基于 VAE 的 Neural Topic Models 研究进展概述

Leilan Zhang
Tsinghua University

摘　要

本文介绍了当前主要的几种神经主题模型——基于 VAE 的 NVDM-GSM、基于 WAE 的 W-LDA、将词向量与主题向量相结合的 ETM、以及基于高斯混合先验的 GMNTM，梳理了各自的架构和目标函数的推导。

1　NVDM-GSM

NVDM-GSM 是基于标准变分自编码器的神经主题模型，由Miao et al. (2017) 提出。变分自编码器的目标是对隐变量 z 的真实后验分布 $p(z|x)$ 进行建模，通过贝叶斯法则（Bayes law）可以将后验概率由似然 $p(x|z)$，先验分布 $p(z)$ 和 x 的边际分布 $p(x)$ 表示为：

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{1}$$

其中分母原则上可通过式（2）

$$p(x) = \int p(z)p(x|z)\mathrm{d}z \tag{2}$$

来进行计算。然而由于式（2）中积分需要遍历 z 的所有取值，在高维空间中往往难以计算。因此，一般不直接求解真实后验分布 $p(z|x)$，而是求解变分后验分布 $q_\phi(z|x)$（ϕ 为变分参数），并不断缩小 $q_\phi(z|x)$ 与 $p(z|x)$ 之间的差异来达到逼近 $p(z|x)$ 的目的。$q_\phi(x)$ 通常选择易于计算的分布族，如高斯分布。本质上，这种方法是将推断问题转换为优化问题，通过最小化变分分布 $q_\phi(z|x)$ 和真实分布 $p(z|x)$ 之间的 KL 散度来求解 $p(z|x)$。分布 $q_\phi(z|x)$ 和 $p(z|x)$ 之间的 KL 散度定义为：

$$D_{\mathrm{KL}}\left[q_\phi(z|x) \| p(z|x)\right] = \sum_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z|x)} = \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p(z|x)}\right] \tag{3}$$

将式（3）中后验分布利用贝叶斯法则替换后得到：

$$\begin{aligned}
D_{\mathrm{KL}}\left[q_\phi(z|x) \| p(z|x)\right] &= \mathbb{E}_{q_\phi(z|x)}\left[\log q_\phi(z|x) - \log \frac{p(x|z)p(z)}{p(x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log q_\phi(z|x) - \log p(x|z) - \log p(z) + \log p(x)\right] \\
&= \log p(x) - \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p(x,z) - \log q_\phi(z|x)\right]}_{\mathrm{ELBO}(\phi)}
\end{aligned} \tag{4}$$

式（4）中右侧括弧中部分为 $logp(x)$ 的变分下界，记为 ELBO（Evidence Lower BOund）。在给定数据 x 之后，x 的分布 $p(x)$ 可视作常数，因此最小化式（4）左侧的 KL 散度的优化目标，

等价于最大化 ELBO。使用插项的技巧，式（4）中 ELBO 可重写为：

$$
\begin{aligned}
\mathrm{ELBO}(\phi) &= -\mathbb{E}\left[\log p(x,z) - p(z) + p(z) - \log q_\phi(z|x)\right] \\
&= \mathbb{E}[\log p(x|z)] - D_{\mathrm{KL}}\left[q_\phi(z|x) \| p(z)\right]
\end{aligned}
\tag{5}
$$

因此，最大化 ELBO 等价于最大化式（5）右侧，其中第一项为似然函数，将迫使解码器将生成样本 x' 尽可能还原为输入样本 x，通常采用交叉熵进行度量；第二项为关于 z 的分布的正则项，将迫使变分后验分布 $q_\phi(z|x)$ 逼近先验分布 $p(z)$。

在实践中，式（5）中的后验分布 $q_\phi(z|x)$ 与先验分布 $p(z)$ 都需要确定为具体的分布才能进行优化，常用的假设是将这两个分布都选定为多元高斯分布，其中后验分布 $q_\phi(z|x)$ 的均值和方差分别假定为 $\mu(x)$ 和 $\Sigma(x)$，并假定其协方差阵为对角阵，先验分布 $p(z)$ 则通常取标准正态分布 $\mathcal{N}(0,1)$。因此，实际的优化目标 $D_{\mathrm{KL}}\left[q_\phi(z|x) \| p(z|x)\right]$ 为：

$$
\begin{aligned}
D_{\mathrm{KL}}\left[q_\phi(z|x) \| p(z)\right] &= D_{\mathrm{KL}}[\mathcal{N}(\mu(x), \Sigma(x)) \| \mathcal{N}(0,1)] \\
&= \frac{1}{2}\left(\operatorname{tr}(\Sigma(x)) + \mu(x)^T \mu(x) - d - \log\det(\Sigma(x))\right)
\end{aligned}
\tag{6}
$$

其中，d 为 z 的维数，$tr(.)$ 为迹运算。图1展示了 VAE 的架构，其中，$q_\phi(z|x)$ 作为编码器，将数据 x 映射为隐变量 z 的分布的均值 $\mu(x)$ 和方差 $\sigma(x)$，从该分布中采样得到 $z \sim \mathcal{N}(\mu(x), \Sigma(x))$，$p_\rho(x|z)$ 则用作解码器，通过隐变量生成样本 x'，分布中的参数 ϕ 与 ρ 分别对应编码器和解码器网络中的权重参数。在对隐变量 z 的采样操作中，直接的采样操作并不可导，网络参数难以更新，为此，VAE 中采用了重参数Kingma et al. (2014) 的技巧，取 $z = \mu + \epsilon * \sigma$，$\epsilon \sim N(0, I)$，则仍有 $z \sim N(\mu, \sigma)$，在保证分布不变的同时也满足了 z 对 μ 和 σ 的可导性。

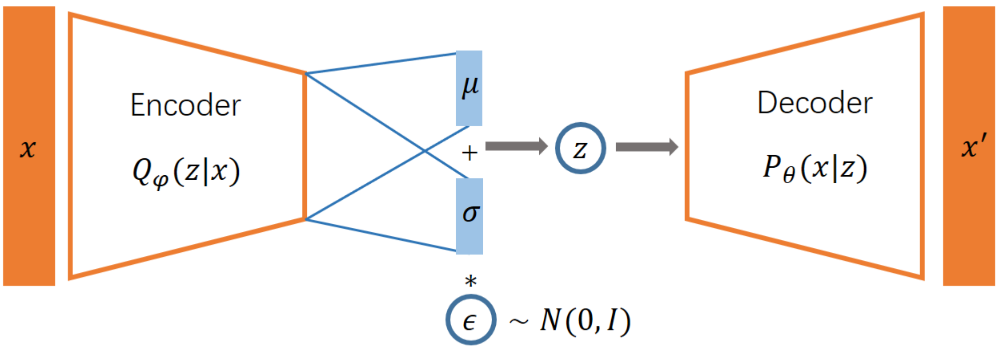

图 1: VAE 的网络架构

主题模型 NVDM-GSM（Gaussian-Softmax）采用了上述标准 VAE 的架构，以文档的词袋表示 BOW（Bag Of Word）作为输入。考虑到分布需满足规范性，因此从隐空间采样得到高斯变量 z 后，还需要将 z 归一化才能作为主题分布，NTM-GSM 采取的方法是使 z 通过 Softmax 层，即

$$
\begin{aligned}
z &\sim \mathcal{N}\left(\mu(x), \sigma(x)^2\right) \\
\theta &= \mathrm{Softmax}\left(W_1^T z\right)
\end{aligned}
\tag{7}
$$

其中 W_1 为 $L * K$ 的矩阵，L 为 z 的维数，K 为主题数。由此求得的归一化向量 θ（K 维）作为文档的主题分布向量，在导入解码器 $P(x|z)$ 后得到重构文档 x'。解码器中的权重参数 ρ_{K*V} 即为主题-词分布矩阵，令 θ 取第 k 维为 1 的 One-hot 向量并导入解码器，即可得到第 k 个主题的主题-词分布。

2 W-LDA

为了解决 VAE 的后验坍塌问题，同时能够利用 Dirichlet 分布作为隐空间的先验分布，Tolstikhin 等人Tolstikhin et al. (2018) 提出基于 WAE（Wasserstein Auto-Encoder）构建主题模型。

WAE 基于 Wasserstein 距离来度量先验分布和后验分布的差异。相比于 KL 散度，Wassertein 距离在两个分布没有重叠时仍然能保持连续并反映两个分布的远近。分布 P_x 与 P_y 的 Wasserstein 距离定义为：

$$W\left(P_x, P_y\right)=\inf_{\gamma \in \Pi\left(P_x, P_y\right)} \mathbb{E}_{(x, y) \sim \gamma}[\|x-y\|] \tag{8}$$

式（8）中 $\Pi\left(P_x, P_y\right)$ 为 (x, y) 的联合分布的集合，满足 $x \sim P_x, y \sim P_y$，对于其中 x 与 y 的某个联合分布 γ，可以求得 γ 下所有 x 与 y 的距离的期望，所有期望的下确界（infimum）即定义为 P_x 和 P_y 的 Wasserstein 距离。

作为 Wasserstein 的定义式，式（8）难以直接用于计算，Tolstikhin 等人Tolstikhin et al. (2018) 通过推导，提出可采用式（9）作为实践中 WAE 的优化目标，即：

$$D_{WAE}\left(P_x, P_G\right):=\inf_{Q(z|x) \in \mathcal{Q}} \mathbb{E}_{P_x} \mathbb{E}_{Q(z|x)}[\|x-G(z)\|]+\lambda \cdot \mathcal{D}_z\left(Q_z, P_z\right) \tag{9}$$

其中，G 为解码器，Q_z 为经编码器 Q 映射后的边缘分布，与 VAE 不同的是，WAE 可以使用确定性的编码器 $P(z|x)$，而不需通过采样的方法得到隐变量 z，因为不需为每一个样本 x 在隐空间中求得对应的一个分布，只需要 z 的边缘分布接近先验分布即可。G 将隐变量映射为生成样本 $x^{\prime}$，因此式（9）第一项为重构误差，度量了生成样本与输入样本之间的差异，第二项 $D_z(Q_z, P_z)$ 则选用 MMD（Maximum Mean Discrepancy）度量先验分布和变分分布的差异。

WAE 的整体算法为：

Algorithm 1 WAE-MMD process

Parameter: 编码器 Q 参数 ϕ, 解码器 G 参数 θ，正定核函数 k

1: **while** 停止条件不满足 **do**
2: 从训练集中采样 $\{x_1, x_2, ..., x_n\}$
3: 从先验分布 P_z 中采样 $\{z_1, z_2, ..., z_n\}$
4: 从后验分布 $Q_\phi(z|x_i)$ 中采样 $\tilde{z}_i, (i=1,2,...,n)$
5: 通过最小化式（10）来更新 Q_ϕ 和 G_θ：

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \|x_i, G_\theta\left(\tilde{z}_i\right)\|+\frac{\lambda}{n(n-1)} \sum_{\ell \neq j} k\left(z_\ell, z_j\right) \\ & +\frac{\lambda}{n(n-1)} \sum_{\ell \neq j} k\left(\tilde{z}_\ell, \tilde{z}_j\right)-\frac{2 \lambda}{n^2} \sum_{\ell, j} k\left(z_\ell, \tilde{z}_j\right) \end{aligned} \tag{10}$$

6: **end while**

式（10）中后三项为 MMD 的离散形式，在分别从先验分布和后验分布中采样后，通过样本的 MMD 值来估计分布 Q_z 与 P_z 之间的实际 MMD 值。

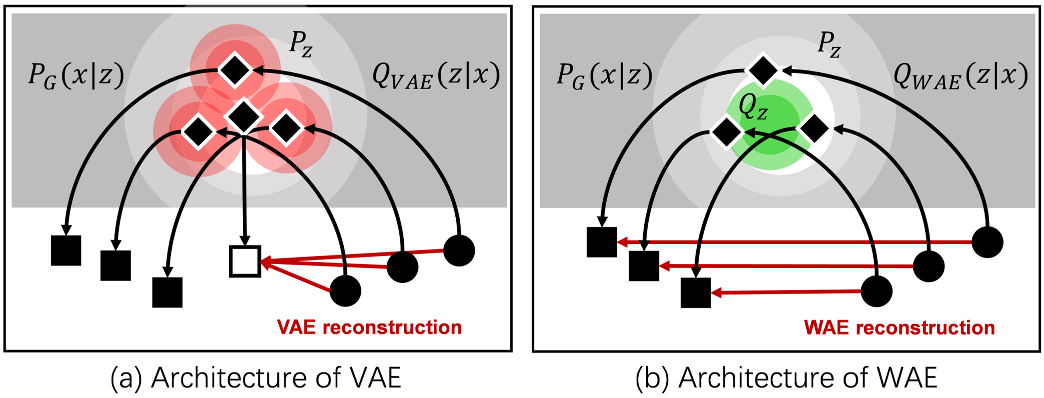

图 2: VAE 与 WAE 的差异对比

图2对比了 VAE 与 WAE 的不同点。WAE 与 VAE 的优化目标都由重构误差，以及隐空间中先验分布 P_z 及变分后验分布 $Q(z|x)$ 间的差异构成的正则项组成。在 VAE 中（图2(a)），每个输入样本对应一个分布（图2(a) 中每一个红色的小圆圈），VAE 的优化目标是使这些分布都趋近标准高斯分布（图2(a) 中白色的大圆圈所示），造成的结果是各个分布之间出现层叠，在层叠区域，一个隐变量 z 会对应多个输入样本，使得重构样本实际上是这些输入样本的平均值，正是由于这个原因，VAE 所生成的图像容易出现模糊；在 WAE 中，每个输入样本对应的是一个隐变量 z 而非一个分布，WAE 的优化目标是使后验的边际分布 $Q(z) = \int P(z|x)dP_x$（图2(b) 中绿色圆圈）趋近先验分布 P_Z（图2(b) 中白色大圆圈所示），各输入样本对应的 z 在隐空间中可以合理分布，避免了分布层叠的问题，可以生成更加清晰的图像。

基于 WAE 构建的主题模型 W-LDA，以文档的词袋表示 BOW 作为输入，编码器由多层感知机构成，并经 Softmax 层生成隐变量 θ 作为主题分布。与 NTM-GSM 中不同，W-LDA 中 θ 由确定性映射得到，无需经过采样操作。

与 LDA 一致，W-LDA 主题模型采用 Dirichlet 分布作为主题的先验分布；由于 θ 需满足归一化约束，即 θ 的可行解空间构成单纯形，因此要求核函数在单纯形上有较好的度量意义，W-LDA 选取了信息扩散核作为 MMD 的核函数：

$$\mathbf{k}\left(\theta, \theta'\right) = \exp\left(-\arccos^2\left(\sum_{i=1}^{d}\sqrt{\theta_i\theta_i'}\right)\right) \tag{11}$$

该核函数度量了将不同向量映射到球面后的像之间的测地线长，相较于常见的 L^2 范数，该核函数对位于单纯形边界处的点更为灵敏，因此更适合于数据稀疏的场景。与 NTM-GSM 相同，解码器的权重参数 β 构成了主题-词分布，因此，令 θ 取遍不同的 One-hot 向量并导入解码器，即可得到对应主题的主题-词分布。

3 ETM

现有主题模型通常通过主题-词分布来描述主题，为了更细致地刻画所挖掘的主题的可解释性，Dieng 等人Dieng et al. (2019) 提出了 ETM（Embedded Topic Model），在神经主题模型中引入了词向量和同样维数的主题向量，使主题与词可在同一个嵌入空间进行表示。

ETM 中定义了一个词向量矩阵 ρ，大小为 $L * V$，其中 L 是词向量的维数，V 是词汇表的

大小；同时定义了一个主题向量矩阵 α，大小为 $L*K$，其中 K 为主题数。与 NVDM-GSM 类似，在 ETM 中，文档的词袋表示 BOW 由编码器映射到隐空间的高斯分布，经采样得到隐变量 z，经过 Softmax 归一化后得到主题分布，即 $\theta = Softmax(Wz+b), z \sim N(\mu(x), \sigma(x))$；主题-词分布反映的是每个词在各主题中的重要程度，ETM 中将第 k 个主题的主题向量与各个词向量进行内积后，再经 Softmax 归一化的结果作为其主题-词分布 β_k，即：$\beta_k = Softmax(\rho^T \alpha_k)$。

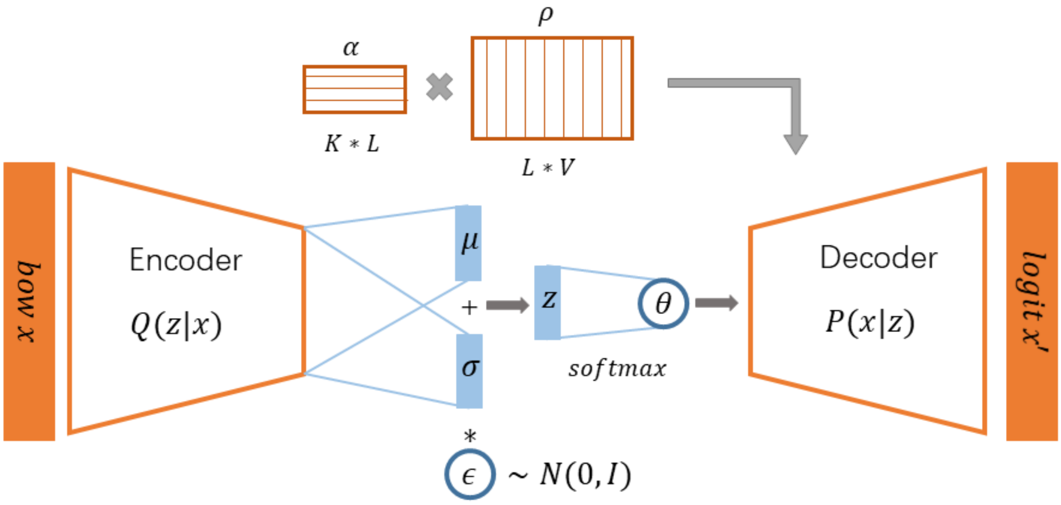

图 3: ETM 的网络架构

ETM 的网络架构如图3所示。对比图1，ETM 与 NVDM-GSM 最主要的区别在于，ETM 中主题-词分布矩阵 β_{K*V} 被分解为主题向量矩阵 α_{K*L} 与词向量矩阵 ρ_{L*V}，词向量的训练与主题模型的训练同步进行，使得主题模型能根据词嵌入表示的调整逐步完善。ETM 的目标函数与 NVDM-GSM 相同，由重构误差及隐变量的后验分布同先验分布的 KL 散度构成。

ETM 的算法整体流程为：

Algorithm 2 ETM process

Parameter: 主题向量矩阵 α，词向量矩阵 ρ

  for $epoch = 1, 2, ..., N$ **do**

    对每一主题 k，计算主题-词分布 $\beta_k = Softmax(\rho^T \alpha_k)$

    从文档集中选择一个 minibatch **B**

    for d ∈**B** **do**

      计算 d 的归一化词袋表示 x_d

      计算均值向量和对数方差向量 $\mu_d = \mu(x_d)$)，$log\sigma_d = log\sigma(x_d)$

      采样隐变量 $z \sim N(\mu_d, \Sigma_d)$，计算主题分布 $\theta_d = Softmax(Wz+b)$

      计算重构样本的生成概率 $p(x^{'}|\theta_d) = \theta_d^T \beta$

    end for

    计算 ELBO 及其导数，更新 $\alpha_{1:K}$，ρ 及网络参数

  end for

---

4 基于高斯混合先验的神经主题模型 GMNTM

在 VAE 中，尽管先验分布的多元高斯分布可以简化计算，但这种单峰的假设也对隐空间形成限制，同时由于 VAE 将各输入样本对应的隐变量分布都逼近标准高斯分布，这些高度层叠的分布可能会降低模型的性能。从主题模型的角度看，相同主题的文档对应的隐变量应该具有较近的距离，而不同主题的文档则应该被映射到隐空间中相距较远的位置，即相似主题的文档应该在隐空间聚成簇，而不同主题的文档应分布在不同的簇中。而 VAE 将所有分布都逼近单一簇的特性，将使得不同的主题对应的隐变量有混杂的趋势，可能导致不同主题的文档不易区分开，进而造成得到的主题在主题多样性和主题连贯性上都不够理想。

在对话文本中，由于话语长度较短，每个话语所包含的主题较为单一，因此对于属于同一主题的话语，以单峰的高斯分布作为先验较合适；属于不同主题的话语，其对应的主题的先验分布则应有好的区分性，尽可能减少分布层叠的情形。

因此，本文提出用高斯混合分布作为主题隐变量的先验分布，以替代 VAE 中以标准高斯分布为先验分布的假设，并基于高斯混合变分自编码器（GMVAEJiang et al. (2017)Dilokthanakul et al. (2016)）来设计主题模型。

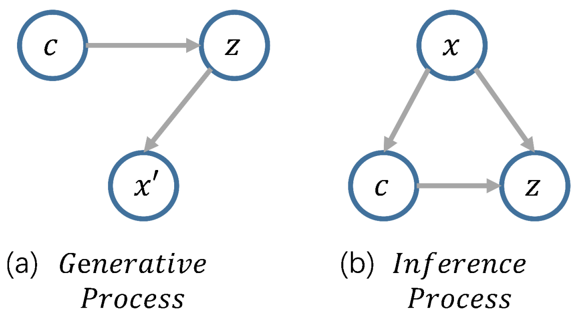

图 4: GMVAE 的概率图模型

GMVAE 引入了离散的类别隐变量 c，用以指示每个输入样本的所属的高斯成分；z 为连续隐变量，产生于 c 所指定的高斯分布。其生成过程如图4(a) 所示，设预先给定的高斯成分个数为 K，GMVAE 首先从多项分布 $Multi(\boldsymbol{\pi})$ 中采样得到类别变量 c，由 c 所对应的高斯分布 $\mathcal{N}(\mu_c, \sigma_c^2\boldsymbol{I})$ 中采样得到隐变量 z，z 再经变换 $f(z;\phi)$ 后得到生成样本 x。

根据图4(a) 所展示的条件依赖关系，x、c 和 z 的联合概率分布 $p(x, z, c)$ 可分解为：

$$p(x, z, c) = p(c)p(z|c)p(x|z) \tag{12}$$

式（12）中各概率满足：

$$\begin{aligned} p(c) &= Multi(\boldsymbol{\pi}) \\ p(z|c) &= \mathcal{N}(z|\mu_c, \sigma_c^2\boldsymbol{I}) \\ p(x|z) &= \mathcal{B}\rceil\nabla(x|\mu_x) \end{aligned} \tag{13}$$

其中 $\mu_x = f(z;\theta)$，$\mathcal{B}\rceil\nabla(x|\mu_x)$ 为参数为 μ_x 的 Bernouli 分布。根据上述生成过程，由 log 函

数的上凸性及 Jensen 不等式，可求得 x 的对数似然的变分下界：

$$\begin{aligned} logp(x) = log\int_z \Sigma_c p(x,z,c)dz = log\int_z \Sigma_c q(z,c|x)\frac{p(x,z,c)}{q(z,c|x)}dz \\ \geq \int_z \Sigma_c q(z,c|x) log\frac{p(x,z,c)}{q(z,c|x)}dz = \mathbb{E}_q(z,c|x)[log\frac{p(x,z,c)}{q(z,c|x)}] = \mathcal{L}_{ELBO} \end{aligned} \tag{14}$$

其中 $q(z,c|x)$ 为变分后验分布，由图4(b) 所展示的推断过程，$q(z,c|x)$ 可分解为 $q(z,c|x) = q(z|x)q(c|x)$，代入式（14）则有：

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(\mathbf{x}) =& E_{q(\mathbf{z},c|\mathbf{x})}\left[\log\frac{p(\mathbf{x},\mathbf{z},c)}{q(\mathbf{z},c|\mathbf{x})}\right] \\ =& E_{q(\mathbf{z},c|\mathbf{x})}[\log p(\mathbf{x},\mathbf{z},c) - \log q(\mathbf{z},c|\mathbf{x})] \\ =& E_{q(\mathbf{z},c|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}|c) \\ & + \log p(c) - \log q(\mathbf{z}|\mathbf{x}) - \log q(c|\mathbf{x})] \end{aligned} \tag{15}$$

与 VAE 中类似，式（15）中的变分后验分布 $q(z|x)$ 设为高斯分布，通过神经网络 $g(x)$ 进行建模，即：

$$\begin{aligned} \left[\tilde{\mu}; \log\tilde{\sigma}^2\right] = g(\mathrm{x};\phi) \\ q(\mathrm{z}|\mathrm{x}) = \mathcal{N}\left(\mathrm{z};\tilde{\mu},\tilde{\sigma}^2\mathrm{I}\right) \end{aligned} \tag{16}$$

将式（13）和式（16）代入式（15），可得到：

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} = & \frac{1}{M}\sum_{l=1}^{M}\sum_{i=1}^{D} x_i \log \boldsymbol{\mu}_x^{(l)}\Big|_i + (1-x_i)\log\left(1-\boldsymbol{\mu}_x^{(l)}\Big|_i\right) \\ & -\frac{1}{2}\sum_{c=1}^{K}\eta_c\sum_{j=1}^{H}\left(\log\sigma_c^2\big|_j + \frac{\tilde{\boldsymbol{\sigma}}^2\big|_j}{\sigma_c^2\big|_j} + \frac{\left(\tilde{\boldsymbol{\mu}}\big|_j - \boldsymbol{\mu}_c\big|_j\right)^2}{\sigma_c^2\big|_j}\right) \\ & + \sum_{c=1}^{K}\eta_c\log\frac{\pi_c}{\eta_c} + \frac{1}{2}\sum_{j=1}^{H}\left(1+\log\tilde{\sigma}^2\big|_j\right) \end{aligned} \tag{17}$$

其中 M 为所采样的 z 的个数，D 为 x 的维数，J 为隐变量 z 的维数，π_c 为第 c 个高斯成分的先验概率，$\eta_c = q(c|x)$，可通过式（18）进行估计（详细推导参见附录）：

$$q(c|x) = p(c|z) = \frac{p(c)p(z|c)}{\Sigma_{c'=1}^{K} p(c')p(z|c')} \tag{18}$$

GMNTM 的整体网络结构如图5所示：

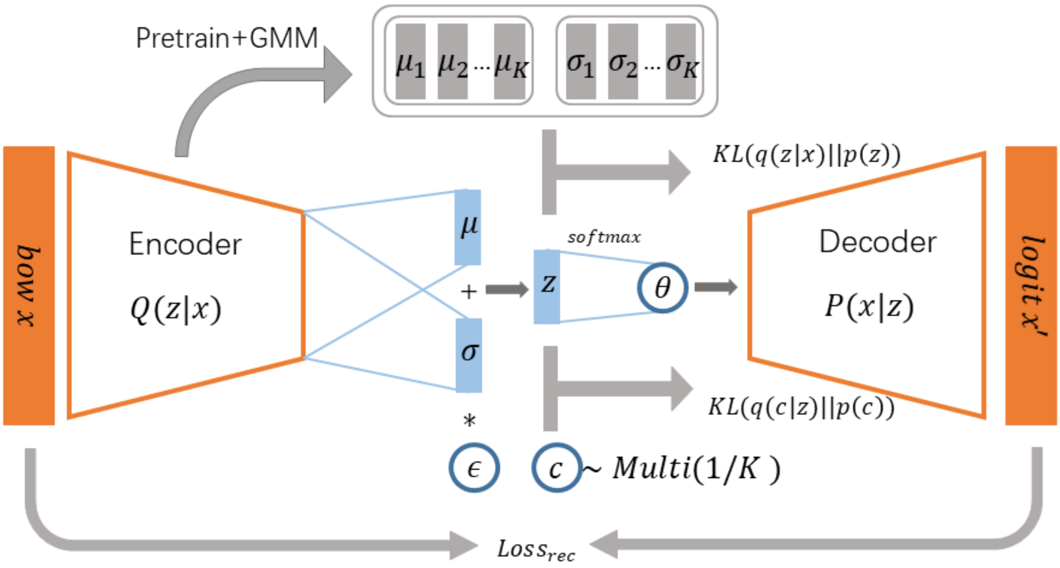

图 5: GMVAE 的网络结构

该网络首先需要经过自编码器的重构过程进行预训练，训练结束后使用高斯混合模型 GMM 进行聚类，将所得的各高斯成分的均值 μ_K 和方差 σ_K 作为隐空间中高斯混合分布的初始值，类别隐变量 c 的先验分布设定为均匀的离散分布，即 $Multi(\frac{1}{K})$。对于输入样本 x，由编码器 Q 映射得到隐变量 z 的分布的均值 $\tilde{\mu}$ 和对数方差 $log\tilde{\sigma}^2$，采样得到 z，由各高斯成分的均值和方差及对应的权重，可求得 z 的先验分布 $p(z) = \Sigma_{i=1}^{K} \pi_i \mathcal{N}(z; \mu_i, \sigma_i^2)$，进而得到 z 的变分后验分布 $q(z|x)$ 与先验 $p(z)$ 的 KL 散度；由式（18），可求得 c 的变分后验分布 $q(c|x)$ 与先验分布 $p(c) = \frac{1}{K}$ 的 KL 散度；重构误差加上上述 KL 正则项即构成了该网络的优化目标。与 NTM-GSM 类似，基于 GMVAE 构建主题模型时，将隐变量 z 经过 Softmax 层后得到的归一化向量 θ 作为主题分布向量，即 $\theta = Softmax(Wz + b)$；将第 k 个高斯成分的均值经变换后得到的 θ_k 导入解码器，所得的归一化输出即为第 k 个主题-词分布。

参考文献

DIENG A B, RUIZ F J R, BLEI D M, 2019. Topic modeling in embedding spaces[J]. CoRR, abs/1907.04907.

DILOKTHANAKUL N, MEDIANO P A M, GARNELO M, et al., 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders[J]. CoRR, abs/1611.02648.

JIANG Z, ZHENG Y, TAN H, et al., 2017. Variational deep embedding: An unsupervised and generative approach to clustering [C]//SIERRA C. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. [S.l.]: ijcai.org: 1965-1972.

KINGMA D P, WELLING M, 2014. Auto-encoding variational bayes[C]//2nd International Conference on Learning Representations, ICLR. [S.l.: s.n.].

MIAO Y, GREFENSTETTE E, BLUNSOM P, 2017. Discovering discrete latent topics with neural variational inference[C]// Proceedings of the 34th International Conference on Machine Learning, ICML. [S.l.]: PMLR: 2410-2419.

TOLSTIKHIN I O, BOUSQUET O, GELLY S, et al., 2018. Wasserstein auto-encoders[C]//Proceedings of the 6th International Conference on Learning Representations, ICLR. [S.l.]: OpenReview.net.

A 公式 (18): $q(c|x) = \mathbb{E}_{q(z|x)}[p(c|z)]$ 的证明

由于 $\mathcal{L}_{ELBO}$ 可表示为:

$$\begin{aligned}
\mathcal{L}_{\text{ELBO}}(\mathbf{x}) &= E_{q(\mathbf{z},\mathbf{c}|\mathbf{x})}\left[\log\frac{p(\mathbf{x},\mathbf{z},c)}{q(\mathbf{z},c|\mathbf{x})}\right] \\
&= \int_{\mathbf{z}}\sum_{\mathbf{c}} q(\mathbf{z},c|\mathbf{x})\log\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|c)p(c)}{q(\mathbf{z},c|\mathbf{x})}d\mathbf{z} \\
&= \int_{\mathbf{z}}\sum_{c} q(c|\mathbf{x})q(\mathbf{z}|\mathbf{x})\log\frac{p(\mathbf{x}|\mathbf{z})p(c|\mathbf{z})p(\mathbf{z})}{q(c|\mathbf{x})q(\mathbf{z}|\mathbf{x})}d\mathbf{z} \\
&= \int_{\mathbf{z}}\sum_{c} q(c|\mathbf{x})q(\mathbf{z}|\mathbf{x})\left[\log\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}+\log\frac{p(c|\mathbf{z})}{q(c|\mathbf{x})}\right]d\mathbf{z} \\
&= \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x})\log\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}d\mathbf{z} - \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x})\sum_{\mathbf{c}} q(c|\mathbf{x})\log\frac{q(c|\mathbf{x})}{p(c|\mathbf{z})}d\mathbf{z} \\
&= \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x})\log\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}d\mathbf{z} - \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x})D_{KL}(q(c|\mathbf{x})\|p(c|\mathbf{z}))d\mathbf{z}
\end{aligned} \tag{19}$$

式 (19) 中第一项与 c 无关,而第二项非负,因此关于 $q(c|x)$ 最大化 $\mathcal{L}_{ELBO}$ 意味着 $D_{KL}(q(c|x)||p(c|z)) = 0$，因此有

$$\frac{q(c|x)}{p(c|z)} = u \tag{20}$$

其中 u 为常数。又因为 $\Sigma_c q(c|x) = 1$ 且 $\Sigma_c p(c|z) = 1$，则有

$$\frac{q(c|x)}{p(c|z)} = 1 \tag{21}$$

两边同时取期望，则有 $q(c|x) = \mathbb{E}_{q(z|x)}[p(c|z)]$。