Оптимизация параметров модели на основе дистилляции знаний

Рассматривается задача дистилляции модели. Будем корректировать траекторию оптимизации на основе двухуровневой задачи оптимизации:

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}\in\mathbb{R}^2} \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\mathbf{val}}} \sum_{k=1}^{K} y^k \log \mathbf{g}(\mathbf{x},\mathbf{w})|_{T=1}$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbb{R}^s} (1-\beta) \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\mathbf{train}}} \sum_{k=1}^{K} y^k \log \mathbf{g}(\mathbf{x},\mathbf{w})|_{T=1} + \beta \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\mathbf{train}}} \sum_{k=1}^{K} \mathbf{f}(\mathbf{x})|_{T=T_0} \log \mathbf{g}(\mathbf{x},\mathbf{w})|_{T=T_0}$$

где $\mathbf{h} = [\beta, T_0]$ — параметры дистилляционного слагаемого.