Neural Attention Reader for Video Comprehension

Ashish Gupta
IIIT Bangalore
Bangalore, India
ashish.gupta@iiitb.org

Rishabh Mehrotra
Spotify Research
London, UK
rishabhm@spotify.com

Manish Gupta
IIIT Bangalore, VideoKen Inc.
Bangalore, India
manish.gupta@videoken.com

ABSTRACT

Despite the increasing availability of informative video content, question answering on videos remains an under-researched and challenging topic. Owing to the free-flowing nature of the verbal content, long duration of videos and lack of clear demarcations on where the context is changing, answering questions from video transcripts remains challenging. We consider the problem of extracting answers for a given question from a pool of videos and propose a novel gated neural attention architecture with content bifurcation module (GABiNet) to infer answers from video content using transcript data. The proposed GABiNet model is efficient enough to consider a large number of candidate videos and jointly learns the question and content representation by incorporating question information into content representation. To deal with the lack of demarcation issue, we propose a number of content bifurcation techniques which enable the neural model to divide the transcript text into different meaningful chunks to enable tractable inference of answers. Based on experiments on a large dataset of educational videos, we investigate the benefits offered by the gating, attention and bifurcation mechanisms and demonstrate significant performance gains over a number of established baselines and state-of-the-art QA(Question Answering) techniques. We contend that our work is among the first to tackle open-domain question answering on video content, and our findings have implications for the design of video-based QA systems.

KEYWORDS

Question answering, Attention model, Content Bifurcation, Video comprehension, Neural model

1 INTRODUCTION

Educators have been recording instructional videos for nearly as long as the format has existed. In the past decade, though, free online video hosting services such as YouTube have enabled people to disseminate instructional videos at scale. For example, Khan Academy videos have been viewed over 300 million times on YouTube [1]. Given the prevalence of such informative video content, an increasing amount of users engage with these videos to find answers. Despite substantial progress in open-domain question-answering (QA), answering questions from video remains challenging. While transcripts are readily available for such video content, given the free flowing and unstructured nature of verbal content, long duration of videos and the lack of clear demarcations on where the context is changing, it becomes prohibitive to find the relevant parts of the video containing the information to answer a given question.

This paper considers the problem of answering both factoid and non-factoid questions in an open-domain setting. Most recent work on question answering techniques has focused on textual content, and the proposed techniques typically assume that a short piece of relevant text is already identified and given to the model. This is not a realistic setting for video QA systems since identifying the right video from a large collection of video is prohibitively challenging. Furthermore, the existing techniques assume demarcation of content into well defined passages (paragraphs, or sections), which is not the case for video transcripts, which makes it harder to use existing techniques for video QA.

In this work, we describe a key research challenge faced during the implementation of a real world video based question answering system currently running on live traffic. We consider the above mentioned challenges with existing QA systems, and propose a novel gated attention based sequential network with content bifurcation module (GABiNet) to find the answer to a given question among a large set of videos. The GABiNet model is composed of five components: (i) a content bifurcation module to divide the video transcript into meaningful chunks, (ii) recurrent network encoder to build representation of questions and chunks, (iii) gated matching layers to match the question and video, as well as to match the question and chunk, (iv) a self-matching layer to aggregate information from a chunk and (v) a pointer network based answer boundary prediction layer.

We investigate a number of different techniques for content bifurcation, including (i) content based, (ii) temporal based, (iii) transition Points and (iv) topical drift convolution. The attention mechanism accounts for the fact that words in the passage are of different importance to answer a particular question. We learn a question-aware representation for each bifurcated chunk, and employ the pointer network for detecting answer spans. The self-matching mechanism along with the pointer network allows the model to aggregate evidence from the whole chunk to infer the answer boundary. Based on a dataset of 15,298 educational videos and their transcripts, we evaluate the performance of the proposed neural model and compare with a number of baselines and state-of-the-art question answering architecture on over 2000 QA pairs generated by domain experts.

Our results show that the proposed GABiNet architecture is able to achieve an F1 score of 0.328, which is a gain of over 14% relative to the best performing baseline(*Topical drift* 0.288). Additionally,

[1] Khan Academy YouTube Channel: http://www.youtube.com/user/khanacademy/about

KDD'18 Deep Learning Day, August 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Keyword	Description
QA	Question Answering
KB	Knowledge Base
SQuAD	Stanford Question Answering Dataset

Table 1: List of abbreviations

we demonstrate that *Topical Drift Convolution* chunking technique performs the best with F1 of 0.288, with over 16% improvement over other bifurcation techniques (*Transition Points* 0.249). Furthermore, comparisons with variants of the proposed model demonstrate the utility and impact of the gating and attention mechanism.

In summary, we make the following contributions:

- We propose a gated attention based architecture (GABiNet) for the problem of open domain question answering, which, to the best of our knowledge, is the first attempt at end-to-end open-domain question answering in the video domain.
- We propose a content bifurcation module with four different techniques, which allows our model to deal with the issues around free-flowing nature of video transcripts.
- While most existing QA systems assume that the relevant text containing the answer is already given, we instead introduce a dual gating mechanisms at video and chunk level, which allows the model to select from a large number of candidate videos, and enables it to automatically select the right chunk (content passage) to select answers from.
- The proposed GABiNet model yields state-of-the-art results against established strong baselines, with over 20% improvement.
- Lastly, we contribute a labeled dataset of over 2000 question answer pairs on video data, which is among the first datasets available on the topic for benchmarking and research purposes.

Apart from being one of the first studies to investigate question answering on a video transcript dataset, our findings have implications on the design on scalable video based QA systems.

2 RELATED WORK

Question Answering was originally defined as finding answers in collections of unstructured documents, following the setting of the annual TREC competitions http://trec.nist.gov/data/qamain.html. With the development of KBs(Knowledge Base), many recent innovations have occurred in the context of QA(Question Answering) from KBs with the creation of resources like WebQuestions Berant et al.[3] and SimpleQuestions Hill et al. [14] based on the Freebase KB Bollacker at al. [4], or on automatically extracted KBs, e.g., OpenIE triples and Fader [10].

The subfield of *machine comprehension*, i.e., answering questions after reading a short text or story, has made considerable progress recently, thanks to new deep learning architectures like attention-based and memory augmented neural networks [1, 5, 11] and release of new training and evaluation datasets like CNN/Daily Mail based on news articles [12]), CBT based on children books [14], or SQuAD [25] and WikiReading [13], both based on Wikipedia. Work done by Chen et al. [6] and Wang et al. [32] also targeted Wikipedia text, which is quite clean and highly structured compared to video transcripts. Significant progress has been made on question answering for datasets like Wikipedia and SQuAD [2, 7, 8, 16, 22, 27]. An objective of our paper is to test how such new methods can perform in a closed domain of educational videos.

Along with close-style datasets, several powerful deep learning models have been introduced to solve this problem [6, 8, 9, 12, 14, 17, 28–30]. Hermann et al.[12] first introduce attention mechanism into reading comprehension. Hill et al.[14] propose a window based memory network for CBT dataset. Kadlec et al.[17] introduce pointer networks with one attention step to predict the blanking out entities. Sordoni et al.[29] propose an iterative alternating attention mechanism to better model the links between question and passage. Trischler et al.[30] solve cloze-style question answering task by combining an attentive model with a reranking model. Dhingra et al.[9] propose iteratively selecting important parts of the passage by a multiplying gating function with the question representation. Cui et al.[8] propose a two-way attention mechanism to encode the passage and question mutually. Shen et al.[28] propose iteratively inferring the answer with a dynamic number of reasoning steps and is trained with reinforcement learning.

Neural network-based models demonstrate the effectiveness on the SQuAD dataset. Wan et al.[32] combine match-LSTM and pointer networks to produce the boundary of the answer. Xiong et al.[34] and Seo et al.[27] employ variant coattention mechanism to match the question and passage mutually. Xiong et al.[34] propose a dynamic pointer network to iteratively infer the answer. Yu et al.[38] and Lee et al.[20] solve SQuAD by ranking continuous text spans within passage. Yang et al.[36] present a fine-grained gating mechanism to dynamically combine word-level and character-level representation and model the interaction between questions and passages. Wang et al.[33] propose matching the context of passage with the question from multiple perspectives.

Video based question answering remains an under-researched area, with prior work done on investigating QA on news videos Yang et al.[35], predicting future events using images from videos Zhu et al.[39] and Li et al.[21]. Despite these efforts, this work is one of the first efforts to investigate open domain machine comprehension for QA on video transcripts.

3 GABINET ARCHITECTURE FOR VIDEO QA

To extract answers from video content, we work with transcript data and propose a gated attention based neural model with content bifurcation module to break down a large piece of transcript text into small chunks and then perform QA on those chunks. The proposed GABiNet model is a composite architecture with different modules responsible for different stages of the answer selection process. For each question, the video retriever (Section 3.1) extracts the top-k videos and tags them with transcripts. For each retrieved video, the content bifurcation module (Section 3.2) extracts meaningful chunks which are then passed onto the gated attention module (Section 3.3). Finally, the answer generation module considers each chunk and finds the answer boundaries and outputs its confidence score for each chunk. The chunk ranking module (Section 3.4) considers all chunks from all videos and ranks the answers based on the model's confidence scores and outputs the final answer. We

Chunk:- I will not comment on the theological implications of this assertion.**The first theoretical calculation of pi was carried out by Archimedes, a great Greek mathematician from Syracuse**, that was about somewhere around 250 BC. And he said that pi was somewhere between 223 divided by 71, and 22 divided by 7. This was amazingly profound. He knew he didn't know what the answer was, but he had a way to give an upper and a lower bound, and say it was somewhere between these two values.

Question:- Who carried out first theoretical calculation of pi?
Answer:- Archimedes
Predicted:- Archimedes

Table 2: An example from our YouTube video data with chunking.

next describe each module in detail. Figure 1 visually describes the different component of the proposed GABiNet architecture.

It noteworthy to mention that for informational video content, in addition to transcript data, raw audio features are also available. We conducted preliminary experiments with raw audio features wherein we convert the sound track into transcripts by extracting MFCC features [19] from audio data from the video and applied enhancement method (phonetic-based transcript error correction). Our preliminary experiments using these features did not perform well, so we preferred working with transcript information for all experiments. We leave integrating audio features as future work.

3.1 Preprocessing & Video Retriever

Given a question and a large pool of videos, we employ an inverted index based retrieval module to first narrow our search space and focus on *reading* only those videos that are likely to be relevant. We use an industrial transcript generator to extract textual transcript for all videos considered. For each sentence in the transcript, we additionally tag the sentence with the corresponding time of utterance. A simple inverted index lookup followed by term vector model scoring performs quite well on this task for many question types, compared to Okapi's BM25 and cosine distance in word embedding space. Chunks of transcript text and questions are compared as TF-IDF weighted bag-of word vectors. We further improve our system by taking local word order into account with n-gram features(especially bigram). To speed up the retrieval process, we hash the transcripts and videos to the database which makes the process fast and memory efficient.

The proposed GABiNet model consumes the top-k retrieved videos with the aim of finding not only the right video but also inferring the correct answer in the video transcript. The selected top-k videos for each question are used throughout the paper for the different experiments.

3.2 Content Bifurcation Module

One key difference between traditional textual content and video transcript is the lack of demarcation of content into meaningful paragraphs or sections. The longer the duration of the video, the longer the transcripted text. To enable tractable inference of answers, we introduce a content bifurcation module with the aim of dividing the long transcripted text into smaller meaningful chunks. We consider a number of bifurcation techniques:

(i) Content based: In this approach, we divide the transcript into equal-sized chunks based on the text size. Each of these chunks are then fed into the neural model to generate answers.

(ii) Temporal based: Often, different concepts are discussed at different timestamps in a video. To capture this insight, we segment the transcript into equal-sized chunks based on time, which are then fed to the neural model to generate answers.

(iii) Transition Points: The start of a new concept is often signified by a transition statement. In this approach, we adopt a sliding window over the sentences of the transcript to detect such transition points and find segmentation boundaries based on the differences in the content of adjacent sentences. We run a sliding window over the sentences, find cosine similarity between each sentence pair in a sequential manner and find transition indexes based on:

$$idx = argmin\ cossim(s_i, s_{i+1}) \tag{1}$$

where *cossim* is cosine similarity between sentences and *idx* is the index where we can break the transcript and collectively merge sentences in a continuous fashion within indices.

(iv) Topical Drift Convolution: We go beyond pairwise sliding windows, and find topical drifts in the video content using a convolutional operator which considers sequential groups of sentences to detect changes in the topical content. We convolute over a group of sequential sentences with a sliding window approach and use a similarity threshold to detect sentence groups which are topically different to warrant a separate chunk.

$$idx = argmin\ cossim(g_i, g_{i+1}) \tag{2}$$

$$g_i = concat\ \{...., s_{i-2}, s_{i-1}, s_i\} \tag{3}$$

$$g_{i+1} = concat\ \{s_{i+1}, s_{i+2}, ..., s_k\} \tag{4}$$

where *concat* is used to group sentences $s_1, s_2, ..., s_{i+1}$, k is the *idx*(this we found in above method of **Transition Points**) of the last sentence if this is end of file or is *idx* of last sentence of a topic where the topic groups are selected based on some threshold value of dissimilarity between sentences. Similar to above method, we use

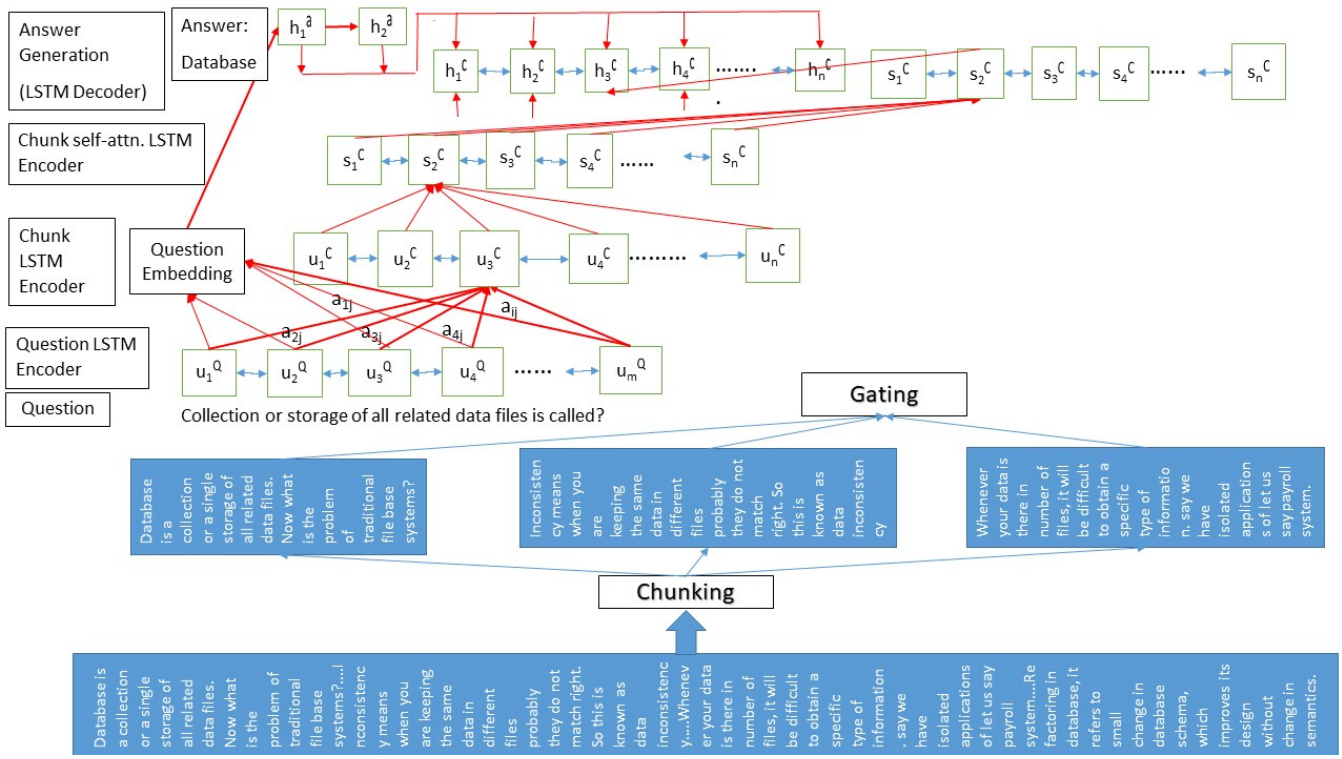

Figure 1: Overview of the proposed GABiNet architecture.

these groups of chunks and apply the sliding window approach to separate the not-so-correlated groups from the correlated groups of sentences. The gated attention model works on the chunks output from the content bifurcation module.

3.3 Gated Attention based Neural Matching Model

Given a question and a bifurcated chunk with tokens $\{q_1, q_2, ..., q_m\}$ and $\{c_1, c_2, ..., c_n\}$ respectively, we develop a gated attention LSTM model which we apply on the chunk, to learn question representation and question-aware chunk representation, which are used to detect answer spans. We next describe the neural matching model in detail.

3.3.1 Question & Chunk Encoding. We begin by converting all tokens of chunk and question text into their word embeddings ($\{e_t^C\}_{t=1}^n$ and $\{e_t^Q\}_{t=1}^m$). We used 300-dimensional Glove word embeddings Pennington et al.[24] trained from 840B Web crawl data. In addition to the embeddings, we consider few hand-crafted features. We convert the tokens in chunks and question to their lowercase and lemmatized form and check whether they can be matched. This basic feature turns out to be extremely helpful. We additionally extract normalized TF(Term Frequency), POS(Part of Speech) and named entity tags for c_i and q_i. These features are concatenated for each token of chunk and question.

After converting the words in question and chunk to their word-level embeddings ($\{e_t^Q\}_{t=1}^m$ and $\{e_t^C\}_{t=1}^n$), we used bi-directional LSTM Hochreiter et al.[15] to produce representation $u_1^C, ..., u_n^C$ and $u_1^Q, ..., u_m^Q$ for all words in the chunk and question respectively.

$$u_t^Q = BiLSTM(u_{t-1}^Q, e_t^Q) \tag{5}$$

$$u_t^C = BiLSTM(u_{t-1}^C, e_t^C) \tag{6}$$

The question and chunk embedding thus obtained are used as their representation for all further steps.

3.3.2 Gated Attention based neural model. We propose a gated attention-based recurrent network to incorporate question information into chunk representation. It is a variant of attention based recurrent networks, with an additional gate to determine the importance of information in the given chunk regarding a question. For a particular question and chunk representation ($\{u_t^Q\}_{t=1}^m$ and $\{u_t^C\}_{t=1}^n$), RocktÃđschel et al.[26] proposed generating sentence-pair representation $\{s_t^C\}_{t=1}^n$ by aligning words in question and chunk as follows:

$$s_t^C = BiLSTM(s_{t-1}^C, c_t) \tag{7}$$

where $c_t = att\,(u^Q, [u_t^C, s_{t-1}^C])$ is attention vector of question u^Q:

$$k_j^t = s^T ReLU(W_u^Q u_j^Q + W_u^C u_t^C + W_s^C s_{t-1}^C) \tag{8}$$

$$a_i^t = exp(k_i^t) / \sum_{j=1}^{m} exp(k_j^t) \tag{9}$$

$$c_t = \sum_{j=1}^{m} a_i^t u_i^Q \tag{10}$$

where Equation (8) performs a ReLU nonlinearity on top of representation of chunk and question. Equation (9) computes the attention score (a_i^t) which captures the similarity between chunk's token c_i and question's token q_j. This is a soft match between the chunk's token and the question's token. These features add soft alignments between similar but non-identical words. Finally, c_t describes aligned embedding of question or rather attention pooled vector of question u^Q.

3.3.3 Chunk Self-Attention Encoder. One problem with the above representation is that it has a very limited knowledge of the context. Moreover, the chunk embedded in the above encoder layer is syntactically divergent to the question given as input. To solve this problem, we propose direct matching the question-aware passage representation with itself. It encodes evidence from question and passage matching words into a new chunk representation h_t^C:-

$$h_t^C = BiLSTM(h_{t-1}^C, [s_t^C, c_t]) \tag{11}$$

where $c_t = att\,(s^C, s_t^C)$ is attention vector of s^C:

$$k_j^t = s^T ReLU(W_s^C s_j^C + W_s^C s_t^C) \tag{12}$$

$$a_i^t = exp(k_i^t) / \sum_{j=1}^{m} exp(k_j^t) \tag{13}$$

$$c_t = \sum_{j=1}^{m} a_i^t s_i^C \tag{14}$$

3.4 Answer Generation & Chunk Ranking

For a given question and bifurcated chunk, we compute probability distribution of the start and end positions of the probable answer in the chunk (P(start) & P(end)) depending on the degree of relevance between chunk's token c_i and the question and then produce an output vector $\mathbf{o}$ which is a weighted combination of that chunk embedding c_i:

$$P_{start}(i) = softmax_i\ \mathbf{h_i^a W_s h_i^C} \tag{15}$$

$$P_{end}(i) = softmax_i\ \mathbf{h_i^a W_e h_i^C} \tag{16}$$

$$\mathbf{o} = argmax_{ij}\ P_{start}(i) * P_{end}(j) \tag{17}$$

where W_s and W_e are bilinear sequential attention matching terms. Equation (15) and (16) are multiplicative attention between hidden representation of passage and question encoding. Using the output vector $\mathbf{o}$, the system outputs the most likely answer using:

$$a = argmax_{chunks}\ \mathbf{o} \tag{18}$$

For each bifurcated chunk, the above equation gives us the selected answer based on the considered chunk.

For a given question, we consider all the retrieved videos and employ a gating mechanism to adaptively control the input into the neural model. Specifically, the chunk-level gating mechanism filters out chunks from across different videos which are not similar to the question. The resulting set of chunks which pass through the chunk-level gating mechanism are then ranked using the scores and the top-k chunks are chosen for evaluation. As described in the Section 4, we used a ranking based evaluation technique wherein the ground-truth relevance label on the answer generated by each chunk is used to compute various metrics.

4 EXPERIMENTAL SETUP

Evaluating the correctness of answers in a video QA domain is challenging since there exists no publicly available labeled dataset. In this section, we describe the dataset created via crowdsourcing for training and evaluating the proposed model (Section 4.1), as well as the different baselines considered (Section 4.2). Additionally, we describe the ranking based evaluation setup used for comparisons (Section 4.3).

4.1 Crowd-sourced Data Collection

While there exist large scale labeled datasets for textual content QA, an important challenge in video based QA is the lack of labeled dataset. In this work, we create a new labeled dataset based on educational videos obtained on Youtube, along side utilizing the large scale labeled SQuAD dataset for pre-training the model.

Our primary dataset is based on YouTube videos on educational content available through NPTEL (National Programme on Technology Enhanced Learning)[2]. The dataset comprises 15,298 videos, with over 200 sentences per video. The average time duration of a video in this dataset is 48 minutes, and the maximum time duration is 1 hour and 17 minutes. Given the educational nature of videos, we obtained a set of over 2000 questions from the curriculum, the answers for which were described in the video lectures.

We employed Amazon Mechanical Turk for getting labeled data for each question. We followed a rigorous process for the video annotation task, wherein for each question, a set of videos were annotated by a minimum of 3 different human judges. Judges were explained and shown examples of question-answering tasks, and hidden quality control measures were employed to remove judgements from incompetent judges. We performed post filtering of judgments based on significant disagreement of the judge with other judges, which resulted in an iterative process following which judgments from over 30 judges were removed from the pool of judges. Overall, we obtained an inter-rater agreement of 0.69 (FleissâĂŹ kappa) which implies substantial agreement among judges.

4.1.1 Dataset for Distant Supervision. We use the SQuAD dataset for pre-training our deep neural model. The Stanford Question Answering Dataset (SQuAD) Rajpurkar et al.[25] is a dataset for machine comprehension based on Wikipedia. The dataset contains 87k examples for training and 10k for development, with a large hidden test set which can only be accessed by the SQuAD creators. Each example is composed of a paragraph extracted from a Wikipedia article and an associated human-generated question. The answer is always a span from this paragraph and a model is given credit if its predicted answer matches it.

4.2 Baselines

We compare the proposed GABiNet model with a number of baselines. The task of question-answering on video content has not been well-researched, and there exists few published baseline approaches to compare with. We instead consider strong baselines and state-of-the-art approach from neural question answering in the traditional textual content setting.

[2] contains video lectures from IITs and IISc

Method	#chunks	F1	MRR	ndcg@1	ndcg@3
Content	4	0.259	0.138	0.067	0.113
	5	0.274	0.137	0.04	0.115
	6	0.268	0.146	0.067	0.115
Temporal	4	0.262	0.142	0.173	0.098
	5	0.271	0.131	**0.212**$^{*\&}$	0.196
	6	0.273	0.148	0.135	0.207
Transition Points	4	0.261	0.141	0.15	0.201
	5	0.276	0.139	0.117	0.298
	6	0.249	**0.281**$^{*\&}$	0.117	0.308
Topical Drift	4	0.259	0.172	0.154	0.277
	5	0.278	0.198	0.147	0.298
	6	**0.288**$^{*\&}$	0.211	0.154	**0.309**$^{*\&}$

Table 3: Performance of different bifurcation techniques without gating for the top 10 videos. * and & signify statistically significant difference between the method and the best performing baseline using χ^2 test with $p \leq 0.05$

Method	bleu@1	bleu@2	bleu@3
Baseline 1	0.392	0.542	0.498
Baseline 2	0.433	0.619	0.578
Baseline 3	0.488	0.665	0.611
GABiNet	**0.513**$^{*\&}$	**0.693**$^{*\&}$	**0.663**$^{*\&}$

Table 4: Performance of different methods based on BLEU scores. * and & signify statistically significant difference between the method and the best performing baseline using χ^2 test with $p \leq 0.05$

- Baseline 1 - (Xiong *et al.*[34]): A dynamic co-attention network for question answering which fuses co-dependent representations of the question and the document and introduces a dynamic pointing decoder iterates over potential answer spans.
- Baseline 2 (Wang *et al.*[31]): a machine comprehension model based on match-LSTM and pointer network, used for question answering of textual content.
- Baseline 3 (Chen *et al.*[6]): a multi-layer recurrent neural network machine comprehension model trained to detect answer spans in documents.
- Baseline 4 (Yang et al.[37]): a reading comprehension model which dynamically combines word-level and character-level representations with the idea of modeling the interaction between questions and paragraphs.

Additionally, to evaluate the benefits offered by different components of the proposed GABiNet architecture, we compare with variants of the proposed model:

- Variant 1: The proposed GABiNet model but without gating and attention mechanism
- Variant 2: The proposed GABiNet model with attention module but without gating mechanism
- Variant 3: The proposed model GABiNet with chunk-level gating only
- Variant 4: The complete GABiNet architecture

4.3 Evaluation Metrics

The goal of the evaluation setup is to evaluate the quality and correctness of the inferred answer. The proposed GABiNet model outputs a ranked list of answers, one for each chunk, which we compare with the ground truth answer obtained via crowd-sourcing, to evaluate the correctness of the inferred answer. Given the ranking setting, we evaluate the performance of the different techniques using a number of metrics, including F1 scores, Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG) estimates. F1 score measures the overlap between the prediction and ground truth answers, which takes the maximum F1 over all of the ground truth answers. MRR evaluates any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness. Normalized DCG measures the usefulness of a document based on its position in the result list, i.e., top results are more relevant than the bottom ones.

We additionally use the BLEU scores (Bilingual Evaluation Understudy) from Papineni et al.[23]. BLEU uses a modified form of precision to compare a candidate answer text against reference text (ground truth answer). BLEUâĂŹs output is always a number between 0 and 1. This value indicates how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts.

4.3.1 Parameter Setting. To extract the initial embeddings for tokens, we used the Glove implementation with 300 dimensions. We used spaCy[3] for tokenization, lemmatization and generating part-of-speech and named entity tags. We used Glove embeddings Pennington et al.[24] for word vector representation. We used 3-layers of bidirectional LSTM with 128 hidden units for both chunk and question encoding. We used a minibatch size of 32, *Adam* for optimization as described in Kingma et al.[18] with hyperparameters such as learning rate(learning_rate) of 0.004, beta_1 as 0.9 and beta_2 as 0.996. Dropout with p = 0.2 applied to word embeddings.

[3] spacy v1.7.x Downloaded from: *https://spacy.io/*

Method	#chunks	F1	MRR	ndcg@1	ndcg@3
Content	4	0.303	0.318	0.24	0.273
	5	0.328	0.319	**0.28**$^{*\&}$	0.215
	6	0.316	0.31	0.227	0.302
Temporal	4	0.252	0.27	0.227	0.112
	5	0.326	0.29	0.253	0.214
	6	0.314	0.29	0.227	0.257
Transition Points	4	0.194	0.23	0.187	0.204
	5	0.21	0.23	0.173	0.242
	6	0.285	0.26	0.2	0.266
Topical Drift	4	0.276	0.28	0.2	0.303
	5	0.306	0.31	0.227	0.294
	6	**0.328**$^{*\&}$	**0.32**$^{*\&}$	0.227	**0.305**$^{*\&}$

Table 5: Performance of different bifurcation techniques with gating. * and & signify statistically significant difference between the method and the best performing baseline using χ^2 test with $p \leq 0.05$

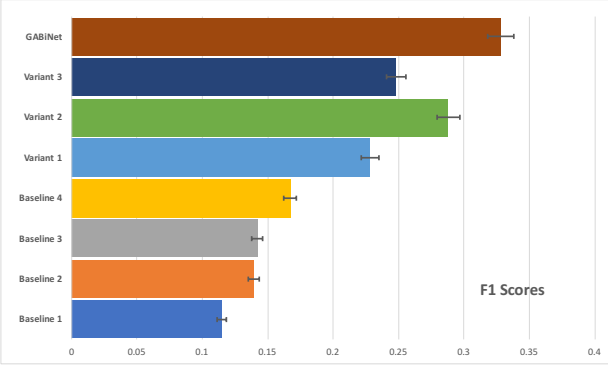

Figure 2: Comparing the performance of the baselines and variants of the proposed approach based on F1 scores on the labeled dataset.

4.4 Results and Analysis

We leverage the experimental setup described above to evaluate and compare the performance of the proposed GABiNet architecture with different baselines. We begin by investigating the impact of pre-training on a different dataset, followed by evaluating the performance on the main task of finding relevant answers.

4.4.1 Impact of pre-training: Given the scarcity of labeled data in video domain, we explore the benefits of pre-training on the SQuAD QA dataset. We pre-train the proposed model on SQuAD dataset, followed by training on the labeled video QA dataset and observed a performance gain of over 16% in F1 scores. All further experiments were performed using this pre-training step.

4.4.2 Finding Relevant Answers. We then examine the performance of the various techniques on the task of finding the correct answer spans on the labeled dataset. Figure 2 presents the F1 score comparing the performance across the baselines and variants of the proposed approach. We observe a substantial gain in performance for the proposed GABiNet architecture over all baselines and variants considered. Further, all variants of GABiNet outperform the baselines, with Variant 2 performing better than variants 1 and 3, hinting at the benefits offered by the gating mechanism. A major portion of the performance boost can be attributed to the introduction of the content bifurcation module, with the complete GABiNet architecture with the bifurcation module improving F1 scores by over 22% over the best performing baseline. The GABiNet architecture used for this evaluation uses the best performing content bifurcation method.

4.4.3 Impact of Content Bifurcation: We investigate the benefit of content bifurcation module on question answering performance. As shown in Figure 2, we observe that content bifurcation with no attention and with attention neural model help boost the F1 score by 11.3% and 14.9% respectively. This demonstrates the importance of bifurcating the transcript text into chunks.

4.4.4 Comparing Different Bifurcation Techniques: Table 5 compares the different content bifurcation techniques. We observe that fixed width content bifurcation performs better than fixed width temporal bifurcation, while detecting transition points based on sentence pairs is not very helpful. However, without gating, detecting topical drift consistently performs better than all other content bifurcation techniques, with over 1.2%, 1.4% and 3.9% improvement in F1 score compared with content, temporal and transition bifurcation techniques respectively. We present results with chunk sizes of 4-6 since they performed better than lower and higher number of chunks across all videos considered. In the future, we intend to explore adaptive chunk sizes based on the video duration.

We additionally observe a similar trend in the relative performance of the different content bifurcation techniques for the NDCG and MRR metrics. From these results, we see that chunking the transcripts using the *Topical Drift* through time approach performs the best.

4.4.5 Impact of Chunk-level Gating. Without the chunk-level gating, all bifurcated chunks from each video were passed to our neural model. With the chunk-level gating, only selected bifurcations are passed through to our model, which reduces the amount of noise the model receives and learns from, which in turn makes the model more efficient to infer the answer. As shown in Table 3, we observe a general decline in all metrics when compared with Table 5 which highlights the importance of the gating mechanism. Indeed, gating mechanism helps reduce noisy input to the model. Further, we observe a similar trend in performance with topical drift chunking performing better than most other techniques. We observe an expected increase in the ndcg@3 estimate over ndcg@1 estimate, with the ranking technique getting more chance to retrieve the correct answer with 3 slots. Overall, we observe a dip in performance to the tune of 12%, which highlights the importance of including the gating mechanism.

4.4.6 Relevance of Answer. Table 4 shows BLEU score for the proposed GABiNet model with the best bifurcation technique (Topical Drift), alongside estimates for the different baselines. We observe that using 6 chunks leads to the most relevant answer in the form of the highest BLEU score. The BLEU score is a strong indicator of the similarity between the inferred answer with the ground truth answer. We observe a similar trend as the ranking evaluation, with the proposed GABiNet architecture performing best, with over 22% improvement over the worst and over 5% boost in BLEU score over the best performing baseline. These results further indicate and support the observation that the proposed GABiNet architecture is better able to infer answers from transcribed video content than well established baselines.

4.5 Discussion

Different metrics are used for evaluating model performance, such as F1 score, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG) and BLEU score. Across all metrics, we observe a substantial improvement in the ability to infer answers for the proposed GABiNet architecture over its variants and baselines. Given the free flow nature of content in videos, and the lack of demarcation, the baseline approaches are unable to perform efficiently while the content bifurcation module helps the GABiNet model to overcome this issue, by breaking transcripted text into meaningful chunks. We also observe performance boost by pre-training on a different unrelated large scale dataset (SQuAD). Even though we pre-trained on the entire SQuAD, it would be interesting to consider a more focused relevant subset of data from it to see if having a more focused distant supervision is more helpful. In terms of attention and gating mechanism, similar to widely published work, we observe similar trends which highlight the benefit of considering attention mechanism. Finally, most existing QA systems assume the input consists of text which already has the right answer. We alleviated this constraint and considered many different videos and relied on the model to not only find the right video and the right chunk, but also answer from the video. We observed that having the chunk-level gating mechanism helped in this regard.

5 CONCLUSION

With an increasing number of online courses and educational video content being generated, video based question answering is an increasingly important problem in an industrial setting. We present a composite neural architecture (GABiNet) comprised of a gated attention model with content bifurcation techniques to answer questions using a video's transcript. Our findings show a clear need for considering content bifurcation when dealing with long duration video content. We further demonstrate that detecting topical shift in video content plays an important role in improving question-answering performance across three different metrics. We envision future research to incorporate raw video content and explore adaptive width bifurcation techniques to jointly learn content and question representations for machine comprehension for videos.

6 ACKNOWLEDGEMENT

This work is partly supported by Infosys Foundation.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[2] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.
[3] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1533–1544.
[4] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 1247–1250.
[5] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075* (2015).
[6] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 1. 1870–1879.
[7] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733* (2016).
[8] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Attention-over-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423* (2016).
[9] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549* (2016).
[10] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1156–1165.
[11] Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401* (2014).
[12] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. 1693–1701.
[13] Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. Wikireading: A novel large-scale language understanding task over wikipedia. *arXiv preprint arXiv:1608.03542* (2016).
[14] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301* (2015).
[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[16] Minghao Hu, Yuxing Peng, and Xipeng Qiu. 2017. Reinforced mnemonic reader for machine comprehension. *CoRR, abs/1705.02798* (2017).
[17] Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547* (2016).

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Jong-Hwan Lee, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee. 2000. Speech feature extraction using independent component analysis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, Vol. 3. IEEE, 1631–1634.

[20] Kenton Lee, Shimi Salant, Tom Kwiatkowski, Ankur Parikh, Dipanjan Das, and Jonathan Berant. 2016. Learning recurrent span representations for extractive question answering. *arXiv preprint arXiv:1611.01436* (2016).

[21] Guangda Li, Zhaoyan Ming, Haojie Li, and Tat-Seng Chua. 2009. Video reference: question answering on YouTube. In *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 773–776.

[22] Boyuan Pan, Hao Li, Zhou Zhao, Bin Cao, Deng Cai, and Xiaofei He. 2017. MEMEN: Multi-layer Embedding with Memory Networks for Machine Comprehension. *arXiv preprint arXiv:1707.09098* (2017).

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[25] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[26] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664* (2015).

[27] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* (2016).

[28] Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1047–1055.

[29] Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245* (2016).

[30] Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. 2016. Natural language comprehension with the epireader. *arXiv preprint arXiv:1606.02270* (2016).

[31] Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905* (2016).

[32] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 189–198.

[33] Zhiguo Wang, Haitao Mi, Wael Hamza, and Radu Florian. 2016. Multi-perspective context matching for machine comprehension. *arXiv preprint arXiv:1612.04211* (2016).

[34] Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604* (2016).

[35] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. 2003. VideoQA: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 632–641.

[36] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2013–2018.

[37] Zhilin Yang, Bhuwan Dhingra, Ye Yuan, Junjie Hu, William W Cohen, and Ruslan Salakhutdinov. 2016. Words or characters? fine-grained gating for reading comprehension. *In proceedings, ICLR 2017* (2016).

[38] Yang Yu, Wei Zhang, Kazi Hasan, Mo Yu, Bing Xiang, and Bowen Zhou. 2017. End-to-end reading comprehension with dynamic answer chunk ranking. *ICLR submission* (2017).

[39] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision* 124, 3 (2017), 409–421.