前所述，MTL 可用于学习仅使用原始任务可能不容易学习到的特征。实现此目的的有效方法是使用提示，即将特征预测为辅助任务。最近在自然语言处理背景下这种策略的例子是 [41]，他们预测输入句子是否包含正面或负面情绪词作为情感分析的辅助任务，以及 Cheng 等人 [13]预测句子中是否存在名称作为名称错误检测的辅助任务。

7.4. 注意力聚集

类似地，辅助任务可用于将注意力集中在网络通常可忽略的图像部分上。例如，为了学习驾驶 [11]，单任务模型通常会忽略车道标记，因为它们仅构成图像的一小部分并且不总是存在。然而，将车道标记预测为辅助任务会迫使模型学习表示它们；这些知识也可以用于主任务。类似地，对于面部识别，可以学习预测面部关键点的位置作为辅助任务，因为这些关键点通常是因人而异的。

7.5. 量化平滑

对于许多任务，训练目标是可量化的，例如，连续的标量可能更适合作为标签，但已有的标签却是离散的。在许多需要人类评估数据收集的场景中就是这种情况，例如预测疾病风险（例如低/中/高）或情感分析（正/中/负）。在这些情况下使用量化的辅助任务可能会有所帮助，因为它们的目标更平滑，因此可以更容易地学习它们。

7.6. 预测输入

在某些情况下，使用某些特征作为输入是不切实际的，因为它们对于预测所需目标没有帮助。但是，他们仍然可以指导学习任务。在这些情况下，这些特征可以用作输出而不是输入。Caruana and de Sa 等人 [12]提出了几个适用的问题。

7.7. 通过未来预测当前

在许多情况下，某些特征仅在应该进行预测后才可用。例如，对于自动驾驶汽车，一旦汽车经过它们，就可以更精确地测量障碍物和车道标记。Caruana 等 [11]也给出了肺炎预测的例子，额外的医学试验结果会在之后给出。对于这些示例，附加数据不能用作特征，因为它在运行时不可用作输入。但是，它可以用作辅助任务，以在训练期间向模型输入额外的知识。

7.8. 表示学习

MTL 中辅助任务的目标是使模型能够学习对主任务共享或有用的表示。到目前为止讨论的所有辅助任务都隐含着一个假设：它们与主任务密切相关，因此学习它们可能允许模型学习有益的表示。对此可以进行更明确的建模，例如通过采用已知使模型能够学习可转移表示的任务。Cheng 等 [13]和 Rei 等 [35]采用语言模型作为辅助任务就是这样的例子。类似地，自编码器的目标也可以用作辅助任务。

7.9. 什么辅助任务是有效的?

在本节中，我们讨论了怎样在即使只关心单一任务的情况下通过辅助任务来获得多任务学习的增益的一些方法。但是，我们仍然不知道在实践中哪些辅助任务是有用的。寻找辅助任务很大程度上是基于这样的假设：辅助任务应该以某种方式与主任务相关，并且它应该有助于预测主要任务。

但是，我们仍然不清楚何时应将两项任务视为相似或相关。Caruana 认为如果使用相同的特征做出决定，则将两个任务定义为相似 [11]。Baxter 等人 [8]则认为相关任务应当共享一个公共的最优假设类，即拥

有相同的归纳偏置。Ben-David 和 Schuller 等人 [9]提出如果两个任务的数据可以使用一组变换 F 从固定的概率分布生成，则两个任务是 F 相关的。虽然这允许推理通过不同传感器收集数据的任务（但是是相同的分类任务，例如使用来自具有不同角度和照明条件的摄像机的数据进行物体识别），但是它不适用于不处理相同问题的任务。Xue 等人 [39]认为，如果分类边界即参数向量接近，那么两个任务是相似的。

尽管在理解任务相关性方面取得了一些早期的理论进展，但我们距离实现这一目标还有很大距离。任务相似性不是二元的，而是存在于频谱上。允许我们的模型学习与每个任务共享的内容或许能够使得我们暂时避免理论的缺乏并更好地利用即使只是弱相关的任务。然而，我们仍然需要针对 MTL 制定更具原则性的任务相似性概念，以便了解应该选择哪些任务。

最近的工作 [2]发现具有紧密、均匀的标签分布的辅助任务对于 NLP 中的序列标注问题非常有帮助，Ruder 等人 [36]也在实验中证实了这一规律。此外，Bingel 和 Søgaard 发现通过非平稳辅助任务，主任务能够迅速达到稳定水平 [37]。然而，迄今为止的这些实验还只是局限在小范围内，最近的发现也只是为更深入地理解深层神经网络中的多任务学习提供了早期的线索。

8. 总结

这这篇综述中，我们回顾了多任务学习的历史以及在深度学习中的更近期的进展。虽然多任务学习已经被频繁地使用，但二十年前的硬参数共享模式仍然在基于深层神经网络的 MTL 中随处可见。近年来的研究开始转而研究让神经网络学习共享的内容。我们对任务，以及任务之间的相似性、关系、层次以及用于 MTL 时的收益，仍然是非常局限的，目前还需要对神经网络中的多任务学习的泛化能力进行更深入的研究和理解。

译者小结. 在过去的五年中，深度学习为自然语言处理领域带来了深刻的变革，人们乐此不疲地探索各种神经网络结构，从 CNN 到 RNN 再到 Transformer。然而，近期随着迁移学习在各大 NLP 任务上取得一致性成功，以及一些简单的模型通过设计巧妙的训练任务和策略获得的出色表现，人们开始注意到训练任务（或者说监督信号）的重要意义，我相信随着对任务更好的理解和建模，深度学习还将在 NLP 领域取得更大的进展。

参考文献

[1] Y. S. Abu-Mostafa. Learning from hints in neural networks. J. Complexity, 6(2):192–198, 1990. 3

[2] H. M. Alonso and B. Plank. When is multitask learning effective? semantic sequence prediction under varying data conditions. arXiv preprint arXiv:1612.02251, 2016. 8

[3] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 6:1817–1853, 2005. 4

[4] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006, pages 41–48. MIT Press, 2006. 3

[5] S. Ö. Arik, M. Chrzanowski, A. Coates, G. F. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Y. Ng, J. Raiman, S. Sengupta, and M. Shoeybi. Deep voice: Real-time neural text-to-speech. In D. Precup and Y. W. Teh, editors, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 195–204. PMLR, 2017. 7

[6] B. Bakker and T. Heskes. Task clustering and gating for bayesian multitask learning. Journal of Machine Learning Research, 4:83–99, 2003. 4

[7] J. Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. Machine Learning, 28(1):7–39, 1997. 2, 6

[8] J. Baxter. A model of inductive bias learning. J. Artif. Intell. Res., 12:149–198, 2000. 3, 7

[9] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In Learning Theory and Kernel Machines, pages 567–580. Springer, 2003. 8

[10] R. Caruana. Multitask learning: A knowledge-based source of inductive bias. In Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June

27-29, 1993, pages 41–48. Morgan Kaufmann, 1993. 2, 6

[11] R. Caruana. Multitask Learning. Machine Learning, 28(1):41–75, 1997. 1, 2, 6, 7

[12] R. Caruana and V. R. de Sa. Promoting poor features to supervisors: Some inputs work better as outputs. In M. Mozer, M. I. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996, pages 389–395. MIT Press, 1996. 7

[13] H. Cheng, H. Fang, and M. Ostendorf. Open-domain name error detection using a multi-task RNN. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, editors, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pages 737–746. The Association for Computational Linguistics, 2015. 7

[14] R. Collobert and J. Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008, volume 307 of ACM International Conference Proceeding Series, pages 160–167. ACM, 2008. 1

[15] L. Deng, G. E. Hinton, and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: an overview. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013, pages 8599–8603. IEEE, 2013. 1

[16] L. Duong, T. Cohn, S. Bird, and P. Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers, pages 845–850. The Association for Computer Linguistics, 2015. 2

[17] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. Journal of Machine Learning Research, 6:615–637, 2005. 4

[18] Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In F. R. Bach and D. M. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of JMLR Workshop and Conference Proceedings, pages 1180–1189. JMLR.org, 2015. 7

[19] R. B. Girshick. Fast R-CNN. In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pages 1440–1448. IEEE Computer Society, 2015. 1, 7

[20] K. Hashimoto, C. Xiong, Y. Tsuruoka, and R. Socher. A joint many-task model: Growing a neural network for multiple NLP tasks. In M. Palmer, R. Hwa, and S. Riedel, editors, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 1923–1933. Association for Computational Linguistics, 2017. 5

[21] T. Heskes. Empirical bayes for learning to learn. In P. Langley, editor, Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000, pages 367–374. Morgan Kaufmann, 2000. 4

[22] L. Jacob, F. R. Bach, and J. Vert. Clustered multi-task learning: A convex formulation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008, pages 745–752. Curran Associates, Inc., 2008. 4

[23] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada., pages 964–972. Curran Associates, Inc., 2010. 4

[24] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In L. Getoor and T. Scheffer, editors, Proceedings of the

28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, pages 521–528. Omnipress, 2011. 4

[25] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 7482–7491. IEEE Computer Society, 2018. 6

[26] N. D. Lawrence and J. C. Platt. Learning to learn with the informative vector machine. In C. E. Brodley, editor, Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004, volume 69 of ACM International Conference Proceeding Series. ACM, 2004. 4

[27] S. Liu, S. J. Pan, and Q. Ho. Distributed multi-task relationship learning. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, pages 937–946. ACM, 2017. 4

[28] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y. Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In R. Mihalcea, J. Y. Chai, and A. Sarkar, editors, NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015, pages 912–921. The Association for Computational Linguistics, 2015. 7

[29] M. Long and J. Wang. Learning multiple tasks with deep relationship networks. arXiv preprint arXiv:1506.02117, 2, 2015. 5

[30] K. Lounici, M. Pontil, A. B. Tsybakov, and S. A. van de Geer. Taking advantage of sparsity in multi-task learning. In COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009, 2009. 3

[31] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. S. Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 1131–1140. IEEE Computer Society, 2017. 5

[32] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 3994–4003. IEEE Computer Society, 2016. 5

[33] S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of 1,∞-regularization. In Proceedings of the 21st International Conference on Neural Information Processing Systems, pages 1161–1168. Curran Associates Inc., 2008. 4

[34] B. Ramsundar, S. M. Kearnes, P. Riley, D. Webster, D. E. Konerding, and V. S. Pande. Massively multitask networks for drug discovery. CoRR, abs/1502.02072, 2015. 1

[35] M. Rei. Semi-supervised multitask learning for sequence labeling. In R. Barzilay and M. Kan, editors, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 2121–2130. Association for Computational Linguistics, 2017. 7

[36] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard. Latent multi-task architecture learning, 2017. 6, 8

[37] A. Søgaard and J. Bingel. Identifying beneficial task relations for multi-task learning in deep neural networks. In M. Lapata, P. Blunsom, and A. Koller, editors, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers, pages 164–169. Association for Computational Linguistics, 2017. 8

[38] A. Søgaard and Y. Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers. The Association for Computer Linguistics, 2016. 5

[39] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. Journal of Machine Learning Research, 8:35–63, 2007. 8

[40] Y. Yang and T. M. Hospedales. Deep multi-task representation learning: A tensor factorisation approach. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. 2, 6

[41] J. Yu and J. Jiang. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In J. Su, X. Carreras, and K. Duh, editors, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pages 236–246. The Association for Computational Linguistics, 2016. 7

[42] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. In L. D. Raedt and S. Wrobel, editors, Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005, volume 119 of ACM International Conference Proceeding Series, pages 1012–1019. ACM, 2005. 4

[43] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006. 3

[44] C.-H. Zhang, J. Huang, et al. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics, 36(4):1567–1594, 2008. 3

[45] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI, volume 8694 of Lecture Notes in Computer Science, pages 94–108. Springer, 2014. 7