面向目标的语义通信中的动态相对表示

Simone Fiorellino^{1,2}, Claudio Battiloro³, Emilio Calvanese Strinati⁴, Paolo Di Lorenzo^{1,2}

¹DIET department, Sapienza University, Rome, Italy. ²CNIT, Parma, Italy.

³Harvard T.H. Chan School of Public Health, Harvard University, Boston, U.S.A.

⁴CEA Leti, University Grenoble Alpes, 38000, Grenoble, France

Emails: {simone.fiorellino,paolo.dil orenzo}@uniroma1.it, cbattiloro@hsph.harvard.edu, emilio.calvanese-strinati@cea.fr

摘要—在未来 6G 无线网络中,通信的语义和有效性方面将 发挥基础性作用,将意义和相关性纳人传输。

然而,当设备使用不同的语言、逻辑或内部表示时,会出现 障碍,导致可能危及理解的语义不匹配。在潜在空间通信中,这 一挑战表现为高维表示中的错位,其中深度神经网络编码数据。 本文提出了一种面向目标的语义通信的新框架,利用相对表示通 过潜在空间对齐来缓解语义不匹配问题。我们提出了一种动态 优化策略,该策略适应相对表示、通信参数和计算资源以实现节 能、低延迟、面向目标的语义通信。数值结果证明了我们的方法 在减轻设备间的不匹配方面以及优化能耗、延迟和有效性方面的 有效性。

Index Terms—目标导向的语义通信,相对表示,深度学习,随机优化。

I. 介绍

经典通信范式优先考虑个体符号或比特的准确传 输,而不明确考虑其潜在意义。正如韦弗所识别的,可 以通过三个层次来理解通信:语法层(专注于信号的可 靠传输)、语义层(关注那些信号传达的意义)和有效性 层(确保消息达到其预期目的)[1]。然而,数据密集型无 线应用指数级增长的需求要求对此方法进行根本性的 转变。自动驾驶汽车和增强现实等应用程序所需的信息 量压垮了传统以比特为中心的通信系统的容量。相比之 下,现代网络越来越多地需要交换具有共同意义的信息 [2]。语义通信(SemCom)通过直接将意义嵌入传输中 来满足这一需求,超越了单纯传输比特的功能[3]-[6]。 为了实现这一点,SemCom 严重依赖人工智能(AI)和 深度神经网络(DNNs),从数据中提取并传输关键的语 义特征,而不是原始数据本身。不专注于原始符号的完 美重建, SemCom 更优先于传输潜在意义, 为噪声和不 完美的传输条件提供了固有的鲁棒性。在接收方一侧, 一个 DNN 解码器重构并解读预期的消息。SemCom 的 灵活性和韧性使其成为理想的选择,适用于设备在动态 且资源受限环境中操作的新应用 [7]。在此想法的基础 上, 文献中提出了新的语义联合源和信道编码 [5] 方法、 语义提取和压缩 [8], [9]、目标导向的系统设计和优化 [7], [10]、语义推理 [11] 以及基于生成式人工智能的语 义通信 [12]。尽管有了这些进展,由语义信息误读导致 的干扰,即语义噪声,仍是一个重大挑战。这种干扰通 常源于不同设备表示和解读数据方式的差异,即使物理 传输无误也会导致理解错误。这个问题在动态边缘环境 中尤为突出,在这种环境下,涉及多种设备且复杂系统 范围内的重新训练是不切实际的。高维数据表示中的不 对齐会在这些条件下严重扰乱语义理解 [13]。这突显了 需要机制来缓解语义错配而无需进行系统更新的需求。 为应对这些挑战,本文提出了一种基于最近的相对表 示(RelReps)概念的新型目标导向语义通信框架 [14], [15]。我们的主要贡献是两方面的: (i) 我们在 SemComs 中引入了相对表示以减轻动态场景中设备具有异构逻 辑时的语义不匹配影响;(ii)我们基于 Lyapunov 随机 优化 [16] 引入了一种动态优化策略, 该策略能够动态地 优化通信(即速率)、计算(即 CPU 频率周期)和学习 资源(即锚定集、编码器),从而实现节能、低延迟的 有效语义通信。所提出的动态资源分配方案具有简单的 闭合形式解,并且不需要事先了解影响系统的随机变量 的统计信息(例如,无线信道)。数值结果评估了我们 的方法在实现语义对齐的同时,在能耗、延迟和目标有 效性之间取得最优折衷方面的性能。

This work was supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART"), and by the SNS JU project 6G-GOALS under the EU's Horizon program Grant Agreement No 101139232.



图 1: 边缘设备(TX)根据最优编码器-锚点集对 (E^* , \mathbb{A}^*) 对数据进行编码。RX 配备了在共享锚点集上训练的相对解码器,以解释传输的相对表示。

II. 相对表示动机

SemCom 旨在交换数据所传达的意义,而不仅仅是 原始数据本身。为了实现这一目标,发射器(TX)和接 收器(RX)通常采用深度神经网络来编码和解码语义 信息,传输绝对表示(即原始的 DNN 编码器输出)。在 此场景中,由于随机因素或设计选择的不同,不同的编 码器在语义上对同一信息进行表征时可能存在差异,从 而产生语义噪声。这些不对齐会在解码器端引入误解, 干扰整体通信目标。

为了解决这个问题,我们在 SemCom 框架中引入了 RelRep,使零样本拼接成为可能——即在不同编码器 之间有效沟通的能力,而无需重新训练。有了 RelRep, TX 可以在飞行中切换编码器, RX 仍然能够解读消息, 而不必为每个新编码器重新训练自己,如图 1所示。接 下来将解释 RelRep 对于 SemCom 的原理。

相对表示。给定一个训练数据集 X = { x_i, y_i }_i, 编码函数 $E(\cdot)$ (例如, DNN) 会将每个样本 $x_i \in X$ 映射到其 绝对表示 $e_{x_i} := E(x_i) \in \mathbb{R}^d$ 。X 的一个子集被定义 为锚点集, A = { $a_1, \ldots, a_{|A|}$ }。一个通用的相似性函数 $sim : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 被用于计算任意两个绝对表示之 间的相似性。两个这样的表示之间的相似性分数 r 由 $r = sim(e_{x_i}, e_{x_j})$ 给出。然后, 样本 $x_i \in X$ 的相对表示 法定义为:

$$\boldsymbol{r}_{\boldsymbol{x}_i}(E,\mathbb{A}) := (sim(\boldsymbol{e}_{\boldsymbol{x}_i}, \boldsymbol{e}_{\boldsymbol{a}_1}), \dots, sim(\boldsymbol{e}_{\boldsymbol{x}_i}, \boldsymbol{e}_{\boldsymbol{a}_{|\mathbb{A}|}})) \quad (1)$$

Timm model name	# Parameters	FLOPs
mobilenetv3_small_100	1.5M	$111.98~{\rm M}$
mobilenetv3_large_100	4.2M	$435.36~{\rm M}$
rexnet_100	$3.5\mathrm{M}$	$799.28\mathrm{M}$
vit_small_patch16_224	21.6M	$8.5~\mathrm{G}$
vit_base_patch16_224	$85.7 \mathrm{M}$	$33.72~\mathrm{G}$
vit_base_resnet 50_384	98.1M	$99.04~\mathrm{G}$

表 I: 图 2中使用的 Timm 模型。

我们将 $D_{\mathbb{A}}: \mathbb{R}^{|\mathbb{A}|} \to \mathbb{R}^{\tau}$ 定义为相对解码器函数。它专 用于下游任务,使其能够解释相对表示中所表达的关 系,并准确提取意义。

语义对齐。RelRep 有效地通过关注数据关系(1)而



图 2: 零样本拼接使用不同预训练编码器和变化锚集大 小在 CIFAR-10 分类任务上的性能。虚线表示端到端 性能。

不是绝对表示来缓解语义不匹配的影响。这种方法促 进了语义等效的潜在空间,并允许基于使用的锚点数 量进行动态语义压缩。相对解码器解释这些关系,即 使在边缘动态切换编码器的情况下也能实现无需重新 训练的零样本拼接。这种灵活性在图 2中得到展示, 该图显示了跨不同编码函数和锚集基数的模型拼接性 能。值得注意的是,在绝对情况下,由于维度不匹配 或错位,零样本拼接经常失败,导致解码器之间行为 不可预测。在图 2中,我们使用了预训练的编码器(通 过从 ImageNet 上训练的模型中移除最终分类器获得 - 请参阅表 I获取完整列表)。相对解码器使用冻结的 mobilenetv3_large_100 编码器在 CIFAR-10 数据集上 进行训练。mobilenetv3_large_100的性能因此是端到 端的,而其他编码器的结果展示了零样本拼接能力。在 RelRep 设计的发展中,我们遵循了 [14] 中提出的方法。 我们将余弦相似性作为相似函数,基于数据表示之间的 角度在整个不同的编码功能下保持一致的假设。锚定点 集是通过对训练数据集中的样本进行均匀随机选择生 成的。这种方法确保了在实现简易性和性能结果稳健性 之间取得平衡。未来的工作可能会探索优化 RelRep 设 计以提高零样本拼接场景下的性能。通过其动态语义压 缩和编码器无关性, RelRep 所提供的灵活性是在资源 受限环境中的一大优势,降低了边缘设备的复杂度和部 署开销。因此,语义理解的任务转移到了接收端。这些 优势促使我们在下一节中提出的资源优化策略,旨在实 现节能、目标导向的语义通信。

III. 动态资源分配

在本节中,参考图 1中的一般通信方案,我们设计 了一种面向目标的包含推理任务的通信中的计算(例 如,CPU时钟频率)、通信(例如,数据速率)和学习 (例如,锚点数量、编码器)资源的动态分配方案,同时 促进对传输信息的语义理解。我们假设在发送方 TX 侧 分别有预训练编码器集合 *E* 和锚集集合 *A*。接收方 RX 被假定知道 *A*,并且拥有一组相对解码器(每个锚集 对应一个)以保证不同预训练编码器之间的语义对齐。 这些相对解码器假设使用从任意编码器生成的 RelReps 进行训练,该编码器可能属于或不属于 *E*。在这种设置 下,诸如信道质量、数据特性以及设备建模信息的方式 (即语义语言)的变化可能会波动。因此,所提出的动态 方法旨在通过在通信中的语义完整性方面工作来实现 功耗、延迟和推理准确性之间的最优权衡。编码器和锚 集的选择直接影响到发送方 TX 可实现的压缩效果和计 算需求:复杂模型可能提取更多有意义的信息但需要更高的计算频率,消耗更多的电力;同样地,更大的锚集 允许更细微的 RelReps 但也增加了需要传输的数据量。

A. 系统模型

我们假设一个简化的 RX,其解码延迟可以忽略不 计。假定发送端 TX 处有完美的信道状态信息 (CSI) 可 用。我们采用基于时隙的模型,在该模型中系统能够在 每个时隙内自适应和动态地优化资源分配。接下来,我 们将详细描述我们的系统模型。

电力消耗。我们考虑 TX 的两个电源消耗来源:本地 计算和传输。在时间 t, TX 用于本地计算的功率为 $p_t^c = \kappa (f_t)^3$,其中 κ 是处理器的有效开关电容, f_t 是 CPU 时钟频率 (Hz) [17]。在时间 t 用于传输的功率由 香农公式计算为

$$p_t^u = \frac{N_0}{|h_t|^2} \left[\exp\left(\frac{R_t}{B} \ln 2\right) - 1 \right],$$

其中, B 是带宽, $|h_t|^2$ 是上行信道功率增益, N_0 是噪 声功率谱密度, R_t 是数据速率。因此, 总能耗由

$$p_t = p_t^u + p_t^c. (2)$$

延迟。给出 每个时隙的总体延迟有两个主要的延迟来 源:本地处理时间和上行通信时间。编码器和锚点集的 选择直接对本地处理时间(由于编码器复杂度)以及上 行通信时间(由于由特定锚点集引起的要传输的相关表 示大小)产生影响。本地处理时间可以通过 $L_t^c = \frac{N_{E_t,A_t}}{f_t}$, 计算得出,其中 N_{E_t,A_t} 是编码消息所需的 CPU 周期 数,并且取决于所选的编码器 $E_t \in \mathcal{E}$ 和时间 t 选择的 锚定集 $A_t \in A$ 。上行通信时间可以通过以下方式计算 $L_t^u = \frac{n_{A_t} \cdot q}{R_t}$ 其中 $n_{A_t} \cdot q$ 是要传输的位数,它取决于相对 表示的大小,即锚集 A_t 的基数。因此,总延迟由下式 给出

$$L_t = L_t^u + L_t^c. aga{3}$$

学习准确性。我们假设 RX 被提供了一个验证集 X^{VAL}。 一个任务相关的函数 G_{E,A} 被引入来衡量学习性能,给 定一对编码器和锚定点集 E,A。作为一个示例,让我们 考虑一个分类任务,其验证准确率可以用来估计学习性 能为

$$G_{E,\mathbb{A}} = \frac{1}{|\mathbb{X}^{\text{VAL}}|} \sum_{x,y \in \mathbb{X}^{\text{VAL}}} \mathbb{I}(\widehat{y}_{E,\mathbb{A}}(\boldsymbol{x}) = y), \quad (4)$$

其中 $\hat{y}_{E,\mathbb{A}}(\boldsymbol{x})$ 是样本 \boldsymbol{x} 的预测结果,如果分别选择编码 器和锚定点集为 $E \in \mathcal{E}$ 和 $\mathbb{A} \in \mathcal{A}$ 。示例展示了不同编 码器和锚定集如何影响公式 (4) 中的性能指标,如图 2 所示。

B. 问题表述

我们现在可以制定面向目标通信的动态资源分配 问题。目标是找到最优策略,以便在每个t分配上行数 据速率 R_t 、设备 f_t 的 CPU 周期、编码器 E_t 和锚集 A_t , 以最小化(2)中的长期平均系统功耗,并且满足在(4) 中的平均推理准确率约束和在(3)中的平均延迟约束。 然后,动态资源分配问题可以表述为:

$$\min_{R_{t},f_{t},E_{t},\mathbb{A}_{t}} \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{p_{\tau}\}$$
subject to
$$(a) \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{\mathbf{1}_{[L_{\tau} \ge L^{\mathrm{IST}}]}\} \le p^{\mathrm{IST}};$$

$$(b) \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{L_{\tau}\} \le \bar{L}; \quad (c) \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{G_{\tau}\} \ge \bar{G};$$

$$(d) R^{\mathrm{MIN}} \le R_{t} \le R^{\mathrm{MAX}}; \quad (e) f_{t}^{\mathrm{MIN}} < f_{t} \le f_{t}^{\mathrm{MAX}};$$

$$(f) E_{t} \in \mathcal{E}; \quad (g) \mathbb{A}_{t} \in \mathcal{A}.$$

$$(5)$$

期望是在系统中与相关随机变量有关的,即信道条件和 发送端可用的计算频率。 $G_{\tau} \ge G_{E_t,A_t}$ 在(4)中的简写 表示。因此,我们隐含假设在验证集上评估的准确性是 推理(测试)准确性的良好替代。如果数据分布没有变 化,这是一个标准且合理的假设。问题的约束条件(5) 具有以下含义:(a)确保瞬时延迟超过某个值 L^{IST} 的概 率不超过阈值 p^{IST} ;(b)确保平均延迟不超过阈值 \bar{L} ; (c)确保系统的平均精度在某个阈值 \bar{G} 之下不会降低。 同时施加约束条件(a)和(b)有助于更好地控制延迟的 统计分布。最后,(d)-(g)中的约束对资源变量 R_t, f_t 施 加了瞬时界限(例如最小速率和最大 CPU 频率),并规 定 E_t 和 A_t 只能从有限集合 \mathcal{E} 和 A 中取值,分别对应 编码器和锚定集。

C. 通过随机优化的算法解决方案

我们现在引入一个动态算法框架来解决长期优化问题(5),通过随机 Lyapunov 优化工具将其重新表述

为稳定性问题 [16]。首先,我们定义虚拟队列 Z_t, Q_t 和 Y_t 分别来强制执行约束条件 (a)-(b)-(c),如

$$Y_{t+1} = \max\{0, Y_t + \epsilon_y(\mathbf{1}_{[L_{\tau} \ge L^{\text{IST}}]} - p^{IST})\}$$
(6)

$$Z_{t+1} = \max\{0, Z_t + \epsilon_z (L_t - \bar{L})\}$$
(7)

$$Q_{t+1} = \max\{0, Q_t + \epsilon_q(\bar{G} - G_t)\}$$
(8)

其中 ϵ_z, ϵ_q 和 ϵ_y 是用于控制算法收敛速度的正步长。此时,我们定义 Lyapunov 函数为 $\mathcal{U}_t = \mathcal{U}(\Phi_t) = \frac{1}{2}(Z_t^2 + Q_t^2 + Y_t^2)$,其中 $\Phi_t = [Z_t^2, Q_t^2, Y_t^2]$ 。然后,漂移加惩罚 函数为 $\Delta_t^p = \mathbb{E}\{\mathcal{U}_{t+1} - \mathcal{U}_t + V \cdot p_t^{TOT} | \Phi_t \}$ 。现在,我们 希望最小化漂移加惩罚的上界 [16]:

$$\Delta_t^p \le \xi + \mathbb{E} \Big\{ Z_t (L_t - \bar{L}) + Q_t (\bar{G} - G_t) + Y_t (L_t + 1 - p^{IST}) + V \cdot p_t^{TOT} | \Phi_t \Big\}$$
(9)

其中 ξ 是一个正有限常数,通过类似 [18] 的过程获得。 最后,使用随机逼近论点 [16],我们优化了漂移加惩罚 的上界,去除了期望,从而得到了以下每时隙问题(省 略所有常数项):

$$\min_{R_t, f_t, E_t, \mathbb{A}_t} \quad (Z_t + Y_t) L_t - Q_t G_t + V \cdot p_t^{TOT}$$
subject to $(f) \ E_t \in \mathcal{E}; \quad (g) \ \mathbb{A}_t \in \mathcal{A};$
 $(d) \ R^{MIN} \le R_t \le R^{MAX}; \quad (e) \ f_t^{MIN} < f_t \le f_t^{MAX}.$

$$(10)$$

现在,由于 (d)-(e)-(f)-(g),(10) 是一个混合整数非线性 优化问题,可能非常复杂难以求解。然而,对于任意给 定的 A_t 和 E_t ,问题 (10) 成为在 R_t 和 f_t 中的凸问题。 因此,我们可以通过施加 KKT 条件 [19] 得到最优上行 速率 R_t^* 和 CPU 频率 f_t^* ,得到

$$R_{t}^{*} = \left[\frac{2B}{\ln 2}W\left(\frac{1}{2}\sqrt{\frac{(Z_{t}+Y_{t})n_{\mathbb{A}_{t}}q|h_{t}|^{2}\ln 2}{BVN_{0}}}\right)\right]_{R^{MAX}}^{R^{MAX}}$$
(11)

$$f_t^* = \left[\sqrt[4]{\frac{(Z_t + Y_t)N_t}{3V\kappa}}\right]_{f_t^{MIN}}^{f_t^{MAX}}$$
(12)

此时,为了确定最优编码器和锚集,算法需要评估每个可能组合的目标函数,将通信和计算参数设置为(11)-(12)中的值。这意味着系统执行了 |*E*|·|*A*| 次独立函数 评估,其中 |*E*|和 |*A*| 分别是编码器和锚点集的基数。在 这种情况下,单个函数评估相关的计算成本通常可以忽

Algorithm 1 资源优化对于 SemCom

输入: 样本 x_t 输出: $E_t^*, \mathbb{A}_t^*, R_t^*, f_t^*$ for $t \ge 0$ do (RX) 更新虚拟队列 Z_t, Q_t, Y_t ; (RX) 计算 f_t^* 和 $R_t^*,$ 如同 (11) 和 (12) 中所示; (RX) 设置 $E_t^*, \mathbb{A}_t^* = \underset{E_t \in \mathcal{E}, \mathbb{A}_t \in \mathcal{A}}{\arg \min \Psi(E_t, \mathbb{A}_t, R_t^*, f_t^*)};$ (TX) 计算并传输 $r_{x_t}(E_t^*, \mathbb{A}_t^*)$; end for



图 3: 长期平均功率与 \overline{L} 的对比,对于不同的 \overline{G} 。

略不计,但可以采用贪婪方法进一步减轻负担。算法概述如 Alg.1所示,其中我们将问题 (10)中的目标表示为 $\Psi(R_t, f_t, E_t, \mathbb{A}_t)$ 。

D. 数值结果

为了评估我们动态资源分配框架的性能,我们在改 变平均延迟 (\overline{L}) 和平均准确性 (\overline{G}) 在 (5) 中的约束条 件时运行算法1。这使我们能够探索功耗、延迟与推理 准确性的内在权衡,突出我们方法的灵活性。如第 II节 所述,由于在零样本拼接设置中绝对情况的行为不可预 测,我们仅关注相对表示。我们的模拟设置与图 2中的 初步研究一致,采用了一组预训练的编码器 \mathcal{E} (见表I) 和锚点集 A。我们在5个随机种子上进行优化,根据以 下 \bar{L} 值[0.03, 0.04, 0.05, 0.06](秒)和 \bar{G} 值[0.7, 0.8, 0.9]。 该信道考虑了路径损耗传播,消息量化(q)设置为32 位,信道带宽为1MHz。为了确保及时响应,我们定义 了一个最大可容忍延迟 $L^{IST} = \overline{L} + 7.5 \times 10^{-3} s$,并设 有一个相应的概率阈值 $p^{IST} = 0.3$ 。然后,在图 3中,我 们展示了平均功率与不同精度约束 G 下的平均延迟约 束 \overline{L} 之间的行为。如图3所示, RelReps 使得所有系统 配置的有效语义信息交换成为可能;正如预期的那样,

更严格的准确性约束需要在给定延迟限制下选择更多 计算成本的选择。最后,图 4说明了特定系统配置(即 $\bar{L} = 0.04 \ s \ \pi \bar{G} = 0.8$)的瞬时行为,顶部显示的是延 迟,底部显示的是准确性,展示了算法1保证长期约束 (5)的能力。

IV. 结论

本文介绍了一个面向目标的和 SemCom 的新框架, 解决了动态环境中语义不匹配带来的挑战。我们的主要 贡献在于使用相对表示实现跨多种设备的稳健通信以 及一种面向目标任务的动态资源优化策略。该策略调整 通信、计算和学习资源以实现节能且低延迟推理,这对 边缘应用至关重要。数值结果验证了本框架在能耗、时 延与推理可靠性之间取得有利平衡的有效性。未来的发 展包括在 RelReps 中选择锚点及相似度函数的最佳方 案,旨在提升面向目标的语义通信性能。



图 4: 延迟和准确性对比 t、 $\bar{L} = 0.04 \ s$ 和 $\bar{G} = 0.8$ 。

参考文献

- C. E. Shannon, "A mathematical theory of communication," The Bell system technical journal, vol. 27, no. 3, pp. 379–423, 1948.
- [2] E. C. Strinati and S. Barbarossa, "6G networks: Beyond shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.
- [3] B. Güler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Transactions on Cognitive Communications and Net*working, vol. 4, no. 4, pp. 787–802, 2018.
- [4] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, "Towards a theory of semantic communication," in 2011 IEEE Network Science Workshop. IEEE, 2011, pp. 110–117.
- [5] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2022.

- [6] E. C. Strinati, P. Di Lorenzo *et al.*, "Goal-oriented and semantic communication in 6G AI-native networks: The 6G-goals approach," *arXiv preprint arXiv:2402.07573*, 2024.
- [7] P. Di Lorenzo, M. Merluzzi, F. Binucci, C. Battiloro, P. Banelli, E. C. Strinati, and S. Barbarossa, "Goal-oriented communications for the IoT: System design and adaptive resource optimization," *IEEE Internet of Things Magazine*, vol. 6, no. 4, pp. 26–32, 2023.
- [8] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 96–102, 2021.
- [9] P. A. Stavrou and M. Kountouris, "The role of fidelity in goaloriented semantic communication: A rate distortion approach," *IEEE Transactions on Communications*, 2023.
- [10] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, "Adaptive resource optimization for edge inference with goal-oriented communications," *EURASIP J. on Advances in Signal Proc.*, no. 1, p. 123, 2022.
- [11] C. K. Thomas, E. C. Strinati, and W. Saad, "Reasoning with the theory of mind for pragmatic semantic communication," arXiv preprint arXiv:2311.18224, 2023.
- [12] S. Barbarossa, D. Comminiello, E. Grassucci, F. Pezone, S. Sardellitti, and P. Di Lorenzo, "Semantic communications based on adaptive generative models and information bottleneck," *IEEE Communications Magazine*, vol. 61, no. 11, pp. 36–41, 2023.
- [13] M. Sana and E. C. Strinati, "Semantic channel equalizer: Modelling language mismatch in multi-user semantic communications," arXiv preprint arXiv:2308.03789, 2023.
- [14] L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodola, "Relative representations enable zero-shot latent space communication," arXiv preprint arXiv:2209.15430, 2022.
- [15] A. Norelli, M. Fumero, V. Maiorca, L. Moschella, E. Rodola, and F. Locatello, "Asif: Coupled data turns unimodal models to multimodal without training," *Proc. of NeurIPS*, vol. 36, 2024.
- [16] M. J. Neely, Stochastic Network Optimization with Application to Communication and Queueing Systems. Morgan & Claypool Publ., 2010.
- [17] T. D. Burd and R. W. Brodersen, "Processor design for portable systems," Journal of VLSI signal processing systems for signal, image and video technology, vol. 13, no. 2-3, pp. 203–221, 1996.
- [18] M. Merluzzi, P. Di Lorenzo, S. Barbarossa, and V. Frascolla, "Dynamic computation offloading in multi-access edge computing via ultra-reliable and low-latency communications," *IEEE Transactions* on Signal and Information Processing over Networks, vol. 6, pp. 342–356, 2020.
- [19] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.