Video Question Answering via Gradually Refined Attention over Appearance and Motion

在VideoQA中，问题可能涉及到视频的不同细节，答案不可能统一。为了得到答案，模型需要仔细的分析问题和关注视频的部分重要信息。

本文中，我们提出了一个端到端的视频问答模型。模型首先采样出视频一系列的帧和片段，从中抽取出纹理和运动特征。随后，模型逐词的读取问题并且利用帧级和片段级相互作用的特征来精炼模型的注意力。当所有问题的单词被处理时，模型产生最终最优的注意力，该注意力融合了纹理和运动特征做为视频的表征。在整个过程中，粗糙的问题特征和精细的单词特征都被利用。

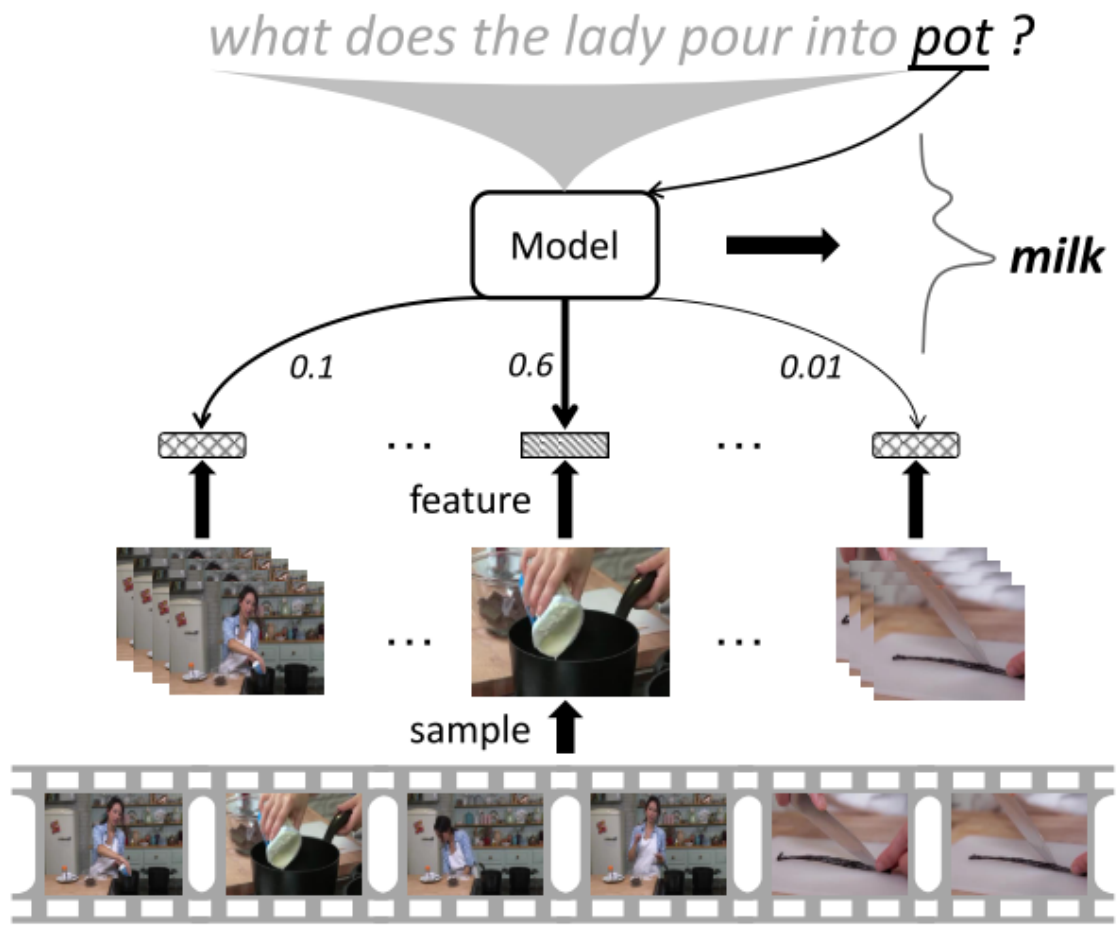

Figure 1: Given the video and the question, our model first samples the video in frame-level and clip-level, then extracts appearance and motion features, while the question is processed in word-level to refine the attention gradually. The numeric values beside the lines indicate the refined attention after the last word being processed.

本文的主要贡献如下：

我们提出利用基于视频的外观和运动信息来解决VideoQA任务。

我们使用粗糙的问题特征结合精细的单词特征做为引导提出精炼视频注意力的模型。

我们利用两个数据集来评估我们提出的模型。

方法

给出视频V和问题Q,VideoQA的目标是给出恰当的回答A。对于一个给定的视频，第一步先从视频中提取出外观和运动信息，随后逐词的分析问题并且在每个时步通过AMU精炼针对这些特征的注意力。问题的最后一个词被处理后，模型产生与视频最相关、最有价值的精炼注意力来回答特定问题。模型使用这个注意力来融合外观和运动特征，得到视频的表征。为了产生回答，其他如问题信息和注意力历史上下文信息也被用做推断。

Feature Extraction

Apperarance:我们的模型，采用VGG network 作为帧级外观特征的提取器。对于给定的视频，我们将它的外观特征表示为$F_a = [f_1^a, f_2^a, \ldots, f_N^a]$，$N$是视频中采样的帧数，$a$表明外观。

Motion:C3D network有捕获视频动态信息的能力，所以我们使用C3D network作为片段级运动特征的提取器。对于给定的视频，通过C3D network采样的片段被处理，采集出来的运动信息表示为$F_m = [f_1^m, f_2^m, \ldots, f_N^m]$，$N$是视频采样的片段数，$m$表明运动。

Question:问题表示为一系列的词汇标注$Q = [q_1, q_2, \ldots, q_T]$，我们使用embedding layer将$q_t$转化成embedding x_t。

AMU

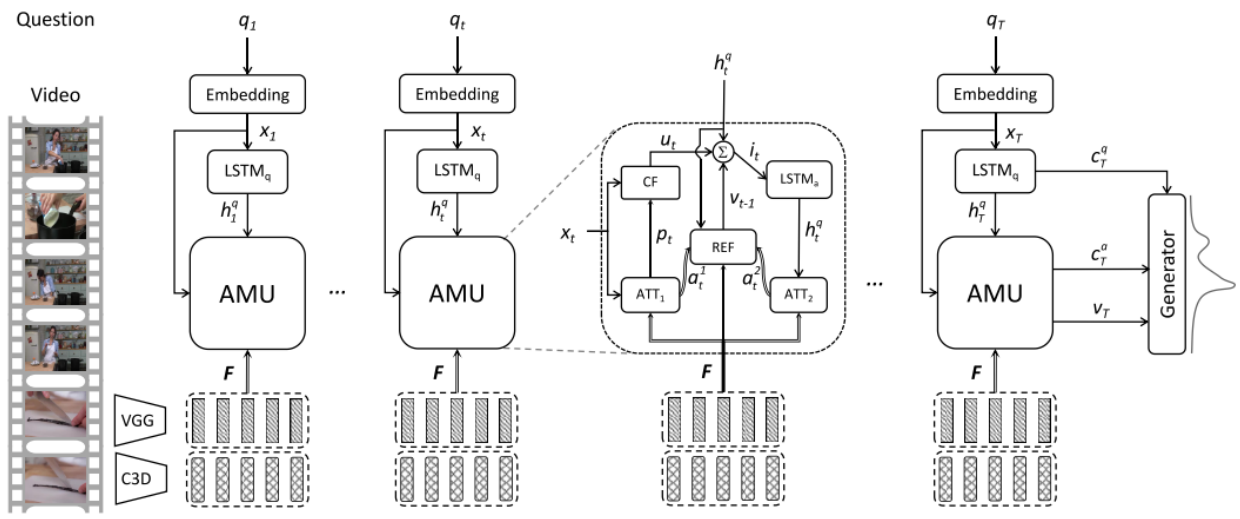

Figure 2: Our model transforms the words by the embedding layer and manipulates the attention in Attention Memory Unit (AMU). The model processes the question word by word while AMU generates and refines the attention over appearance and motion features of the video at each timestep. After all words are processed, the final attention is used to fuse both features as the representation of the video. Together with other contexts, our model outputs the answer. The double line in the figure indicates features in two channels.

问题的单词按顺序被处理，同时一个新颖的注意力机制在过程中被应用。模型首先使用embedding layer 将输入单词转化成embedding x_t，这保存当前单词的语义信息。embedding x_t随后喂入LSTM_q，这一步操作可认为是将处理过的问题信息保存下来。将x_t和h_t^q输入到AMU单元，针对外观和运动特征产生并精炼注意力。如图2所示，AMU将当前的word embedding，question information，and video features作为输入，随后对视频特征执行几轮精炼注意力。为了清楚起见，我们使用双线表示包含2个通道的特征。

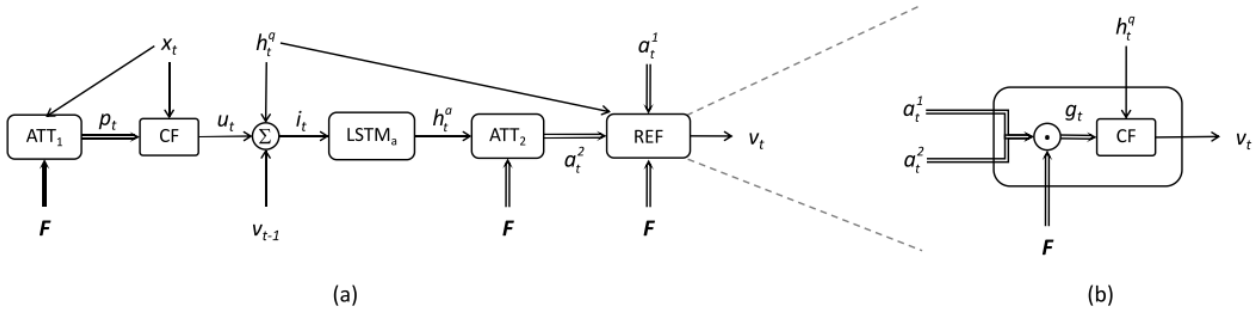

Figure 3: The operation blocks in AMU is unrolled based on its execution order from left to right in (a), and the details of operation REF is presented in (b). The double line in the figure indicates features in two channels.

AMU中包含4种操作

为了清楚起见，我们按照执行顺序，展开AMU的操作模块。

ATT$_1$:基于x_t对F执行初始化注意力并关注与当前单词的相关的视频特征。（即与当前词越相关的信息越容易保留下来）

CF:将外观特征p_t^a和运动特征p_t^m通过CF加权求和，CF给每个通道权重分数并得到中间变量融合表征u_t。

LSTM$_a$:h_t^q、之前产生的视频表征v_{t-1}和u_t相加做为LSTM$_a$的输入，LSTM$_a$保存了所有执行过的注意力操作。(包含了上一时步的信息)

ATT$_2$:基于F，ATT$_2$使用h_t^a执行第二次注意力机制。

REF:第一次注意力权重a_t^1和第二次注意力权重a_t^2在REF被精炼，并且产生视频表征v_t。

Attention

给定一个关于视频的问题，帧或者片段只有一个小的集合里，大多数时间是相关的。这些特征对于给出答案是有用的。注意力机制旨在分别给视频的外观和运动特征分配权重同时结合它们的权重得到有用的特征。AMU中有两个注意力操作ATT$_1$和ATT$_2$。以ATT$_1$为例，ATT$_1$利用x_t对视频特征F执行注意力机制。为了简单起见，我们省略了外观和运动的标注。每个特征通道，该操作都被执行。注意力机制公式表述如下：

$$e_i = tanh(W_f f_i + b_f)^T tanh(W_x x_t + b_x) \tag{1}$$

$$a_i = \frac{exp(e_i)}{\sum_{i=1}^{N} exp(e_i)} \tag{2}$$

权重a_i反应的是当前词与第i个特征的相关程度，W_f和W_x用来将word embeddings 和 video features转化到相同的潜在embeding space。通过a_i，计算融合特征p_t，公式如下：

$$p_t = \sum_{i=1}^{N} a_i tanh(W_f f_i + b_f) \tag{3}$$

p_t是此问题下当前单词的视频表征。当回答这问题时，p_t是ATT$_1$给予当前单词的影响力。之后ATT$_2$将使用h_t^a执行另外一个注意力操作并产生第二个注意力权重。

Channel Fusion

在得到p_t之后，实际上是由p_t^a和p_t^m组成，这两个特征融合形成中间视频表征u_t。因为问题中的单词可能和外观和运动的关联有不同的强度，模型使用当前的单词给每个通道特征分配权重并融合：

$$s_t^a, s_t^m = softmax(W_m x_t + b_m) \tag{4}$$

$$u_t = s_t^a p_t^a + s_t^m p_t^m \tag{5}$$

计算得到的关联强度分别为s_t^a,s_t^m，融合的表征u_t吸收了基于当前单词的视频外观和运动通道的信息。

Memory

我们使用LSTM$_a$去控制第二次注意力操作的输入并保存注意力历史。h_t^q、之前产生的视频表征v_{t-1}和u_t求和做为LSTM$_a$的输入，产生的h_t^a用来作为ATT$_2$的输入。

Refine

在执行完ATT$_2$后，模型产生了基于F的第二个注意力权重a_t^2。两次注意力权重用来精炼注意力。公式如下：

$$a_t = (a_t^1 + a_t^2)/2 \tag{6}$$

$$g_t = \sum_{i=1}^{N} a_t^i tanh(W_f f_i + b_f) \tag{7}$$

$$v_t = CF(h_t^q, g_t) \tag{8}$$

g_t实际上包括来自外观和运动的g_t^a和g_t^m，v_t是时步t的最终视频融合表征。

上述注意力机制过程中，模型使用了精确的词汇信息和粗糙的问题信息来逐步精炼视频外观和运动特征的注意力。当前词嵌入的注意力能够提升藏在问题向量特征中的关键词信息，当融合这些特征时，问题信息可以给出更加通用的指导。在AMU处理完所有问题的词汇时，回答问题最相关的、最显著的精炼视频表征随之产生。

Answer Generation

在时步T，问题的最后一个词被处理之后，我么得到视频融合表征v_T。我们有另外两个上下文信息。问题LSTM$_q$得到的包含问题信息的记忆向量c_T^q和AMU得到的包含注意力过程的记忆向量c_T^a。

我们能够准备好预先定义好的回答集，并产生一个简单的softmax分类器。回答公式如下：

$$answer = arg\,max\ softmax(W_g(W_x c_T^q . c_T^a . v_T)) \tag{9}$$

回答也能够通过LSTM network产生。问题信息c_T^q和注意力过程c_T^a被用来初始化LSTM network，同时精炼视频表征v_T用来作为它的第一个输入。回答的每个词可以像上述公式一样产生。

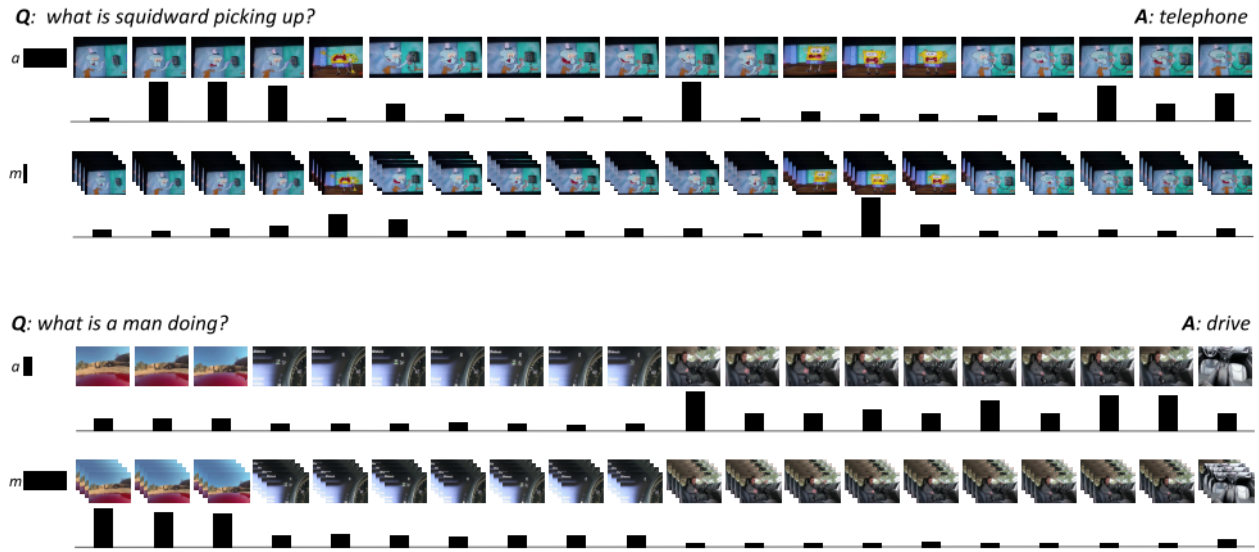

Figure 5: Visualization of the attention for two examples. *a* stands for appearance and *m* stands for motion.