波束成形

A Deep Learning Framework for Hybrid Beamforming Without Instantaneous CSI Feedback

- 为了降低复杂性，可以利用通道统计信息，这样只需要不频繁地更新通道信息。
- 一种基本方法是从预定义的码本中选择射频预编码器和合成器的列，其中包括接收/发射路径角的阵列响应或者DFT码本，然而在毫米波信道中，接收路径角的确定非常困难，阵列响应难以获得。
- 统计混合波束形成(SHB)结构，利用信道统计设计波束形成器，通常使用二阶统计量，即信道协方差矩阵(CCMs)。通过 CCM 采集，基站(BS)只知道信道统计信息，信道信息反馈很少，但是没有即时的 CSI 反馈。因此可以实现更低的反馈开销。
- 信道未知下的波束成形设计：

  -> 基站端在发送导频信号时只开启一个射频链路，只形成一个波束，波束成形矩阵$\overline{\mathbf{f}}_u \in \mathbb{C}^{N_{\mathrm{T}}}, u=[1,\ldots,M_T]$一般从DFT码本$\mathbb{C}^{M_T\times M_T}$中选择，其中$M_T$为DFT码本大小，码本的每列截断为$N_t$以适应天线数目。

  -> 接收端开启全部射频链路来接收导频信号，每个射频链路的$\overline{\mathbf{w}}_u \in \mathbb{C}^{N_{\mathrm{R}}}, u=[1,\ldots,M_R]$同样从从DFT码本$\mathbb{C}^{M_R\times M_R}$中选择，码本的每列截断为$N_r$以适应天线数目。由于接收端射频链路数目小于天线数，每次只能从码本中选取N_{RF}^R个码字构成$W \in C^{N_R\times N_{RF}^R}$，用于接收导频，通过多次选取不同码字构建$W$实现从全部DFT码本定义的码本反向接收导频信号。

  -> 在接收端使用全部DFT码字接收导频信号后，发送端更换$\overline{\mathbf{f}}_u$,以实现以DFT码本定义的全部方向发送导频信号。总计可以获得$M_{\mathrm{T}}\left\lceil\frac{M_{\mathrm{R}}}{N_{\mathrm{R}}^{\mathrm{RF}}}\right\rceil$个数据。发送端只开启一个射频链路的导频传输过程可以捕获毫米波信道中的主要路径，虽然在发射端同时激活多个不同波束的射频链可以加快导频传输过程，但它不能捕获主要路径，导致信道估计性能较差。

$$\overline{\mathbf{Y}}=\overline{\mathbf{W}}^{\mathrm{H}}\mathrm{H}\overline{\mathbf{F}}\overline{\mathbf{S}}+\tilde{\mathbf{N}} \tag{1}$$

  其中 $\overline{\mathbf{F}}=\left[\overline{\mathbf{f}}_1,\overline{\mathbf{f}}_2,\ldots,\overline{\mathbf{f}}_{M_{\mathrm{T}}}\right]\in\mathbb{C}^{N_{\mathrm{T}}\times M_{\mathrm{T}}}$，$\overline{\mathbf{W}}=[\overline{\mathbf{w}}_1,\overline{\mathbf{w}}_2,\ldots,\overline{\mathbf{w}}_{N_{\mathrm{R}}}]\in\mathbb{C}^{N_{\mathrm{R}}\times M_{\mathrm{R}}}$，$\overline{\mathbf{S}}=\operatorname{diag}\{\bar{s}_1,\ldots,\bar{s}_{M_{\mathrm{T}}}\}$，$\tilde{\mathbf{N}}=\overline{\mathbf{W}}^{\mathrm{H}}\overline{\mathbf{N}}$，$\overline{\mathbf{S}}$可以设为单位阵即发送全1的导频信号，最终得到原始信道估计Y为：

$$\begin{gathered}
Y=\mathbf{T}_{\mathrm{T}}\overline{\mathbf{Y}}\mathbf{T}_{\mathrm{R}}, \\
\mathrm{T}_{\mathrm{T}}=\begin{cases}\overline{\mathbf{W}}, & M_{\mathrm{R}}<N_{\mathrm{R}}\\ \left(\overline{\mathbf{W}}\overline{\mathbf{W}}^{\mathrm{H}}\right)^{-1}\overline{\mathbf{W}}, & M_{\mathrm{R}}\geq N_{\mathrm{R}}\end{cases} \\
\mathrm{T}_{\mathrm{R}}=\begin{cases}\overline{\mathbf{F}}^{\mathrm{H}}, & M_{\mathrm{T}}<N_{\mathrm{T}}\\ \overline{\mathbf{F}}^{\mathrm{H}}\left(\overline{\mathbf{F}}\overline{\mathbf{F}}^{\mathrm{H}}\right)^{-1}, & M_{\mathrm{T}}\geq N_{\mathrm{T}}.\end{cases}
\end{gathered} \tag{2}$$

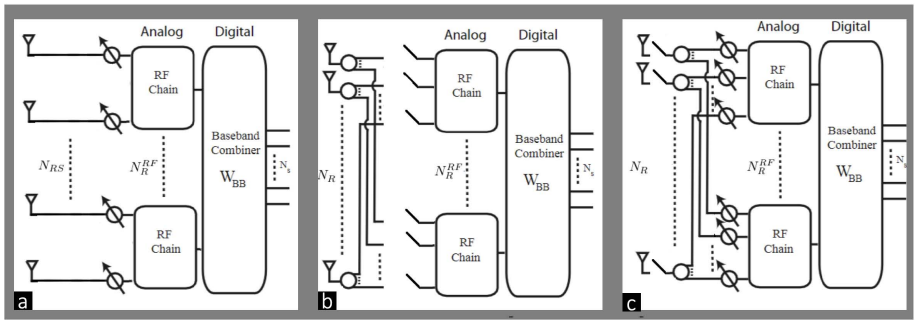
  -> $Y\in\mathbb{C}^{Nr\times N_t}$拆解为实部和虚部两通道数据作为神经网络的输入实现波束成形/信道估计设计。

- 神经网络的输入：实部、虚部、绝对值、复数的角度。
- 通过在原始数据上加高信噪比噪声的仿真扩充数据。

- 在线更新时只更新网络的高层(比如最后的全连接层)，因为它们更依赖于环境。下层(即卷积层)保持完好或冻结，因为它们通常与问题相关而与环境无关，卷积层擅长从输入中提取新特征，全连接层在将输入数据映射到输出方面更强大。此外更新只在环境剧烈变化时进行。综上在线更新发生不频繁，每次更新的数据集较小，待更新的网络参数较少（高层网络参数），所以复杂度较低。
- 因为深层网络缺乏精确度，本质上是有偏的估计器。如果在训练数据包含太多噪声，神经网络将难以学习输入数据的特征，从而难以建立输入数据到输出的映射关系，所以训练数据应该具有较少的噪声以提供良好的精度。
- 角度失配问题：网络在训练阶段使用的是无失配信道:H->NN->(F,W)->SE(F,W,H);测试时由于信道估计的误差使用失配信道：H+delta->NN->(F,W)->SE(F,W,H);传统算法同样受到信道估计误差的影响：H+delta->Algo->(F,W)->SE(F,W,H)。可以通过添加更多具有不同角度信息的通道实现来丰富训练数据，可以实现神经网咯更强的稳健性。
- 但环境的射线集群数、集群内部的角度弥散变化较大时，可以使用在线学习策略，在环境变换较大时以新的信道数据训练神经网络以使用环境的变化。

Joint Antenna Selection and Hybrid Beamformer Design Using Unquantized and Quantized Deep Learning Networks

- 子阵列单元的优化选择降低了模拟移相器和低噪声放大器(LNA)的功耗。
- 天线选择架构：
  1. 部分链接架构可以视为一种天线选择架构，只是天线子阵固定无法改变。
  2. 开关网络架构，每个天线可以只定义其与每个射频链路的连接情况，但未使用移相器。
  3. 带移相器的开关网络架构，其架构大体与全连接相似，开关接在天线的输入端，控制天线是否激活。$\mathbf{H}_{\text{sub}} = \mathbf{QH}, \mathbf{Q} \in C^{N_{RS} \times N_R}$，其中Q为天线选择矩阵，其每行只有一个元素为1其余为0，该元素列号为被选择天线的下标。



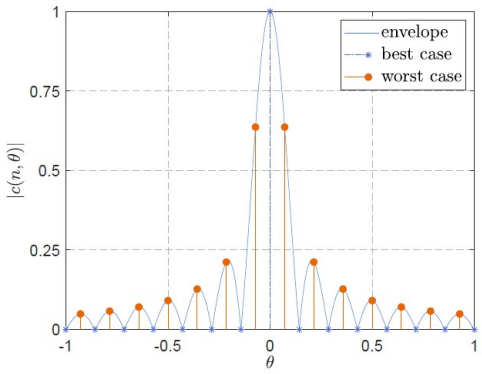
- 天线子阵共有选择$S = C_{N_R}^{N_{RS}}$种选择，通过遍历每种天线组合计算该子信道下全数字的频谱效率，选择全局最优的天线子阵。注意通过贪心搜索，接收端依次选择全部天线中的一个以单天线单流计算频谱效率，最终得最优的N_{RS}天线并不是全局最优的。
- 当N_R较大时子阵选择的计算复杂度高，可以使用BAB策略，将S个子阵组合划分为多个互不相交的块，选择每块中最优的子阵，综合所有块可以得到全局最优解。
- 天线选择神经网络实现分类输入：$[real(H), img(H), abs(H)] \in R^{N_r \times N_T \times 3}$,输出：所有子阵各自为最优解的概率$P \in R^{S \times 1}$。混合波束成形网络实现回归输入：$[real(H_s), img(H_s), abs(H_s)] \in R^{N_{rs} \times N_T \times 3}$，输出：

$$\left[\text{vec}^T\{\angle \mathbf{F}_{RF}\}, \text{Re}\left\{\text{vec}^T\{\mathbf{F}_{BB}\}\right\}, \text{Im}\left\{\text{vec}^T\{\mathbf{F}_{BB}\}\right\} \text{vec}^T\{\angle \mathbf{W}_{RF}\}, \text{Re}\left\{\text{vec}^T\{\mathbf{W}_{BB}\}\right\}, \text{Im}\left\{\text{vec}^T\{\mathbf{W}_{BB}\}\right\}\right]^T \in R^{N_{RS}N_R^{RF} + 2N_S\left(N_T^{RF} + N_R^{RF}\right)}$$

- 数据生成共$N * N_c * N_n$个信道，其中N_c表示可选射线集群数集合的大小，N_n表示在在原始信道上加上噪声以生成的噪声信道个数，并重复上述过程N次，生成$N * N_c$个原始信道，在加上噪声后扩展到$N * N_c * N_n$。在N_n取值较大时容易收敛，比如$N_n = 200$。
- 训练数据中的噪声将会限制性能，因为如果输入数据受到太多噪声，则网络不能区分输入数据，无法建立输入与标签的映射关系，所以训练阶段和测试阶段对信道加的噪声不能过大。
- 在训练阶段输入多种环境参数下的信道比如不同的射线集群数、射线族内角度弥散，对训练数据加上高信噪比的噪声，以增强网络对不同环境的适应力，以及增强网络对不完美信道估计的鲁棒性。
- 网络压缩：通过减少表示卷积和完全连接层的每个权重所需的比特数来压缩原始网络，当网络参数至少为5bit数据时才能保证较好的性能。

Deep Learning for Beam Training in Millimeter Wave Massive MIMO Systems

- 用深度神经网络(DNN)来处理毫米波通信中信道功率泄漏的非线性和非单调特性。
- 由于毫米波信号传输是高度定向的，因此识别最强的信道路径非常重要，该路径通常是信道的视距(LOS)路径。基于毫米波天线阵列形成的波束，目标是找到与毫米波最强路径最匹配的发射和接收波束组合。
- 由有限数量的码字引起的信道功率泄漏。虽然可以使用预定义的码字来形成指向不同方向的不同波束，但是码本中的码字数量是有限的，这表明只能指向有限的方向，但是信道的AOA或AOD在角度空间中连续分布。波束可能不能与通道AOA或AOD精确对准，从而导致功率泄漏。



最佳情形下主瓣只有一个峰值，没有旁瓣，没有通道功率泄漏。最坏情况下主瓣中会出现两个相等的峰值，且每个副瓣中的峰值最高。大多数情况介于最好和最坏情况之间，即在主叶有两个高峰值，在副叶有几个短峰。

显然主瓣中的这两个峰值占据了信道的大部分功率，实际中可能选择对应于第二高峰值而不是最高峰值的码字，这导致波束训练失败并且降低了成功率。然而可实现的速度可能只会受到轻微的影响，因为这两个峰值的高度非常相似，或者至少在一个数量级上。所以可达速率来表征波束训练通常比成功率更有说服力。NA越小信道功率泄漏越大，对成功率的影响越大，但对可达速率的影响越小。

随着NA的增加，码字的数量越来越大，码本的分辨率也越来越高，从而导致较小的信道功率泄漏。

- 在初始测试中，仅测试发射和接收波束组合的一部分。然后根据初始测试的结果，确定了功率最大的两个相邻行和两个相邻列，以获得最佳的波束组合，围绕该波束组合的几个波束组合需要进行额外的测试。
- 码本中的码字数量是有限的，这表明我们只能指向有限的方向。但是信道的AOA或AOD在角度空间中连续分布。波束可能不能与通道AOA或AOD精确对准，从而导致功率泄漏。

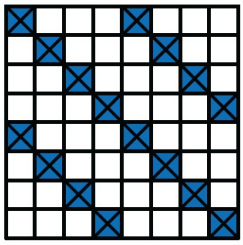
- 单用户单流下优化目标：

$$\max_{\substack{\boldsymbol{f}\in\mathcal{F}_c\\ \boldsymbol{w}\in\mathcal{W}_c}} \log_2\left(1+\frac{P\left|\boldsymbol{w}^H\boldsymbol{H}\boldsymbol{f}\right|^2}{\sigma^2}\right) \tag{3}$$

其中$\mathcal{F}_c,\mathcal{W}_c$一般取DFT码本。由于噪声的存在以及信道未知，在实际中无法直接优化问题(6)，通过收发导频的方式优化问题为：

$$\max_{\substack{\boldsymbol{f}\in\mathcal{F}_c\\ w\in\mathcal{W}_c}} |\sqrt{P}\boldsymbol{w}^H\boldsymbol{H}\boldsymbol{f}x+\boldsymbol{w}^H\boldsymbol{\eta}|^2. \tag{4}$$

- 使用f和w的组合对y的每一次测量被称为单束训练测试。问题(3)和(4)的解在大多数情况下是一致的，也可能由于信道噪声而不同。如果相同，称束训练成功；否则认为束训练失败。成功率被定义为成功的射束训练测试次数与射束训练测试总数的比率。
- 波束扫描:通过穷举f和w的所有组合来求解问题(4)以找到最佳组合，称为波束扫描。其性能最优，但复杂度较高。
- 分层码本：通常由在码本上层覆盖广角的少量低分辨率码字和在码本下层提供高定向增益的大量高分辨率码字组成。由上层码字形成的每个波束通常覆盖由下层码字形成的M个更窄的波束，即，根码字覆盖M个叶码字，其中M称为分级因子，通常设置M=2，因为天线的数量通常是2的整数阶。
- 收发双方的DFT码本两两组合构成可选波束对集合B，通过在集合中等间隔取码本组合B_s：



在该波束对下收发导频信号$[\boldsymbol{y}_T]_i=\sqrt{P}\boldsymbol{w}_{n_i}^H\boldsymbol{H}\boldsymbol{f}_{m_i}x+\boldsymbol{w}_{n_i}^H\boldsymbol{\eta}$，将得到的导频信号作为神经网络的输入。输出为B中各个波束对为最佳波束对的概率P，数据标签由波束扫描产生，将问题转换为多分类问题。

- 为增强算法性能，进行额外的训练：

  -> 通对B_s中的波束对进行波束训练，得到$S_1=\{|y|^2\}$,作为神经网络的输入，得到B中波束对为最佳波束对的概率P。

  -> 将P中与B_s对应的已经进行训练的波束对的概率设为0，防止对波束对重复训练。

  -> 从P中选取概率最大的K个波束对构成集合B_k，B_k的构架利用到了神经和网络的输出概率优于随机选择码本进行波束训练，利用公式(4)对B_k进行额外的波束训练，得到接收导频功率$S_2=\{|y|^2\}$。

  -> 找到$\{S_1,S2\}$中导频功率最大的波束对最为输出。

Deep CNN-Based Channel Estimation for mmWave Massive MIMO Systems

- 传统的信道估计方法在更实际、更复杂的信道模型中往往表现不佳，而且复杂度也很高。相比之下，深卷积神经网络(CNN)能够更好地从海量数据中提取信道矩阵的内在特征，并为利用高效的并行计算方法以更低的复杂度更精确地估计信道提供了可能。
- 信道估计设计：

-> 基站端在发送导频信号时只开启一个射频链路，只形成一个波束，波束成形矩阵$\overline{\mathbf{f}}_u \in \mathbb{C}^{N_\mathrm{T}}, u=[1,\ldots,M_T]$一般从DFT码本$\mathbb{C}^{M_T\times M_T}$中选择，其中$M_T$为DFT码本大小，码本的每列截断为$N_t$以适应天线数目，简单起见可以设置为$M_T=N_T$。

-> 接收端开启全部射频链路来接收导频信号，每个射频链路的$\overline{\mathbf{w}}_u \in \mathbb{C}^{N_\mathrm{R}}, u=[1,\ldots,M_R]$同样从从DFT码本$\mathbb{C}^{M_R\times M_R}$中选择，码本的每列截断为$N_r$以适应天线数目。由于接收端射频链路数目小于天线数，每次只能从码本中选取N_{RF}^R个码字构成$W\in C^{N_R\times N_{RF}^R}$，用于接收导频，通过多次选取不同码字构建$W$实现从全部DFT码本定义的码本反向接收导频信号。

-> 在接收端使用全部DFT码字接收导频信号后，发送端更换$\overline{\mathbf{f}}_u$,以实现以DFT码本定义的全部方向发送导频信号。总计可以获得$M_\mathrm{T}\left\lceil\frac{M_\mathrm{R}}{N_\mathrm{R}^\mathrm{RF}}\right\rceil$个数据。发送端只开启一个射频链路的导频传输过程可以捕获毫米波信道中的主要路径，虽然在发射端同时激活多个不同波束的射频链可以加快导频传输过程，但它不能捕获主要路径，导致信道估计性能较差。

$$\overline{\mathbf{Y}}=\overline{\mathbf{W}}^{\mathrm{H}}\mathrm{H}\overline{\mathbf{F}}\overline{\mathbf{S}}+\tilde{\mathbf{N}} \tag{5}$$

其中 $\overline{\mathbf{F}}=\left[\overline{\mathbf{f}}_1,\overline{\mathbf{f}}_2,\ldots,\overline{\mathbf{f}}_{M_\mathrm{T}}\right]\in\mathbb{C}^{N_\mathrm{T}\times M_\mathrm{T}}$，$\overline{\mathbf{W}}=[\overline{\mathbf{w}}_1,\overline{\mathbf{w}}_2,\ldots,\overline{\mathbf{w}}_{N_\mathrm{R}}]\in\mathbb{C}^{N_\mathrm{R}\times M_\mathrm{R}}$，$\overline{\mathbf{S}}=\mathrm{diag}\left\{\bar{s}_1,\ldots,\bar{s}_{M_\mathrm{T}}\right\}$，$\tilde{\mathbf{N}}=\overline{\mathbf{W}}^{\mathrm{H}}\overline{\mathbf{N}}$，$\overline{\mathbf{S}}$可以设为单位阵即发送全1的导频信号，最终得到原始信道估计Y为：

$$\begin{aligned}
Y&=\mathbf{T}_\mathrm{T}\overline{\mathbf{Y}}\mathbf{T}_\mathrm{R},\\
\mathrm{T}_\mathrm{T}&=\begin{cases}\overline{\mathrm{W}}, & M_\mathrm{R}<N_\mathrm{R}\\ \left(\overline{\mathbf{W}}\overline{\mathbf{W}}^{\mathrm{H}}\right)^{-1}\overline{\mathbf{W}}, & M_\mathrm{R}\geq N_\mathrm{R}\end{cases}\\
\mathrm{T}_\mathrm{R}&=\begin{cases}\overline{\mathbf{F}}^{\mathrm{H}}, & M_\mathrm{T}<N_\mathrm{T}\\ \overline{\mathbf{F}}^{\mathrm{H}}\left(\overline{\mathbf{F}}\overline{\mathbf{F}}^{\mathrm{H}}\right)^{-1}, & M_\mathrm{T}\geq N_\mathrm{T}.\end{cases}
\end{aligned} \tag{6}$$

-> $Y\in\mathbb{C}^{N_r\times N_t}$拆解为实部和虚部两通道数据作为神经网络的输入实现信道估计设计。

- 添加批归一化(BN)层以避免梯度扩散和过拟合。输出层使用双曲正切激活函数将输出映射到区间[−1，1]。采用SF-CNN进行信道去噪，将每层特征映射的大小设置为$N_R\times N_T$。使用卷积隐藏层来完全揭示信道的固有结构,采用尺寸非常小的多卷积滤波器，以较低的复杂度实现良好的信道估计性能。
- 离线训练的SF-CNN对大多数以前没有观察到的新信道统计数据非常稳健,当面对训练信道情形与测试情形不一致（如传播环境、多径数不一致）时仍有较好的性能。

Joint Deep Reinforcement Learning and Unfolding:Beam Selection and Precoding for mmWave Multiuser MIMO With Lens Arrays

- 框架包括一个基于深度强化学习的神经网络和一个深度展开的神经网络，分别用于优化波束选择和数字预编码矩阵。对于基于drl的神经网络，我们将波束选择问题表述为马可夫决策过程问题，并提出了一种双深度q网络算法来解决该问题。基站被认为是一个代理，其中的状态，行动和奖励功能是精心设计的。针对数字预编码矩阵的设计，提出了一种迭代加权最小均方误差算法诱导的深度展开神经网络。
- DLA 通常由两部分组成: 电磁透镜和匹配的天线阵列。它将传统的MIMO空间信道转换为具有角度依赖能量聚焦能力的波束空间信道。实际上由于波束空间信道的稀疏性，只能选择少量增益较大的波束。此外每个射频链选择一个单光束，从而所需的射频链数量可以大大减少。
- 波束选择矩阵$F\in Z^{N_t\times N_{RF}}$，每个RFChian只连接单个天线，每个天线最多连接一个RFChian（F每列只有一个元素为1，每行最多一个元素为1,$F^HF=I$）。每个RF链馈送单个波束，从N_t个波束中选择N_{RF}个波束来服务用户。

Fast Millimeter Wave Beam Alignment: a Phaseless and Low Cost AI-driven Method

- 首先基于稀疏波束分组准则将波束对齐问题转化为波束分组博弈问题。人工智能驱动的波束对准方法在第一阶段将波束均匀分组，在第二阶段采用第一次检测信息来搜索波束方向。
- 分层搜索在波束形成前存在低信噪比的问题，并且容易受到多径干扰的影响。此外继续搜索过程还需要每个判决的反馈，这增加了额外波束信息数据链路的通信负担。
- 首先接收器从准全向波束开始，发射器扫描波束空间以寻找最佳发射器波束。接收机保持准全向，而发射机扫描波束空间以识别最佳波束。其次接收机扫描波束空间以寻找最佳接收机波束，同时发射机使用之前扫描到的最佳波束。
- 采用离散傅里叶变换(DFT)码本来形成预编码矩阵。稀疏毫米波信道h可以写成稀疏形式h=Dx，其中D是DFT矩阵，x是一个K稀疏向量（元素值只有K个是非零的），其中非零条目对应于稀疏信道路径，将BA过程简化为选择列索引。$F_{RF}=D^{*}S$其中$S\in z^{N_T\times N_{RF}}$为列选择矩阵每列只有一个元素非0.
- 由于载波频率很高，在毫米波通信中不能忽略CFO。相位噪声和载波频率偏移导致随时间变化的测量的未知相移，对于鲁棒性波束对准，只考虑对接收信号的幅度估计。
- BA问题被简化为识别接收信号中具有最大幅度的元素。目标是设计一个测量矩阵A并开发一种快速有效的恢复算法来恢复稀疏向量的幅度|x|，形成压缩无相位测量y=|r|=|Ax+n|，其中r代表接收信号，n代表相位噪声。A是一个稀疏矩阵，A的每一行最多包含N_r^{ft}个非零元素。
- N天线，R射频链路，K稀疏度，N个DFT码字。第一阶段利用DL将码字均匀分到R个组中，每个组P=N/R个码字，将每个组分成三类，1：组中只有一个码字与某条路径对齐，2：组中没有码字与某条路径对齐，3：组中只有多个码字与某条路径对齐。每个组相当于射频链路的工作码字，情形 1 射频链路正常工作，情形 2 射频链路无法形成对齐的有效波束，情形 3 一个射频链路只能实现一个波束方向，无法实现多方向对齐。将 2 、 3 合并

EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

- 卷积神经网络（ConvNets）通常是在固定的资源预算下发展起来的，如果有更多的资源可用的话，则会扩大规模以获得更好的精度，比如可以提高网络深度(depth)、网络宽度(width)和输入图像分辨率(resolution)大小。但是通过人工去调整 depth, width, resolution 的放大或缩小的很困难的，在计算量受限时有放大哪个缩小哪个，这些都是很难去确定的，换句话说，这样的组合空间太大，人力无法穷举。基于上述背景，该论文提出了一种新的模型缩放方法，它使用一个简单而高效的复合系数来从depth, width, resolution 三个维度放大网络。
- 将整个卷积网络称为 N，它的第 i个卷积层可以表示为：$Y_i=F_i(X_i)$，X_i代表输入张量，Y_i代表输出张量。整个卷积网络由 k 个卷积层组成，可以表示为：$\mathcal{N}=\mathcal{F}_k\odot\ldots\odot\mathcal{F}_2\odot\mathcal{F}_1(X_1)=\bigodot_{j=1\ldots k}\mathcal{F}_j(X_1)$

  在实际中，通常会将多个结构相同的卷积层称为一个 stage，每个 stage 中的卷积层结构相同(除了第一层为降采样层)，以 stage 为单位可以将卷积网络 N 表示为：

  $$\mathcal{N}=\bigodot_{i=1\ldots s}\mathcal{F}_i^{L_i}\left(X_{\langle H_i,W_i,C_i\rangle}\right)$$

  $F_i^{L_i}$ 表示第 i 个 stage，它由卷积层 F_i从复 L_i次构成。通常的 ConvNet 设计主要关注寻找最佳的网络层 F_i，模型缩放尝试扩展网络深度（L_i，网络层数）、宽度（C_i通道数）或输入数据的分辨率（W_i, H_i），而不改变基线网络中预定义的 （指 kernel size 等每一个层内的参数，缩放只对 depth, width, resolution 进行组合调整，不对每一个层内具体的方式做改变）。优化目标就是在资源有限的情况下，要最大化 Accuracy，其中d, w, r为深度、宽度、分辨率缩放因子。

$$
\begin{aligned}
\max_{d,w,r} \quad & Accuracy\big(\mathcal{N}(d,w,r)\big) \\
s.t. \quad & \mathcal{N}(d,w,r) = \bigodot_{i=1\ldots s} \hat{\mathcal{F}}_i^{d\cdot\hat{L}_i}\big(X_{\langle r\cdot\hat{H}_i, r\cdot\hat{W}_i, w\cdot\hat{C}_i\rangle}\big) \\
& \text{Memory}(\mathcal{N}) \le \text{target\_memory} \\
& \text{FLOPS}(\mathcal{N}) \le \text{target\_flops}
\end{aligned}
$$

- 更宽的网络往往能够捕获更细粒度的特征，并且更容易训练，但是极宽但很浅的网络往往难以捕获更高级别的特征。当网络变得更宽且w更大时，准确率迅速饱和。

  更深入的ConvNet可以捕获更丰富、更复杂的特征，并很好地概括新任务。然而由于梯度消失的问题，更深层次的网络也更难训练。非常深的网络的精度收益会降低。

  有了更高分辨率的输入图像，ConvNet可能会捕获更细粒度的模式，但对于非常高的分辨率，精度增益会降低。

- 更大的网络具有更大的宽度、深度或分辨率，往往可以获得更高的精度，但精度增益在达到 80% 后会迅速饱和，这表明了只对单一维度进行扩张的局限性，模型扩张的各个维度之间并不是完全独立的，对于更大的分辨率图像，应该使用更深、更宽的网络，这就意味着需要平衡各个扩张维度，而不是在单一维度张扩张。实验表明在三个维度等比例的缩放可以取得较好的效果。

- α, β, γ是需要求解的一组参数，分别衡量着 depth, width 和 resolution 的比重，其中 β, γ在约束上会有平方，因为增加宽度或分辨率两倍，其计算量是增加四倍，但是增加深度两倍，其计算量只会增加两倍。

  Same 卷积：$(h_1, w_1, c_1) \to c_2 * (k_1, k_2, c_1) \to (h_1, w_1, c_2)$，需要 $h_1 * w_1 * k_1 * k_2 * c_1 * c_2$ 次乘法，$h_1 * w_1 * (k_1 * k_2 * c_1 - 1) * c_2$ 次加法，如果考虑偏置还需要 $h_1 * w_1 * c_2$ 次加法。参数个数：$w = k_1 * k_2 * c_1 * c_2, b = c_2$。

- 假设可用资源翻倍：ϕ决定了可用的新增资源，α, β, γ决定了新增资源在三个维度上的分配比例。

$$
\begin{aligned}
\text{depth: } & d = \alpha^\phi \\
\text{width: } & w = \beta^\phi \\
\text{resolution: } & r = \gamma^\phi \\
\text{s.t. } & \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\
& \alpha \ge 1, \beta \ge 1, \gamma \ge 1
\end{aligned}
$$

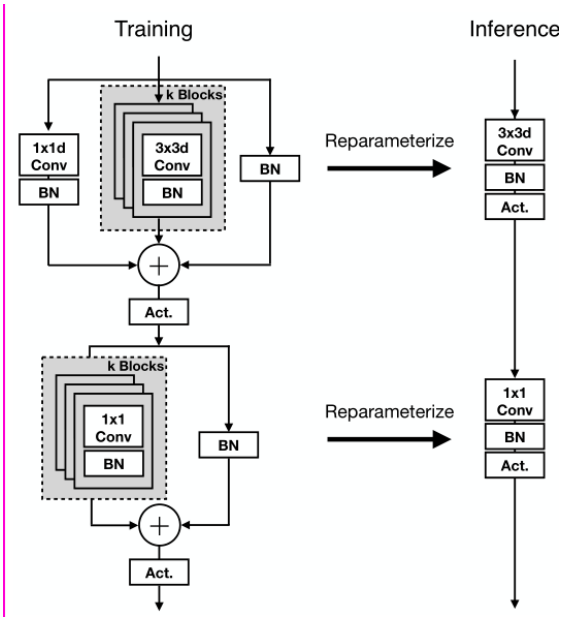
  求解方式：

  - 固定公式中的φ=1，假设可用资源翻倍，然后通过网格搜索（grid search）得出最优的α、β、γ，得出最基本的模型 EfficientNet-B0.
  - 固定α、β、γ的值，使用不同的φ，得到 EfficientNet-B1, ..., EfficientNet-B7。当φ=1 时，得出了一个最小的最优基础模型；增大φ时，相当于对基模型三个维度同时扩展，模型变大，性能也会提升，资源消耗也变大。上述操作将使得FLOPS变为之前的$(\alpha * \beta^2 * \gamma^2)^\phi$倍，限定$\alpha * \beta^2 * \gamma^2 = 2$，这样总的FLOPS约增加为原来的$2^\phi$倍。

An Improved One millisecond Mobile Backbone

- 高效率网络具有更强的实用价值，但学术界的研究往往聚焦于FLOPs或者参数量的降低，而这两者与推理效率之间并不存在严格的一致性。比如FLOPs并未考虑访存消耗与计算并行度，像无参操作(如跳过连接导致的Add、Concat等)会带来显著的访存消耗，导致更长推理耗时。

- 延迟的快慢与模型的参数量或者FLOPs的相关性较弱，在CPU端相关性更弱。
- 不同激活函数导致的延迟差异极大。当采用单分支结构时，模型具有更快的速度。
- 影响运行时性能的两个关键因素是内存访问成本和并行度。在多分支体系结构中，存储器访问成本显著增加，因为必须存储来自每个分支的结果以计算图中的下一个张量。强制同步的体系结构块，如Squze-Excite块中使用的全局池操作，也会由于同步成本而影响总体运行时间。
- 在训练中使用训练时间可重参数分支和正则化动态松弛，有助于缓解在训练小模型时遇到的优化瓶颈。
- 重参数:分离随机变量的不确定性，使得原先无法求导/梯度传播的中间节点可以求导。可以理解为用一个简单的网络结构去等效替代一个较复杂的结构，其优点是可以降低模型计算开销。



- 卷积层被分解为纵深和逐点的层。基本块构建在3x3深度卷积和1x1逐点卷积的MobileNet-V1块之上。
- 由于我们的模型在推理时没有多分支体系结构，因此不会产生前面部分讨论的数据移动成本。与竞争对手的多分支架构(如MobileNet-V2、EfficientNets等)相比，这使我们能够积极扩展模型参数，而不会产生显著的延迟成本。

Tricks in Deep Neural Networks

- 由于深度网络需要对大量的训练图像进行训练才能达到令人满意的性能，如果原始图像数据集包含的训练图像有限，则最好进行数据增强以提高性能。例如流行的水平翻转、随机裁剪和颜色抖动。可以尝试多个不同处理的组合。

Self-Supervised Learning via Maximum Entropy Coding

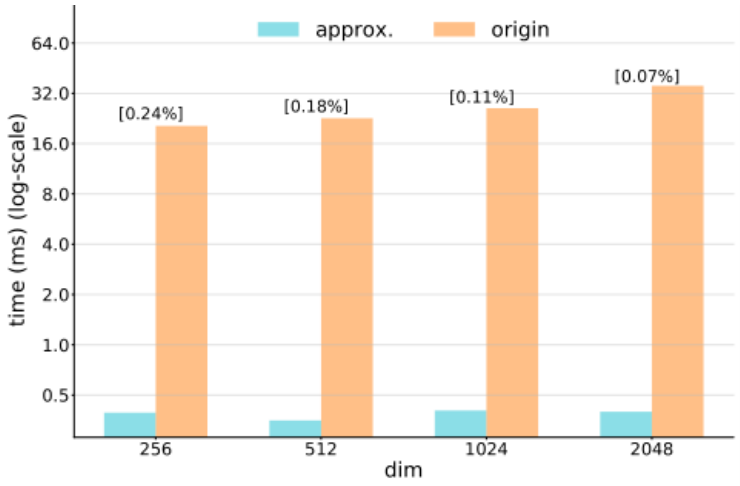
1，高维矩阵的对数行列式是非常昂贵的，并且可能导致病态矩阵的数值不稳定结果，这阻碍了其在大规模预训练中的应用

2，熵近似计算，加快计算过程，避免数值不稳定性：

$$\begin{aligned}\mathcal{L} &= \mu 1 \log\det\left(\mathcal{I}_m + \lambda \mathcal{Z}^\top \mathcal{Z}\right) \\ &= \mathrm{Tr}\left(\mu \log\left(I_m + \lambda \mathcal{Z}^\top \mathcal{Z}\right)\right)\end{aligned} \tag{7}$$

$$= \mathrm{Tr}\left(\mu \sum_{k=1}^{\infty} \frac{(-1)^{\kappa+1}}{k} \left(\lambda Z^{\mathsf{T}} Z\right)^{k}\right)$$
$$\approx \mathrm{Tr}\left(\mu \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k} \left(\lambda Z^{\mathsf{T}} Z\right)^{k}\right)$$

收敛条件：$\left|\left|\lambda Z^{\mathsf{T}} Z\right|\right|_2 < 1$，可以利用$\lambda$实现收敛条件。



调参：epoches，batch-size，n，$\left|\left|\lambda Z^{\mathsf{T}} Z\right|\right|_2 = [0.1, 0.2, 0.4, 0.6, 0.8, 1]$，$lr = \frac{bs * init_lr}{256}$，余弦衰减学习率，$\lambda$由小到大变化

Neural Networks Based Beam Codebooks: Learning mmWave Massive MIMO Beams That Adapt to Deployment and Hardware

1，根据具体部署、周围环境、用户分布和硬件特性调整码本波束模式。神经元权重直接模拟模拟移相器的波束成形权重，考虑到关键的硬件限制。通过在线和自我监督的训练来学习编码本波束，避免了对明确的信道状态信息的需求。

2，经典的波束转向码表有几个缺点：（i）它们通过扫描所有可能的方向而产生高的波束训练开销，即使其中许多方向可能永远不会被使用，（ii）它们通常有单叶波束，可能不是最佳的，特别是在非视线（NLOS）的情况下，以及（iii）它们通常是预定义的，不考虑可能的硬件缺陷（如相位不匹配或任意阵列几何形状）与非常昂贵的校准过程。

3，改进。(i) 对于任意的用户分布，我们的方法学习如何调整波束以集中在用户所在的位置，并大大减少所需的波束训练开销，(ii) 对于具有多个同等强度路径的NLOS场景，所开发的编码本学习解决方案学习多叶波束，以实现更高的数据速率，以及(iii) 对于具有硬件损伤或未知几何形状的阵列，我们的神经网络学习如何调整给定阵列的波束模式并减轻硬件损伤的影响。

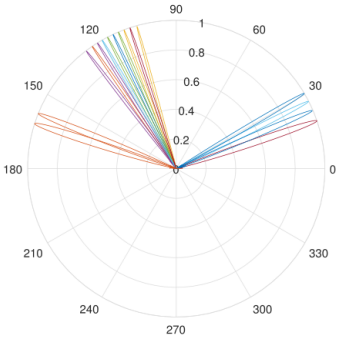
4，线性阵列响应建模，其中d_i为天线间距，$\Delta\theta$为移相器偏移

$$\mathbf{a}(\phi_\ell) = [e^{j(kd_1 cos(\Phi_l)+\Delta\theta_1)}, \ldots, e^{j(kd_M cos(\Phi_l)+\Delta\theta_M)}] \tag{8}$$

5，学习的是该环境场景下的码本，而非指向特定用户的波束，码本与环境强相关，用户码数从码本中选择。训练更新与批处理中的组合接收信号向量相关的波束，而不更新码本中的所有波束，这使得它在性能和收敛方面更稳定。环境适应码本不像DFT波束那样在整个方位面上扩展，这使得码本中的每一个波束都能为特定的用户群服务，并且这些波束都不会被任何方式“浪费”，没有光束浪费在根本没有用户的地方。

$$\mathcal{W}_{\text{opt}} = \arg\max_{\mathcal{W}} \sum_{k \in \mathcal{K}} \left(\max_{w_n \in \mathcal{W}, n=1,\ldots,N} |w_n^H h|^2 \right)$$

$$\mathbf{s.t.}\,[[\mathbf{w}_n]_m] = \frac{1}{\sqrt{M}},\ \forall m = 1,\ldots,M,\ n = 1,\ldots,N, \tag{9}$$



6，移相器相位调节误差+天线间距不均匀：自适应调节训练，将硬件缺陷转移到信道上，等效于信道发生细小变化，硬件无缺陷。

7，实数-复数求导：$w = w^r + jw^{im} \in \mathbb{C}, z = f(w) \in \mathbb{R}$，为实现$\frac{\partial \mathbf{z}}{\partial \mathbf{w}}$可导，将实部虚部分开作为两个独立的变量$w^{\mathbf{r}}$，$w^{\mathbf{im}} \in \mathbb{R}$，分别求导分别更新：

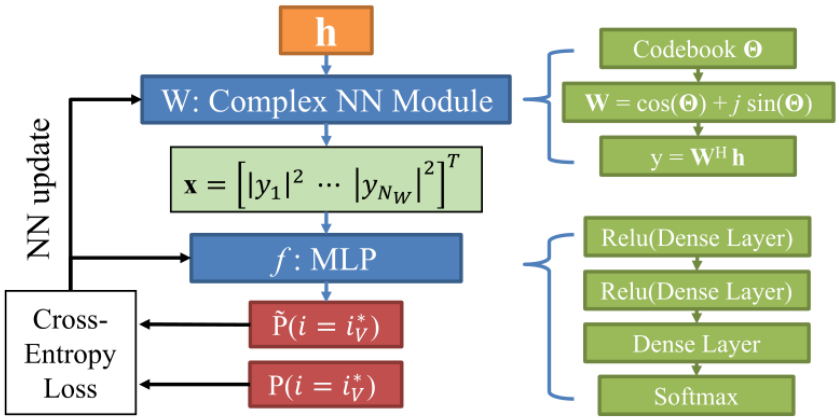
$$\begin{aligned}
\nabla z &= \left[\frac{\partial}{\partial w^{\mathrm{r}}} f(w), \frac{\partial}{\partial w^{\mathrm{im}}} f(w)\right]^T \\
w^r_{\mathrm{new}} &= w^r_{\mathrm{cur}} - \eta \cdot \frac{\partial z}{\partial w^r_{cur}} \\
w^{im}_{\mathrm{new}} &= w^{im}_{\mathrm{cur}} - \eta \cdot \frac{\partial z}{\partial w^{im}_{cur}} \\
w_{new} &= w^r_{new} + j \cdot w^{im}_{new}
\end{aligned} \tag{10}$$

Learning Site-Specific Probing Beams for Fast mmWave Beam Alignment

1，模型由表示模拟波束码本的复杂层和充当波束选择器的多层感知器（MLP）组成。网络以信道h为输入，以较小规模的探测码本W作为全连接层参数，获得环境的粗略信息，以此为依据利用MLP完成从完整码本V中选择通信窄波束，为提高性能可以使用$top-k$补测，选择输出概率最高的k个窄波束码本进行补测。测试阶段去除全连接层参数构建探测码本，用户信道扫描作为MLP输入，完成窄波束选择。

探测码本应基于整个环境向波束选择功能提供有用信息。NN以端到端的方式进行训练，而不是直接优化探测波束的BF增益。预测的最佳波束分布和真实最佳波束分布之间的交叉熵被用作损失函数。探测码本被隐式优化以辅助下游波束选择功能。

W中码字个数作为超参数，可以手动调节，同时w随着训练的进行不断优化以使用环境。窄波束码本V为固定的DFT码本不随环境变化。



2，通过使用探测码本执行波束扫描，高维信道向量$h \in \mathcal{C}^{N_t \times 1}$被转换为接收信号功率值$x \in \mathcal{R}^{N_w \times 1}$的特征向量，实现了维度压缩。具有相同最佳窄波束的信道应该在变换的子空间中彼此接近，而具有不同最佳窄波束信道应该彼此原理，这类似于ML中的表示学习问题，该问题通常寻求学习根据数据标签表现出自然聚类的高维数据的低维表示。

使用轮廓系数衡量聚类性能，其中$a(i)$为数据i与所处集群内其他点的距离的均值，$b(i)$表示数据i与属于其它聚类点的最小距离。系数越接近1表示聚类性能更好，更高的轮廓系数表示数据的更好的聚类和更好的可分离性，这可能会使最佳波束的分类更容易，系数越接近-1表示聚类性能更差。

$$S(\mathcal{D}) = \mathop{\mathrm{E}}_{i \in \mathcal{D}} \left[\frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \right] \tag{11}$$

3，调差：概率码本大小，量化移相器，信道估计误差，MLP层数，MLP/CNN，使用宽波束或者稀疏码本作为概率码本，补测数量，用户数，窄波束码本大小。