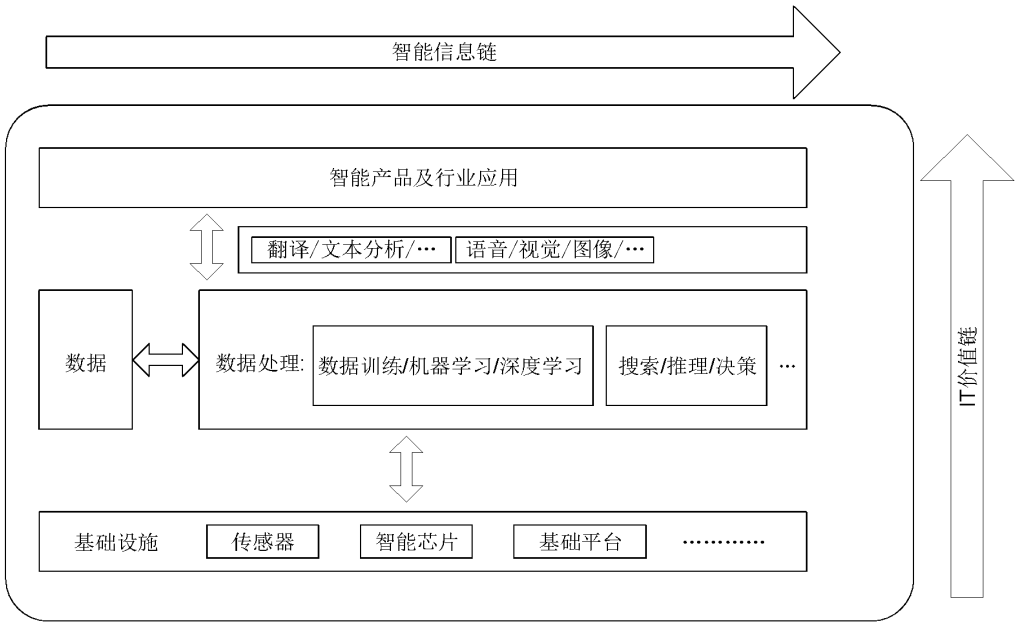
智能信息链
智能产品及行业应用
翻译/文本分析/…
语音/视觉/图像/…
数据
数据处理:
数据训练/机器学习/深度学习
搜索/推理/决策
…
基础设施
传感器
智能芯片
基础平台
……
IT价值链

图 1

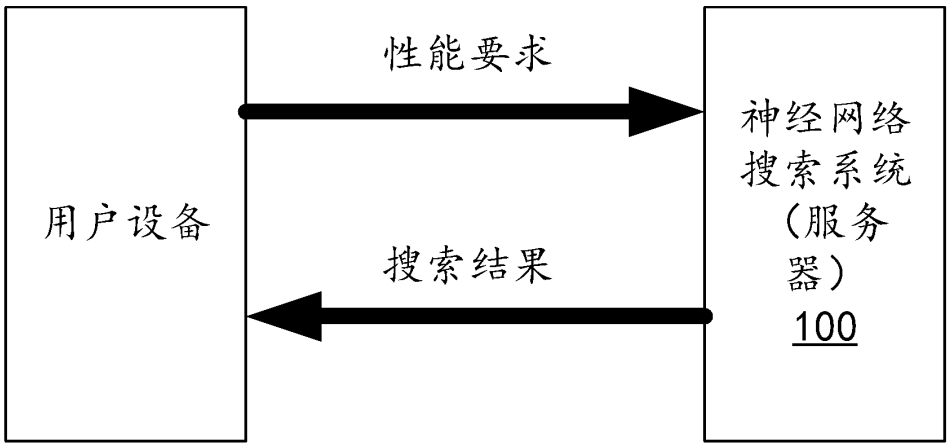
性能要求
用户设备
神经网络搜索系统（服务器）
100
搜索结果

图 2

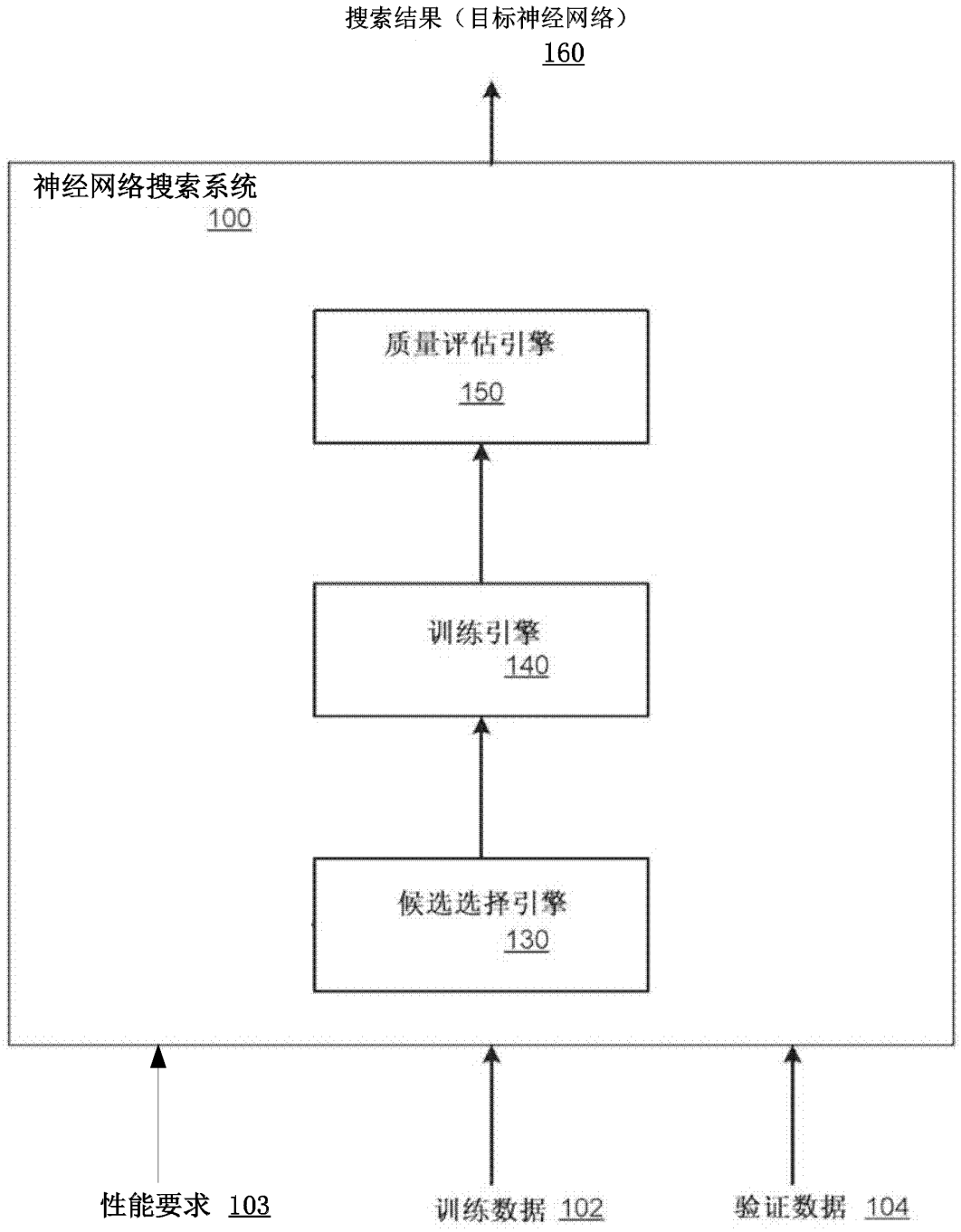
搜索结果（目标神经网络）
160
神经网络搜索系统
100
质量评估引擎
150
训练引擎
140
候选选择引擎
130
性能要求 103
训练数据 102
验证数据 104

图 3

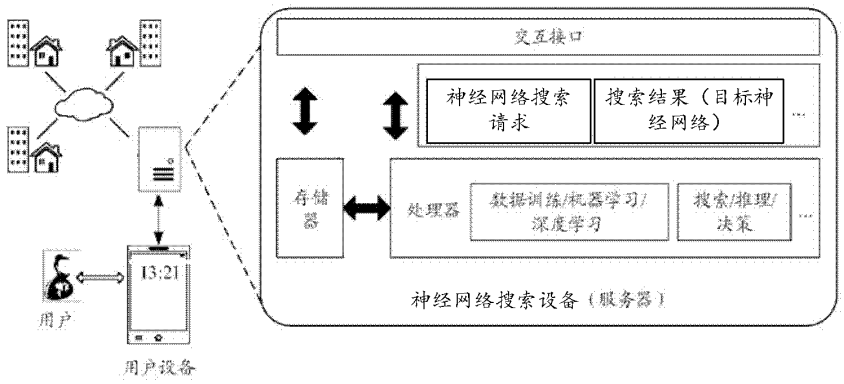
交互接口
神经网络搜索请求
搜索结果（目标神经网络）
存储器
处理器
数据训练/机器学习/深度学习
搜索/推理/决策
神经网络搜索设备（服务器）
用户
用户设备

图 4

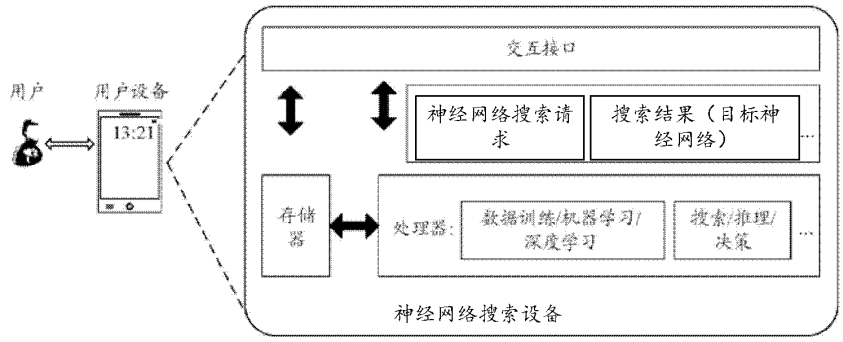
交互接口
用户
用户设备
神经网络搜索请求
搜索结果（目标神经网络）
存储器
处理器
数据训练/机器学习/深度学习
搜索/推理/决策
神经网络搜索设备

图 5

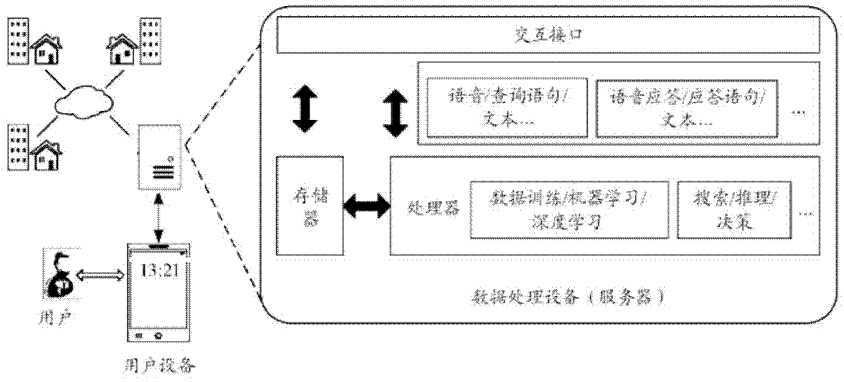
交互接口
语音/查询语句/文本...
语音应答/应答语句/文本...
存储器
处理器
数据训练/机器学习/深度学习
搜索/推理/决策
数据处理设备（服务器）
用户
用户设备

图 6

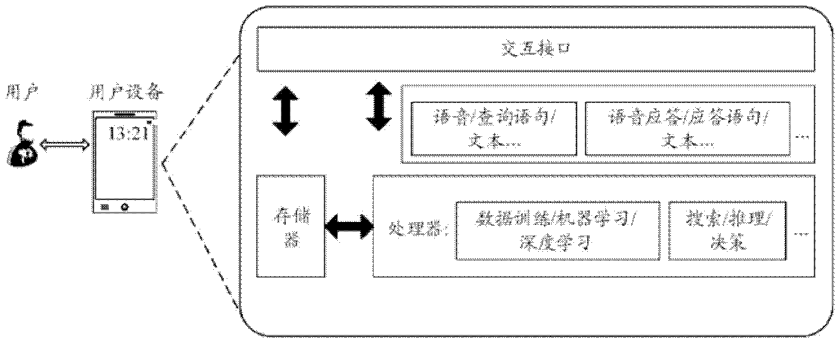
用户
用户设备
13:21
交互接口
语音/查询语句/文本...
语音应答/应答语句/文本...
存储器
处理器:
数据训练/机器学习/深度学习
搜索/推理/决策

图 7

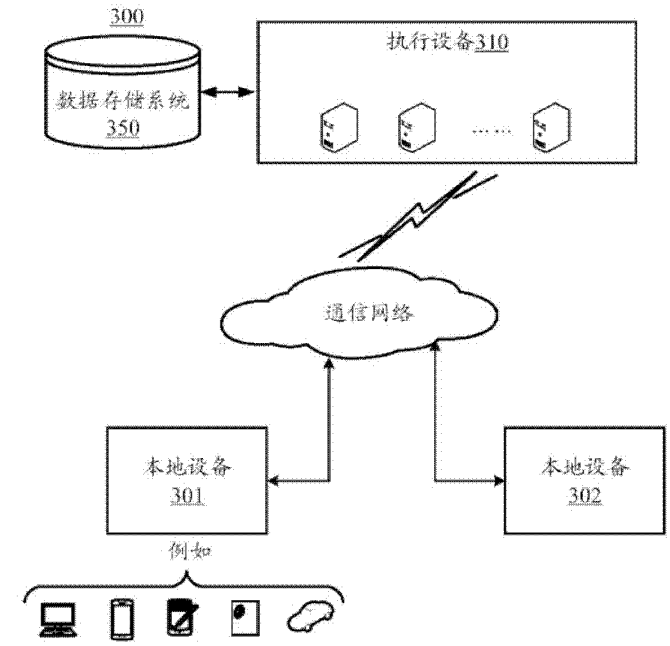
300
数据存储系统
350
执行设备310
通信网络
本地设备
301
本地设备
302
例如

图 8

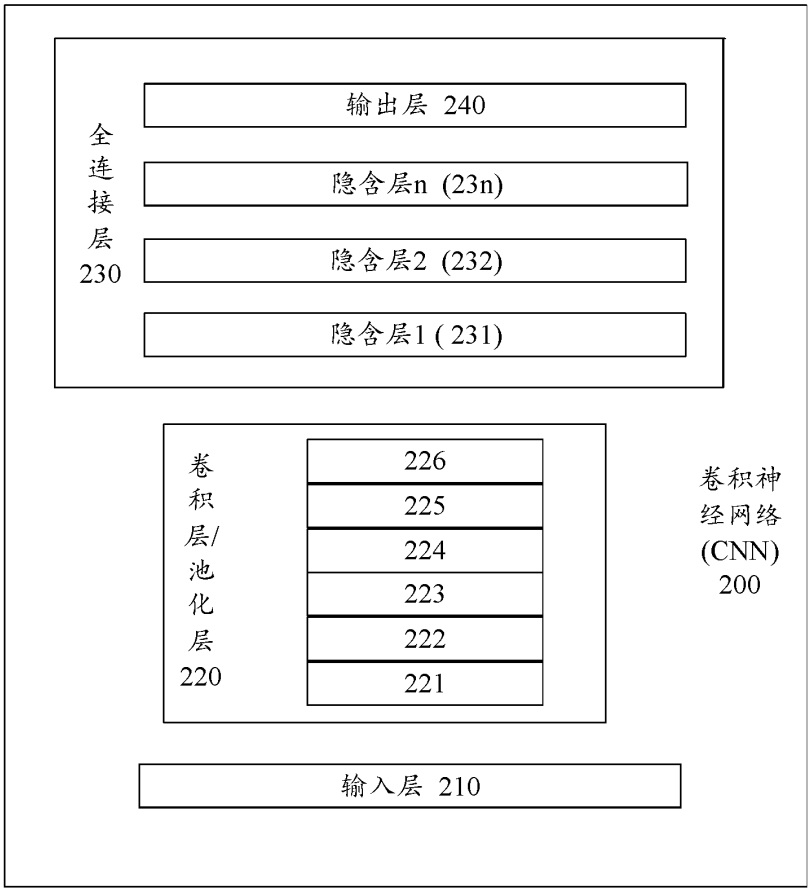
输出层 240
隐含层n (23n)
隐含层2 (232)
隐含层1 (231)
全连接层 230
226
225
224
223
222
221
卷积层/池化层 220
卷积神经网络 (CNN) 200
输入层 210

图 9

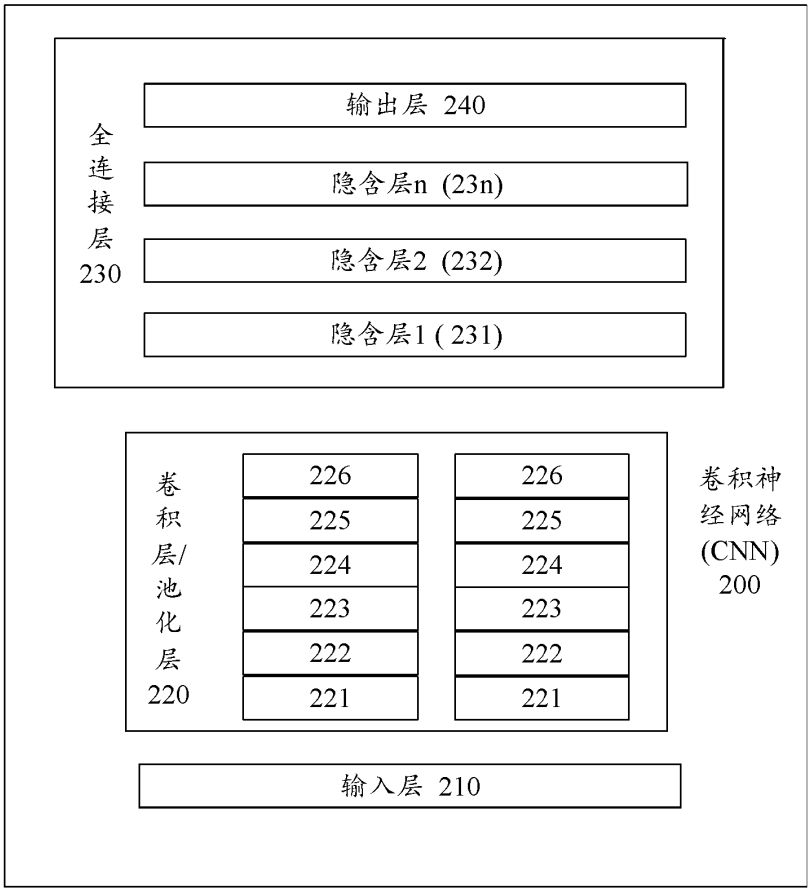
全连接层 230
输出层 240
隐含层n (23n)
隐含层2 (232)
隐含层1 (231)
卷积层/池化层 220
226
225
224
223
222
221
226
225
224
223
222
221
卷积神经网络 (CNN) 200
输入层 210

图 10

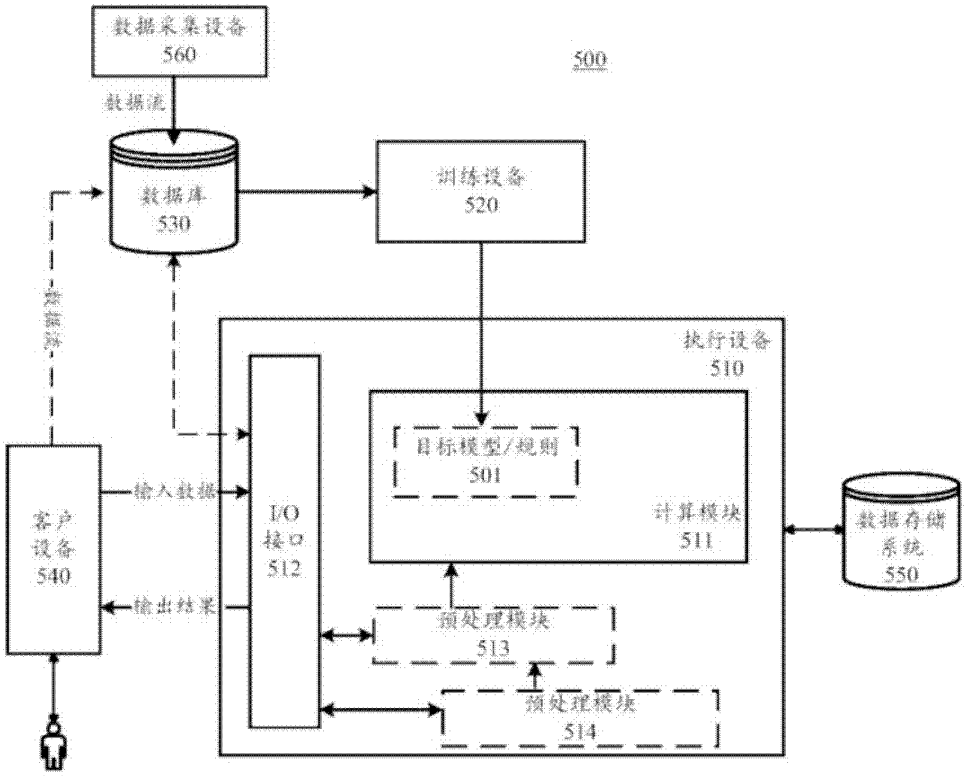
数据采集设备
560
500
数据流
数据库
530
训练设备
520
执行设备
510
目标模型/规则
501
计算模块
511
I/O
接口
512
客户
设备
540
输入数据
输出结果
预处理模块
513
预处理模块
514
数据存储
系统
550

图 11

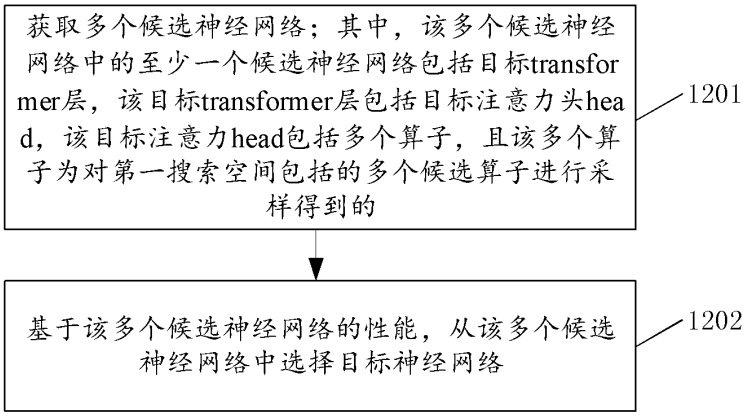
获取多个候选神经网络；其中，该多个候选神经网络中的至少一个候选神经网络包括目标transformer层，该目标transformer层包括目标注意力头head，该目标注意力head包括多个算子，且该多个算子为对第一搜索空间包括的多个候选算子进行采样得到的
1201
基于该多个候选神经网络的性能，从该多个候选神经网络中选择目标神经网络
1202

图 12

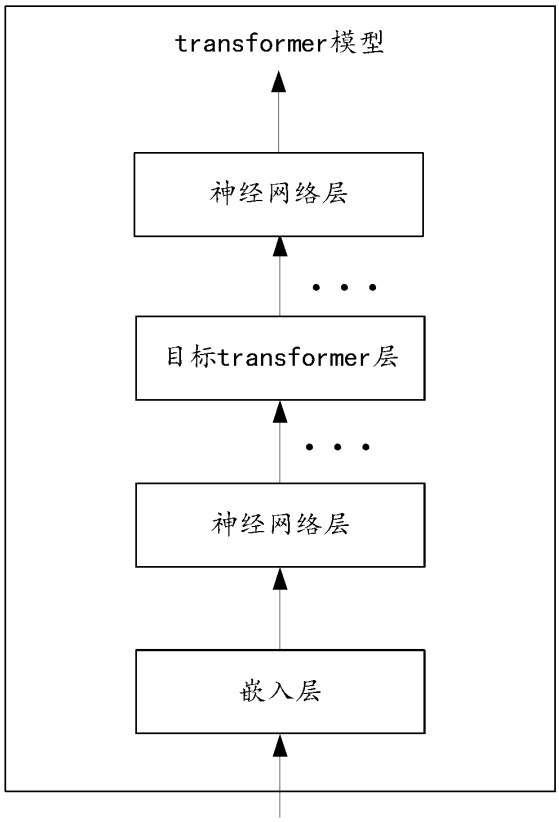

图 13

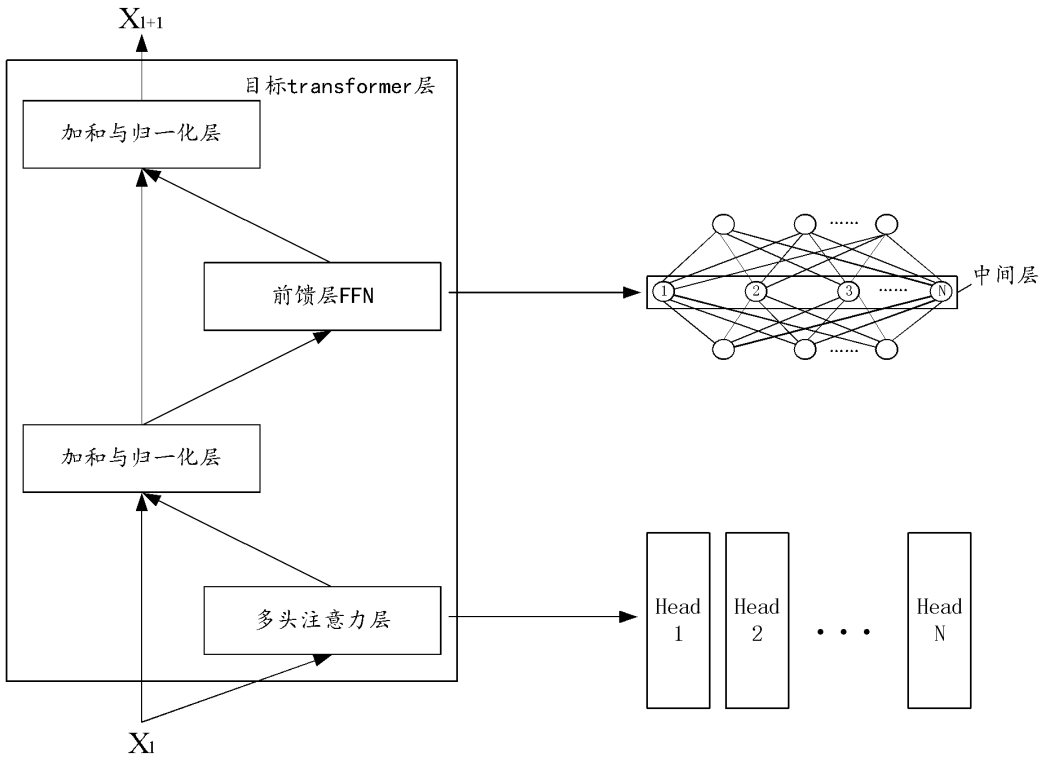
Xl+1
目标transformer层
加和与归一化层
前馈层FFN
加和与归一化层
多头注意力层
Xl
中间层
Head 1
Head 2
Head N

图 14

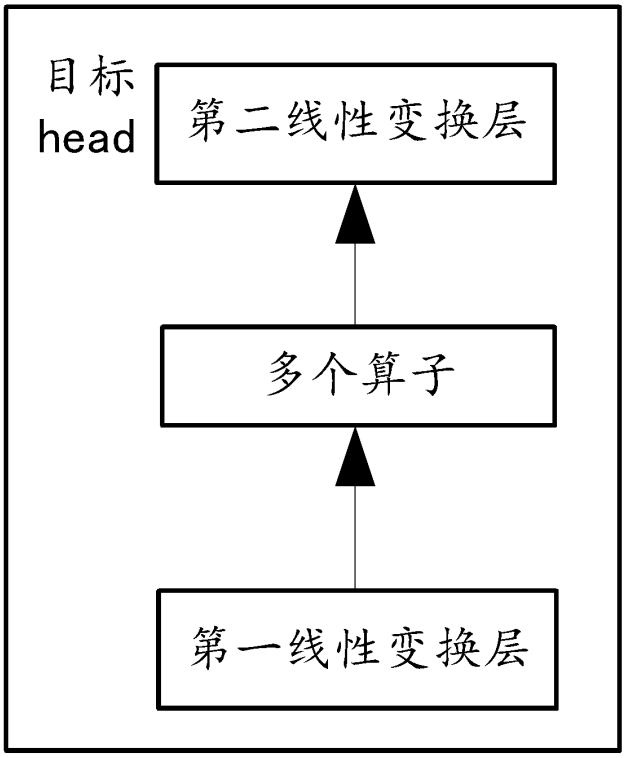

图 15

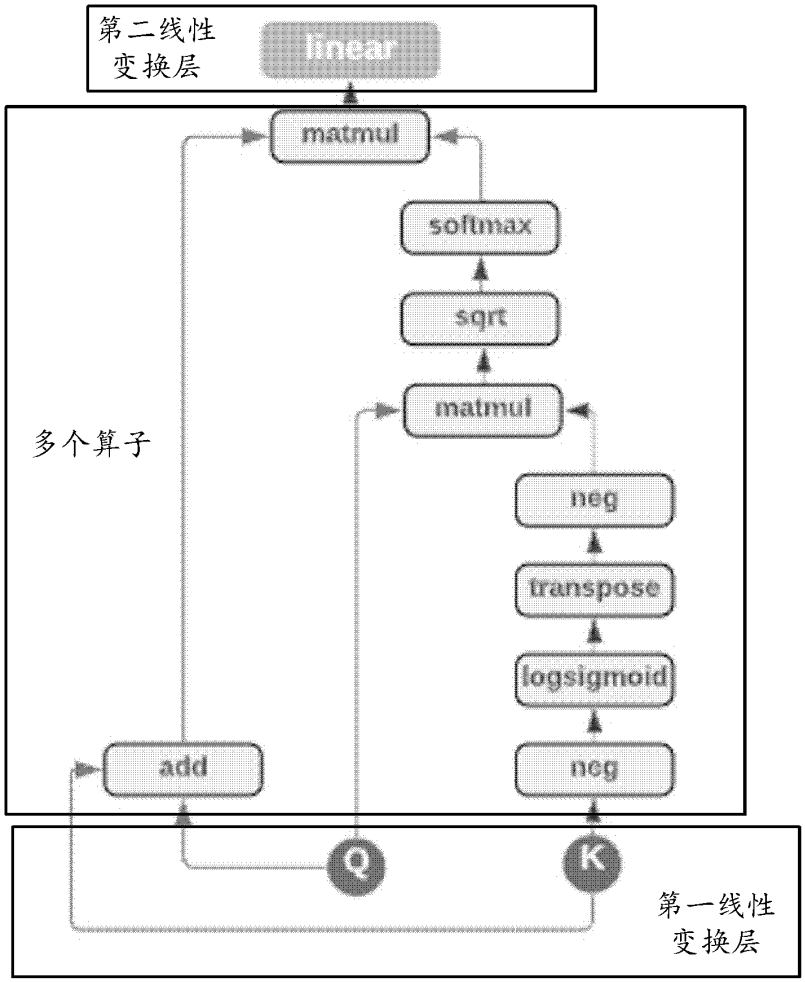
第二线性
变换层
linear
matmul
softmax
sqrt
matmul
多个算子
neg
transpose
logsigmoid
neg
add
Q
K
第一线性
变换层

图 16

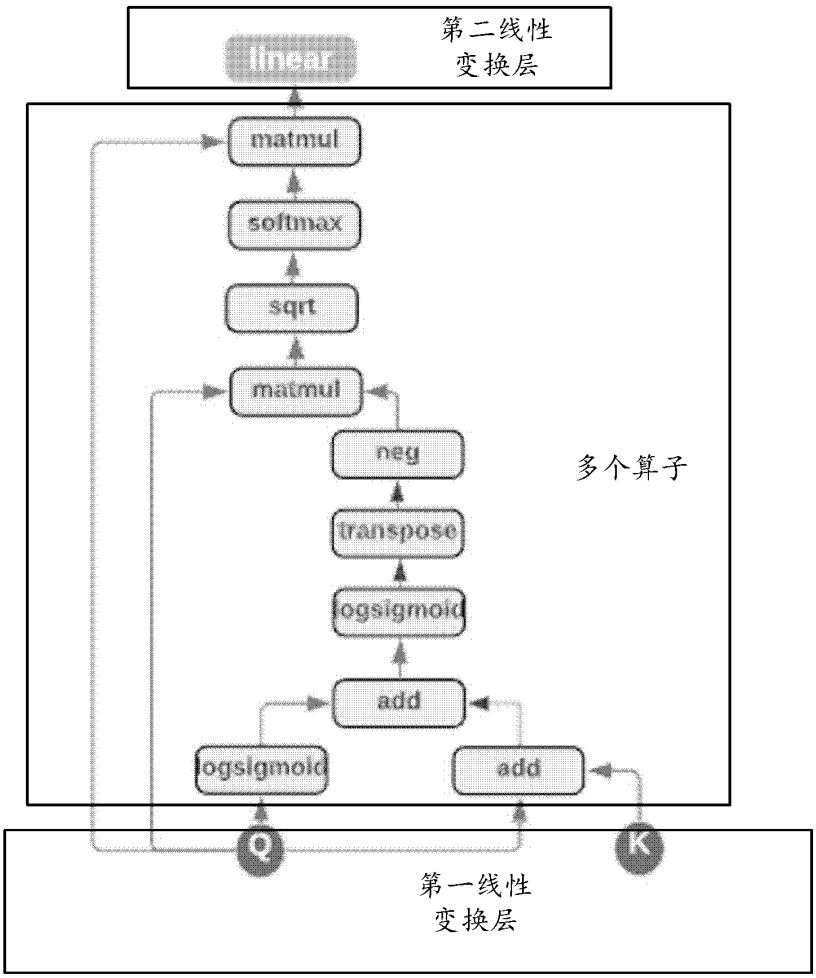
第二线性
变换层
linear
matmul
softmax
sqrt
matmul
neg
多个算子
transpose
logsigmoid
add
logsigmoid
add
Q
K
第一线性
变换层

图 17

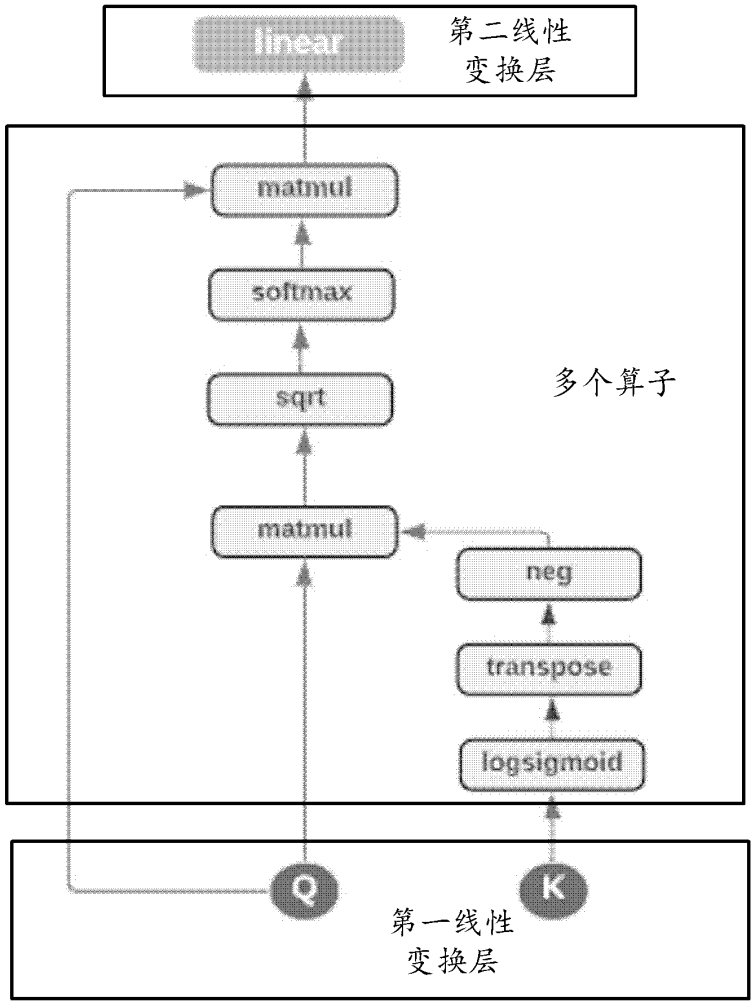
linear
第二线性
变换层
matmul
softmax
sqrt
多个算子
matmul
neg
transpose
logsigmoid
Q
K
第一线性
变换层

图 18

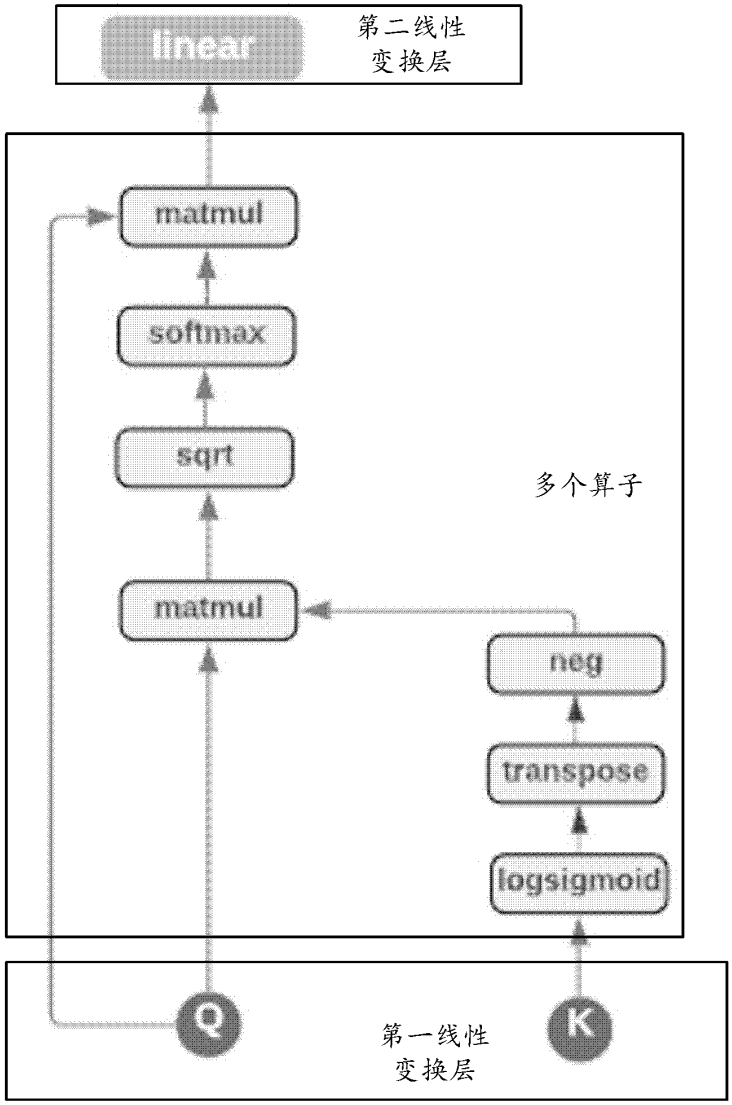
linear
第二线性
变换层
matmul
softmax
sqrt
多个算子
matmul
neg
transpose
logsigmoid
Q
K
第一线性
变换层

图 19

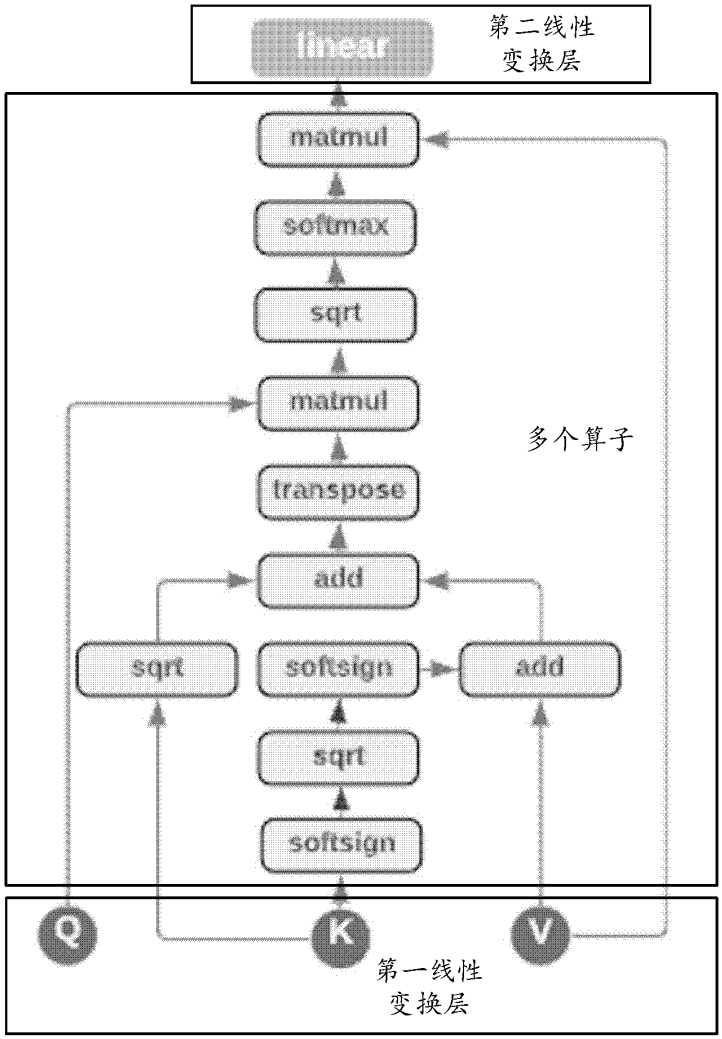
linear
第二线性
变换层
matmul
softmax
sqrt
matmul
多个算子
transpose
add
sqrt
softsign
add
sqrt
softsign
Q
K
V
第一线性
变换层

图 20

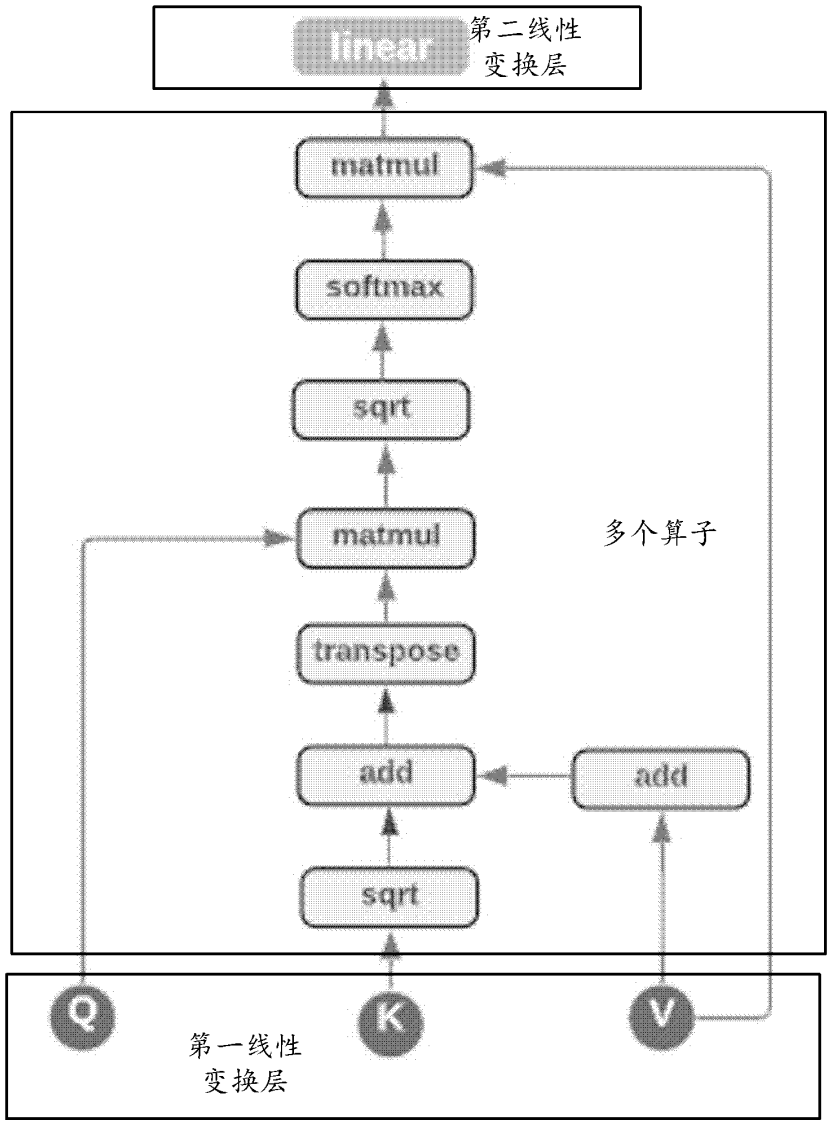
linear
第二线性
变换层
matmul
softmax
sqrt
matmul
多个算子
transpose
add
add
sqrt
Q
K
V
第一线性
变换层

图 21

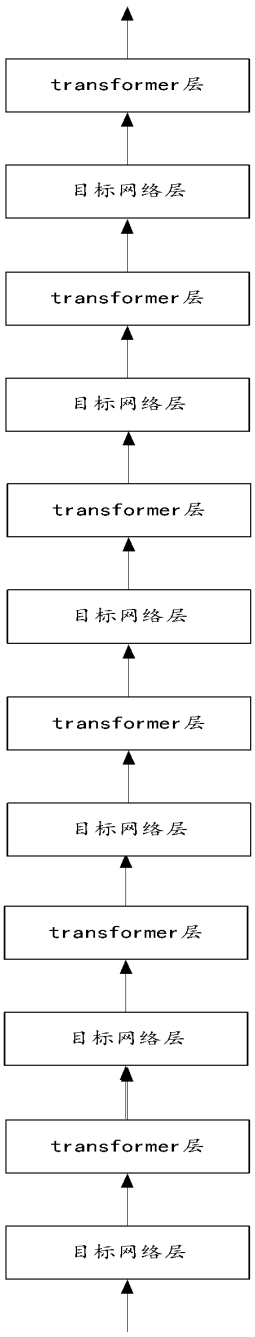
transformer层
目标网络层
transformer层
目标网络层
transformer层
目标网络层
transformer层
目标网络层
transformer层
目标网络层
transformer层
目标网络层

图 22

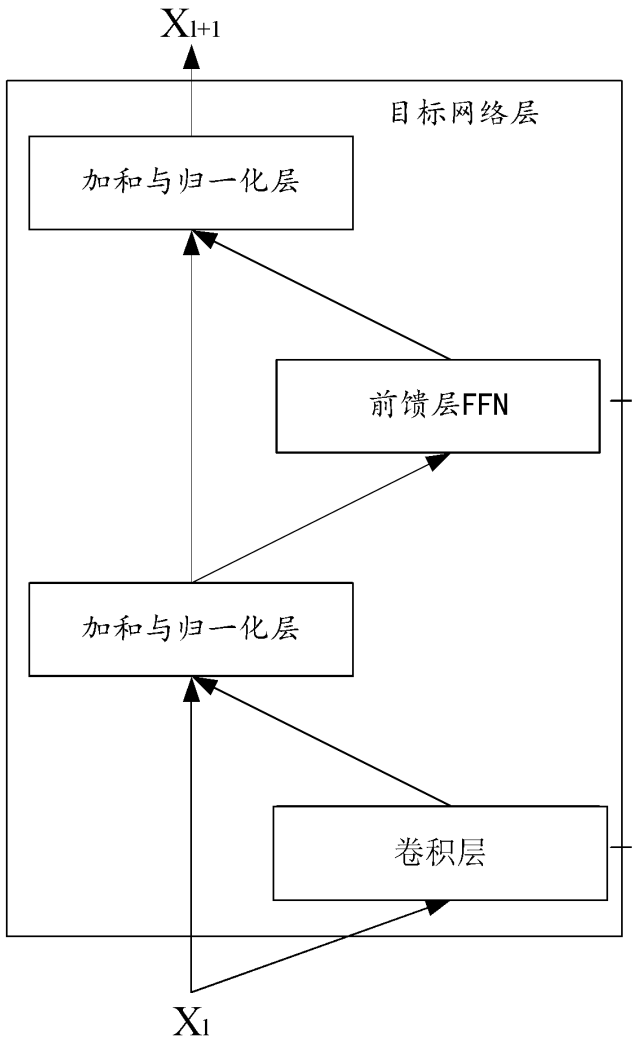
X_{l+1}
目标网络层
加和与归一化层
前馈层FFN
加和与归一化层
卷积层
X_l

图 23

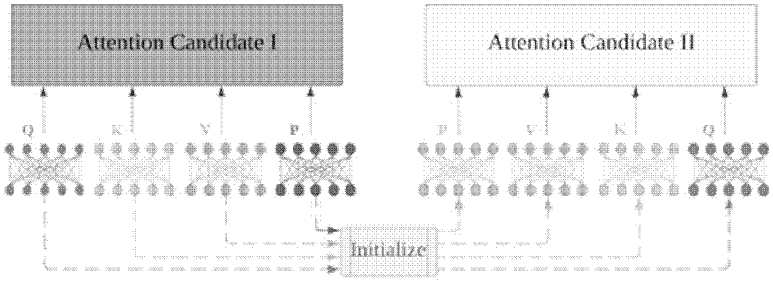
Attention Candidate I
Attention Candidate II
Q
K
V
P
Initialize

图 24

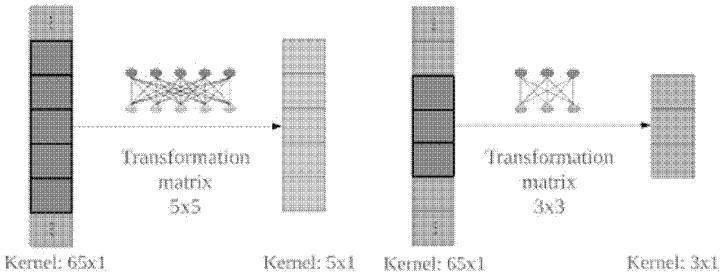
Transformation matrix 5x5
Transformation matrix 3x3
Kernel: 65x1
Kernel: 5x1
Kernel: 65x1
Kernel: 3x1

图 25

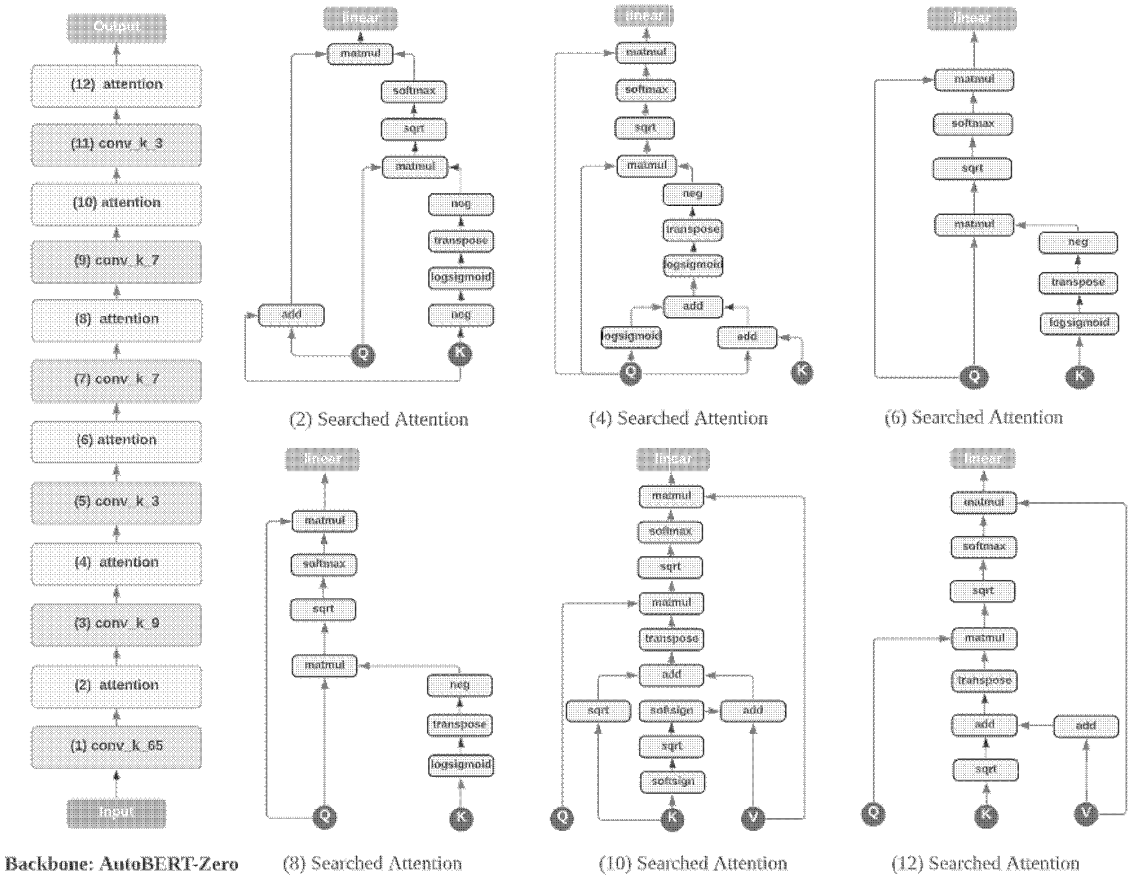
Output
(12) attention
(11) conv_k_3
(10) attention
(9) conv_k_7
(8) attention
(7) conv_k_7
(6) attention
(5) conv_k_3
(4) attention
(3) conv_k_9
(2) attention
(1) conv_k_65
Input
Backbone: AutoBERT-Zero
linear
matmul
softmax
sqrt
neg
transpose
logsigmoid
add
Q
K
V
softsign
(2) Searched Attention
(4) Searched Attention
(6) Searched Attention
(8) Searched Attention
(10) Searched Attention
(12) Searched Attention

图 26

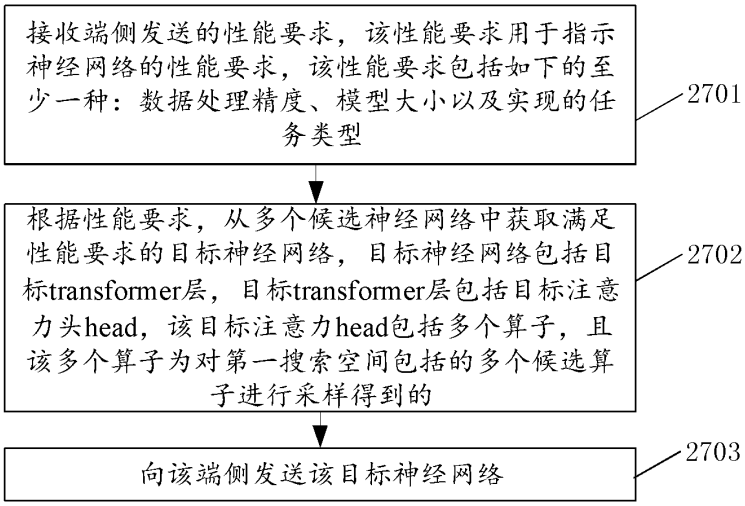
接收端侧发送的性能要求，该性能要求用于指示神经网络的性能要求，该性能要求包括如下的至少一种：数据处理精度、模型大小以及实现的任务类型
2701
根据性能要求，从多个候选神经网络中获取满足性能要求的目标神经网络，目标神经网络包括目标transformer层，目标transformer层包括目标注意力头head，该目标注意力head包括多个算子，且该多个算子为对第一搜索空间包括的多个候选算子进行采样得到的
2702
向该端侧发送该目标神经网络
2703

图 27

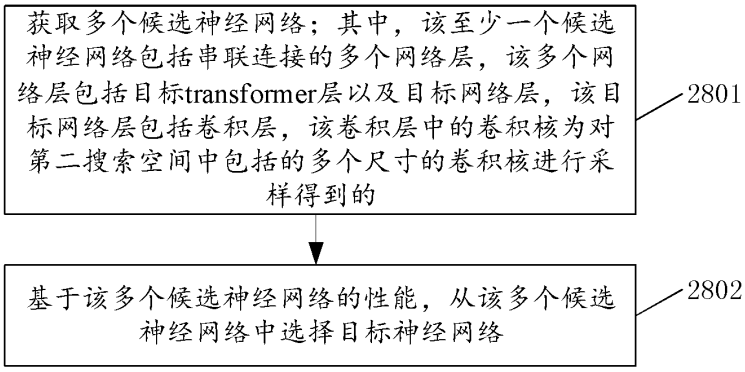
获取多个候选神经网络；其中，该至少一个候选神经网络包括串联连接的多个网络层，该多个网络层包括目标transformer层以及目标网络层，该目标网络层包括卷积层，该卷积层中的卷积核为对第二搜索空间中包括的多个尺寸的卷积核进行采样得到的
2801
基于该多个候选神经网络的性能，从该多个候选神经网络中选择目标神经网络
2802

图 28

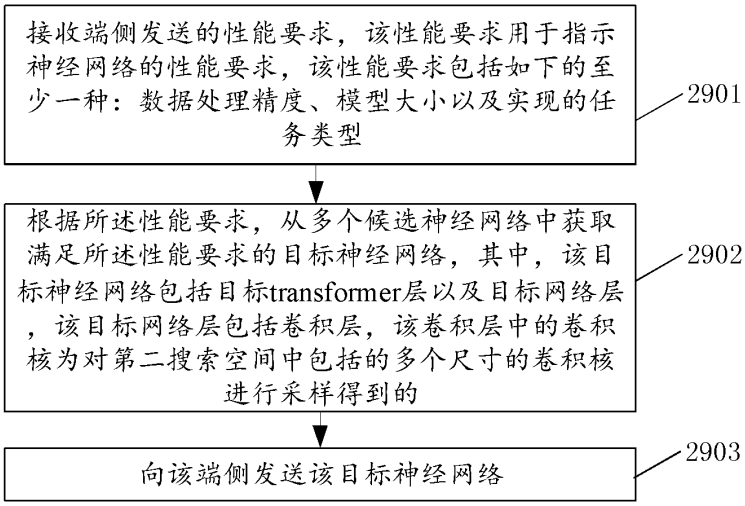
接收端侧发送的性能要求，该性能要求用于指示神经网络的性能要求，该性能要求包括如下的至少一种：数据处理精度、模型大小以及实现的任务类型
2901
根据所述性能要求，从多个候选神经网络中获取满足所述性能要求的目标神经网络，其中，该目标神经网络包括目标transformer层以及目标网络层，该目标网络层包括卷积层，该卷积层中的卷积核为对第二搜索空间中包括的多个尺寸的卷积核进行采样得到的
2902
向该端侧发送该目标神经网络
2903

图 29

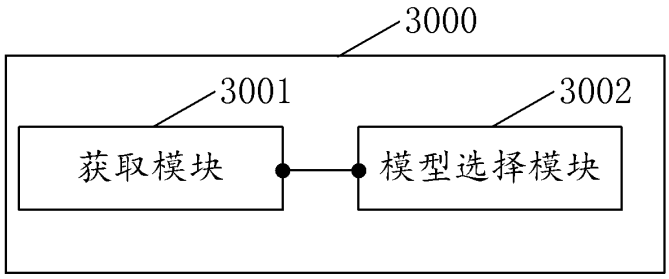
3000
3001
获取模块
3002
模型选择模块

图 30

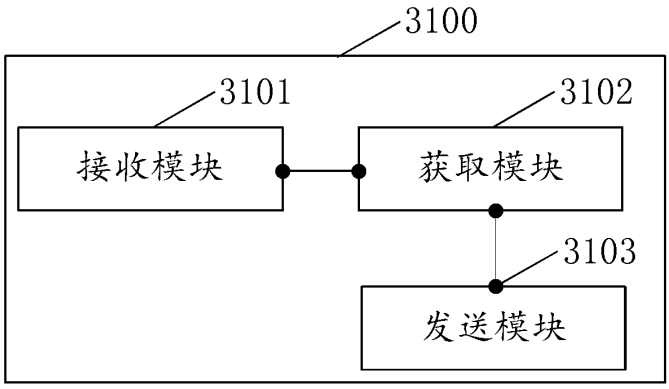
3100
3101
接收模块
3102
获取模块
3103
发送模块

图 31

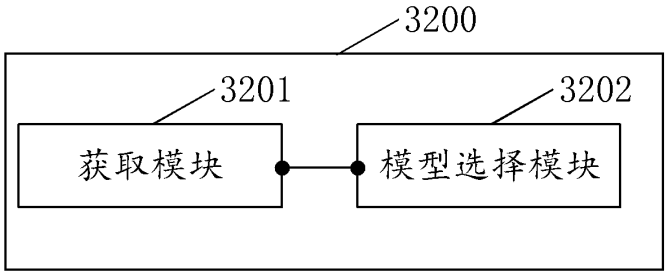
3200
3201
3202
获取模块
模型选择模块

图 32

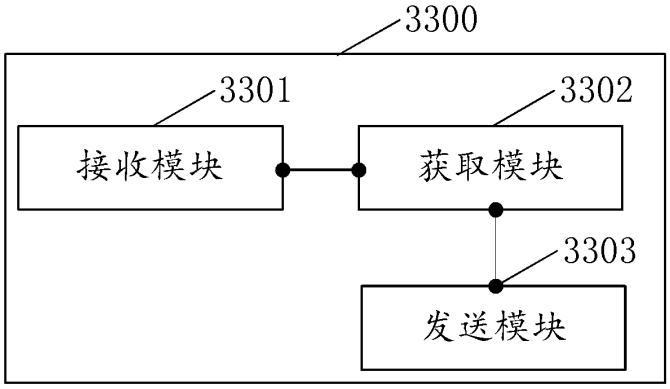
3300
3301
3302
接收模块
获取模块
3303
发送模块

图 33

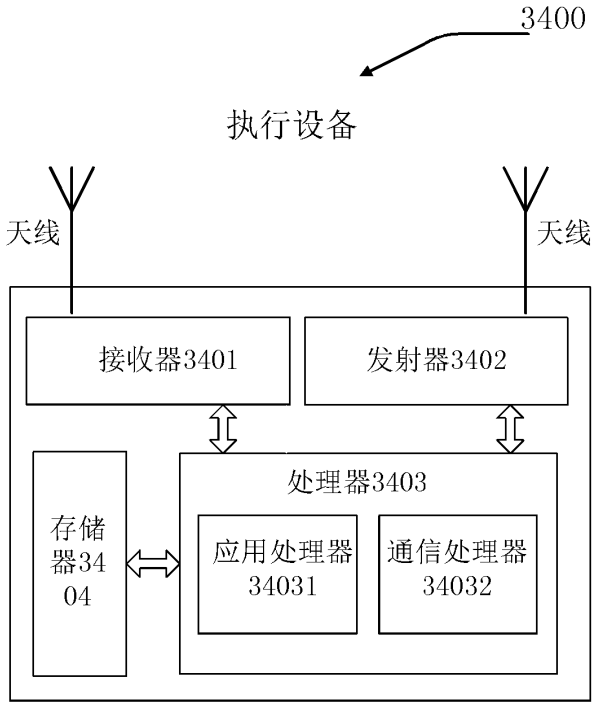
3400
执行设备
天线
天线
接收器3401
发射器3402
处理器3403
存储器3404
应用处理器34031
通信处理器34032

图 34

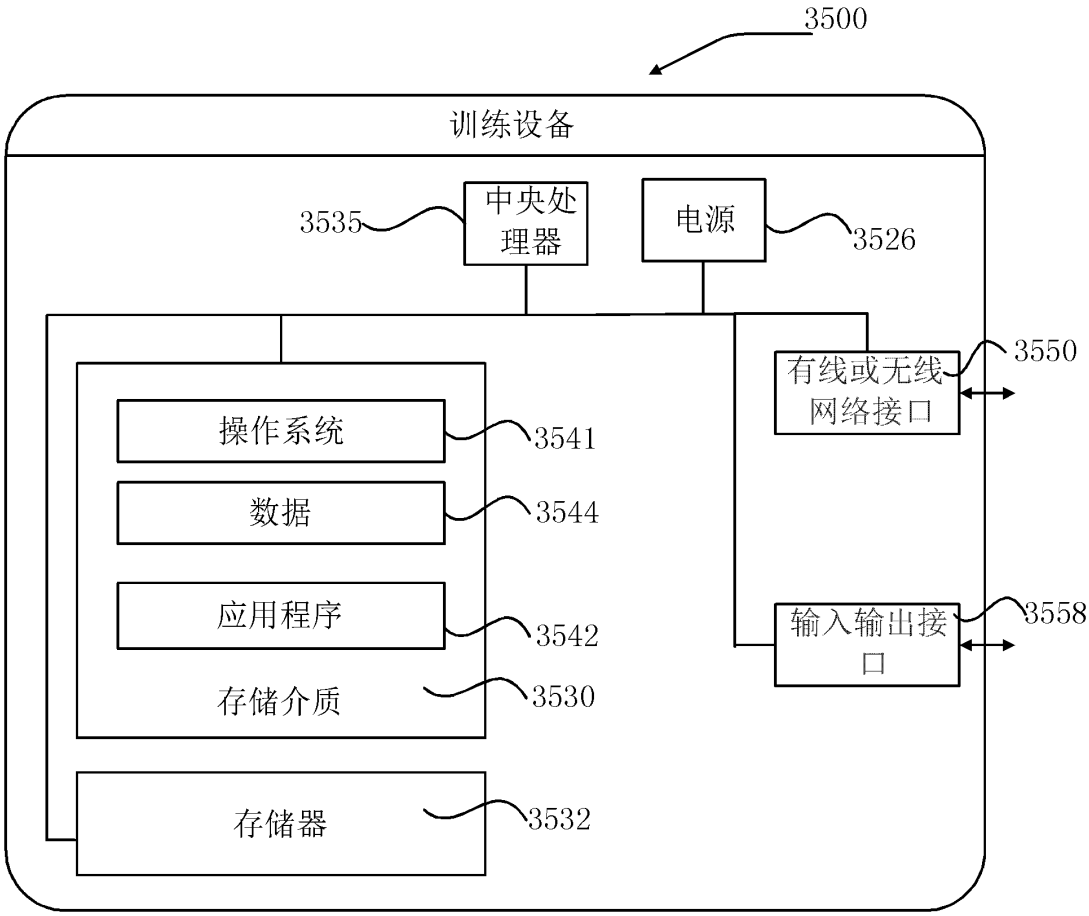
3500
训练设备
3535
中央处理器
电源
3526
有线或无线网络接口
3550
操作系统
3541
数据
3544
应用程序
3542
存储介质
3530
输入输出接口
3558
存储器
3532

图 35

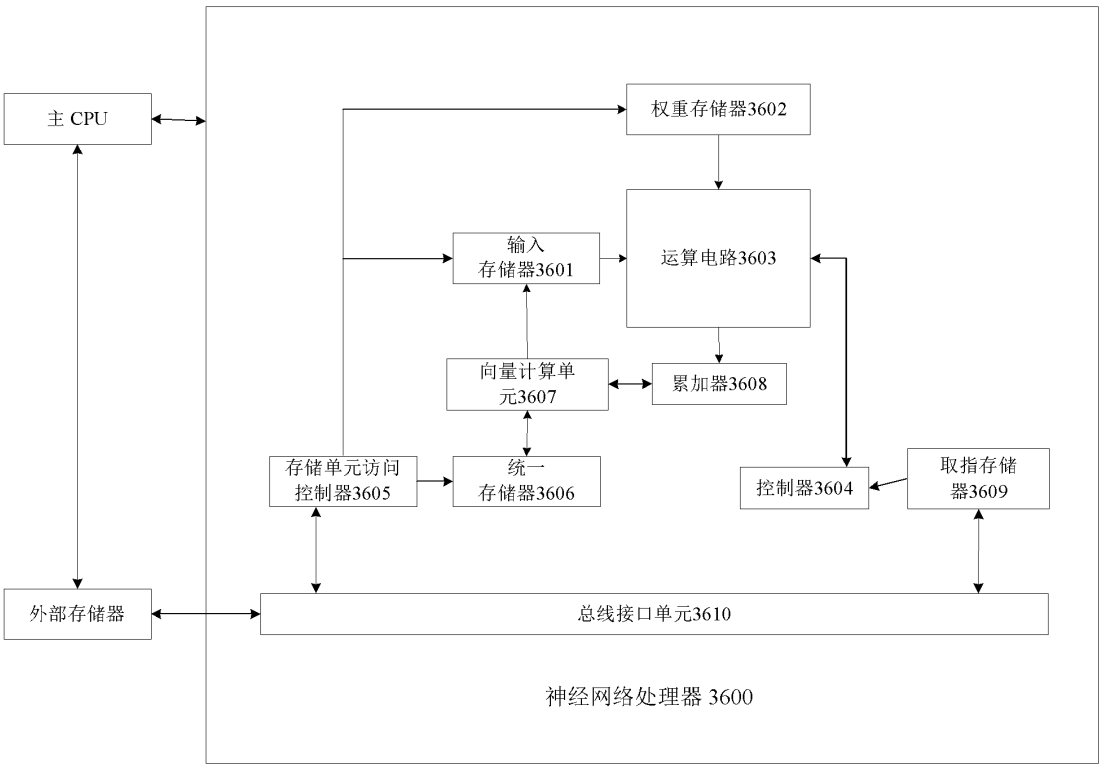
主 CPU
权重存储器3602
输入
存储器3601
运算电路3603
向量计算单
元3607
累加器3608
存储单元访问
控制器3605
统一
存储器3606
控制器3604
取指存储
器3609
外部存储器
总线接口单元3610
神经网络处理器 3600

图 36