УДК 004.89

А. С. Подкладкин, Н. Е. Пособилов

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ГЛУБОКОГО ОБУЧЕНИЯ С ЦЕЛЬЮ ФОРМИРОВАНИЯ КОММЕНТАРИЕВ К ВИДЕО

Нижегородский государственный технический университет им. Р. Е. Алексеева

Предложен подход к решению задачи генерации комментариев к видео с использованием нейронных сетей глубокого обучения. Была разработана модель сети с комбинацией сверточной и рекуррентной структур. Проведено сравнение предлагаемого метода и существующих с помощью метрик генерации выражений.

Ключевые слова: нейронная сеть, глубокое обучение, обработка естественных языков, комментирование видео.

На сегодняшний день цифровой контент по своей природе является мультимедиа: текст, аудио, изображения, видео и т.д. В частности, видео становится новым способом общения между интернет-пользователями. В связи с распространением недорогих устройств мобильной записи, количество видео-контента растет, поэтому возникает потребность в разработке технологий автоматизированного анализа видео. Для большинства людей задача просмотра видео и описание происходящего в нем (словами) является легкой задачей. Для машин же извлечение смысла из видео и генерация комментариев – очень сложная задача. Фундаментальный вопрос, который лежит в основе успеха, – понимание содержимого видео.

Формирование комментариев к видео – новая задача, которая получает все большее внимание исследователей в таких областях, как компьютерное зрение и обработка естественных языков. Это логическое продолжение задачи описания статических изображений. Описание видео на естественном языке еще более сложная задача, так как модель описания видео должна быть достаточно мощной, чтобы не только распознавать действия и объекты, но и иметь возможность моделировать их пространственно-временные отношения, выраженные на естественном языке.

Перспективным направлением для решения этой задачи является использование нейронных сетей глубокого обучения, так как их применение позволило достигнуть высоких результатов в описании статических изображений.

Рассмотрим основные типы архитектур нейронных сетей

Сверточная нейронная сеть (СНС) – специальная архитектура искусственных нейронных сетей, предложенная Я. Лекуном и нацеленная на эффективное распознавание изображений, входит в состав технологий глубинного обучения. Использует некоторые особенности зрительной коры, в которой были открыты так называемые простые клетки, реагирующие на прямые линии под разными углами, и сложные клетки, реакция которых связана с активацией определённого набора простых клеток. Таким образом, идея свёрточных нейронных сетей заключается в чередовании свёрточных слоев и субдискретизирующих слоев. Структура сети – однонаправленная (без обратных связей), принципиально многослойная. Для обучения используются стандартные методы, чаще всего, метод обратного распространения ошибки. Функция активации нейронов (передаточная функция) – любая, по выбору исследователя.

Особенность сверточной нейросети заключается в том, что в ней нейроны первых уровней упорядочены в особую структуру, а именно: на первых слоях нейроны разбиты на изображения определенного размера (их еще иногда называют картами), разные карты внутри одного слоя соответствуют нейронам разного типа, реагирующим на разные особенности

изображений. И вычисления активации следующего слоя в сверточных нейросетях бывают двух основных типов. В первом типе вычислений активация нейронов следующего уровня вычисляется как линейная комбинация активаций нейронов предыдущего уровня, причем веса этих линейных активаций зависят только от взаимных положений нейронов, типов нейронов, но не зависят от положения данного нейрона внутри карты.

Во втором типе вычислений активация нейронов на следующем уровне просто повторяет активацию нейронов на предыдущем уровне, но изображение становится меньшего размера за счет того, что активация рядом расположенных нейронов заменяется на их максимум или их среднее – так называемая процедура пулинга.

AlexNet – архитектура сверточной сети (рис. 1), созданная в 2012 г., которая выиграла конкурс по классификации изображений в этом же году базы ImageNet [1]. Основные особенности: 7 скрытых слоев, ReLu в качестве функций активации, 60000000 параметров.

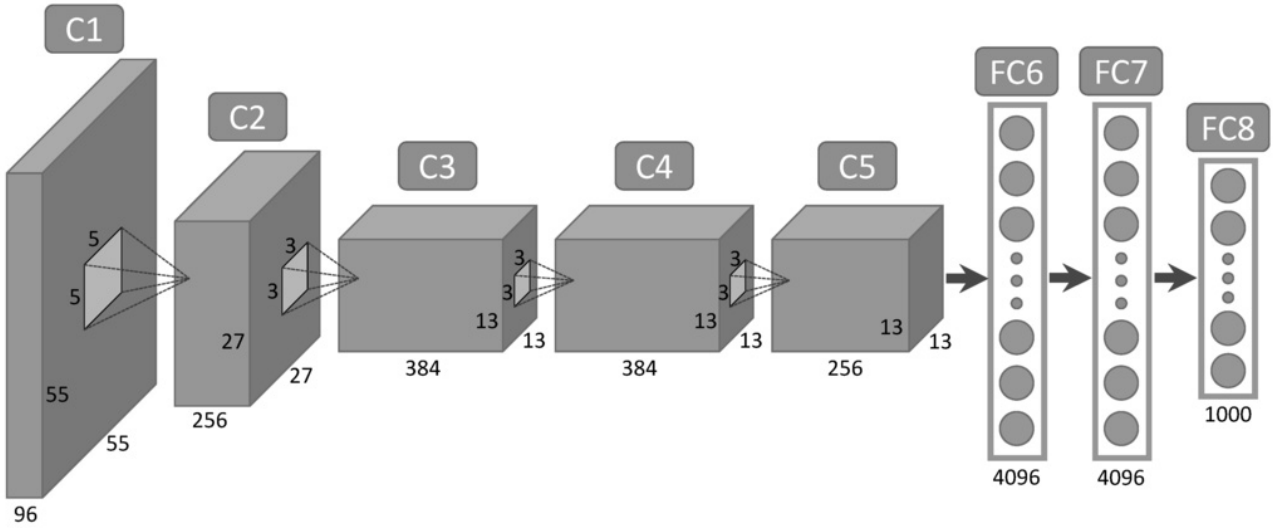

Рис. 1. Архитектура нейронной сети AlexNet (caffe)

Рекуррентные нейронные сети (РНС) - вид нейронных сетей, в которых имеется обратная связь. При этом под обратной связью подразумевается связь от логически более удалённого элемента к менее удалённому. Наличие обратных связей позволяет запоминать и воспроизводить целые последовательности реакций на один стимул. Стандартные РНС учатся отображать последовательность входов (x_1, ..., x_t) в последовательность скрытых состояний (h_1, ..., h_t) и от скрытых состояний к последовательности выходов (z_1, ... , z_t) на основе следующих повторений:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1}),$$

$$z_t = g(w_{zh}h_t),$$

где f и g - элементарные нелинейные функции, такие как сигмоида или гиперболический тангенс; x_t - векторное представление фиксированной длины; $h_t \in R^N$ - скрытое состояние с N значениями; W_{ij} - веса, соединяющие слои нейронов; z_t - выходной вектор.

Рекуррентные нейронные сети могут научиться отображать последовательности, для которых известно соответствие между входами и выходами, однако неясно, могут ли они применяться к задачам, когда входы (x_i) и выходы (z_i) имеют разную длину. Другая известная проблема с РНС заключается в том, что их трудно обучать на большом объеме данных. Для решения этих проблем была предложена архитектура LSTM сетей.

Сети долго-краткосрочной памяти (Long Short Term Memory) - обычно просто называют LSTM - особый вид РНС, способных к обучению долгосрочным зависимостям. Они были предложены Хохрейтером и Шмидхубером в 1997 г. [2]. Сети работают невероятно хорошо на большом разнообразии проблем и в данный момент широко применяются. LSTM

специально спроектированы таким образом, чтобы избежать проблемы долгосрочных зависимостей. Запоминать информацию на длительный период времени - это практически их поведение по умолчанию.

Все рекуррентные нейронные сети имеют форму цепи повторяющих модулей (repeating module) нейронной сети. LSTM тоже имеют такую цепную структуру, но повторяющий модуль другого строения. Вместо одного нейронного слоя их четыре, причем они взаимодействуют особым образом.

В данной работе использовалась реализация LSTM сети с повторяющимся блоком, представленным на рис. 2.

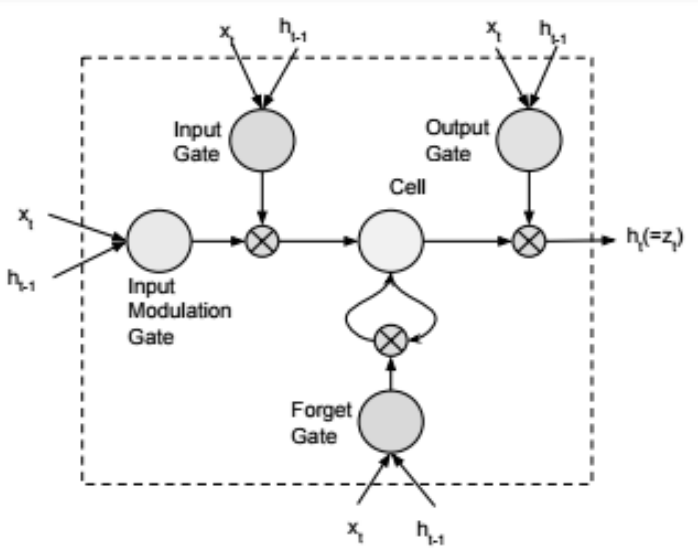

Рис. 2. Архитектура LSTM блока

В основе модели LSTM лежит ячейка памяти *c*, которая на каждом шаге кодирует значения входов, вычисленные до этого шага. Ячейка модулируется воротами, имеющими сигмоидальную функцию активации с диапазоном [0, 1] и применяющимися мультипликативно. Ворота определяют, сохраняет ли LSTM значение из затвора (если значение 1) или отбрасывает его (если значение 0). Три вида ворот: входные ворота (*i*), – контролируют, рассматривает ли LSTM текущий вход (x_t); забывающие ворота (*f*), позволяющие LSTM забыть свою предыдущую память (c_{t-1}); выходные ворота (*o*) – решают, какую часть памяти необходимо передать в скрытое состояние (h_t). Работа LSTM блока описывается так:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1}),$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1}),$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1}),$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \varphi(W_{xc}x_t + W_{hc}h_{t-1}),$$
$$h_t = o_t \odot \varphi(c_t),$$

где σ – сигмоидальная функция активации; φ - гиперболическая функция активации; $\odot$ – произведение с учетом значения ворот; W_{ij} – весовые матрицы (являются обученными параметрами).

На рис. 2 показана предлагаемая модель для генерации комментария к видео. Система основана на комбинации сверточной и рекуррентной нейронной сети. Сначала с помощью СНС генерируется одномерное векторное представление видео. Затем используются РНС, в частности LSTM, чтобы "декодировать" вектора в предложение (т.е. последовательность слов).

Наиболее вероятное описание для видео определяется обучением модели, чтобы максимизировать логарифмическую функцию правдоподобия выражения *S*, учитывая соответствующее видео *V* и параметры модели θ:

$$\theta^* = arg\max \sum_{(V,S)} \log(S|V;\theta) \quad (1)$$

Так как модель генерирует одно слово в предложении на каждом временном шаге, то естественно использовать совместную вероятность последовательных слов. Таким образом, логарифмическая функция правдоподобия предложения определяется суммой логарифмических функций всех слов и может быть выражена следующим образом:

$$logP(S|V) = \sum_{t=0}^{N} logP(S_{w_t}|V, S_{w_1}, \dots, S_{w_{t-1}}),$$

где S_{w_t} представляет *i*-е слово; N – общее число слов. Параметр θ отброшен для удобства.

Необходимо максимизировать функцию правдоподобия выражения *S*. Эта функция рассчитывается и оптимизируется на всем тренировочном наборе данных. Выбирают слово с максимальной вероятностью на каждом временном шаге и устанавливают его на вход РНН сети для следующего временного шага.

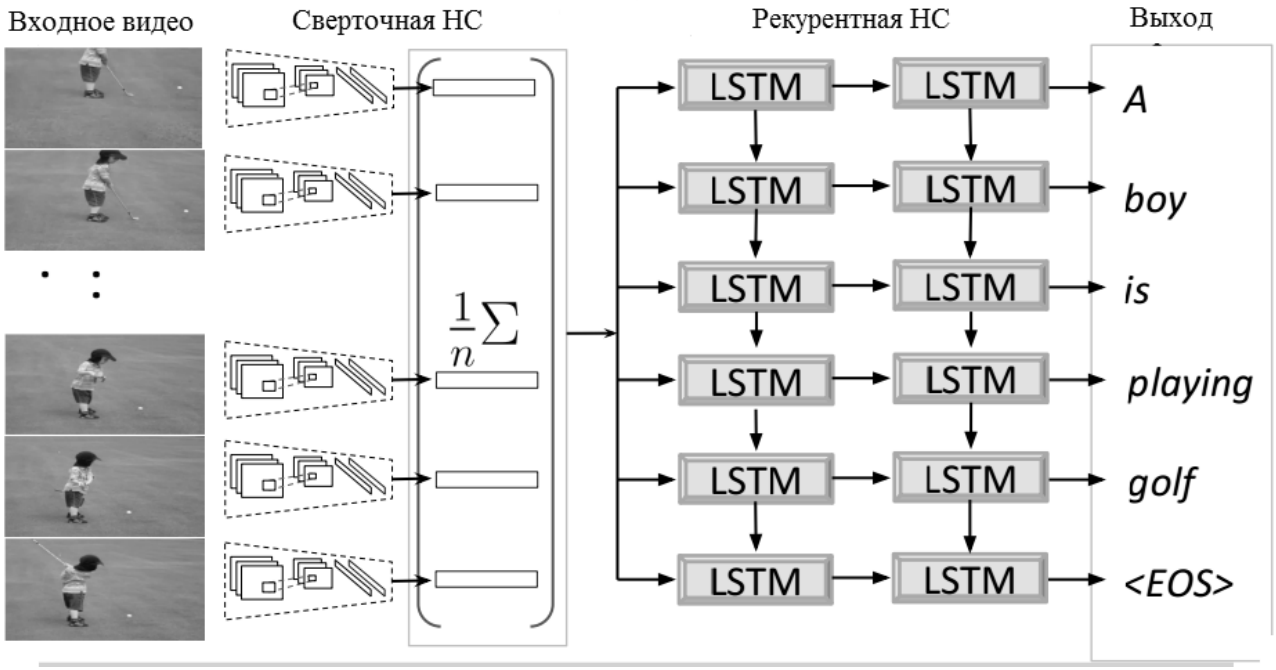

Рис. 3. Структура модели для генерации комментариев к видео

В работе используется LSTM сеть для генерации последовательности, так как сети такой архитектуры показывают высокую эффективность в задачах распознавания речи и машинного перевода. В данном случае используется два слоя LSTM, как показано на рис. 3. LSTM сеть “декодирует” вектор визуальных признаков, характеризующий видео.

Первым шагом в процессе генерации комментария к видео является создание вектора фиксированной длины, который эффективно обобщает происходящее на видео. Для этого используется СНС, в частности архитектуру AlexNet. Эта сеть предварительно обучена на 1.2 млн изображений набора данных ImageNet и, следовательно, имеет надежную инициализацию для распознавания объектов и позволяет сократить время обучения. Берем выборочные кадры из видео (1 из каждых 10 кадров) и извлекаем выход из 7 полносвязного слоя (fc7 на рис.1), выполняем усреднение по кадрам и получаем вектор размерностью 4096 для каждого видео. Этот вектор подается на вход первому слою LSTM сети, а состояние первого слоя является входом второго слоя. Слово из предложения является выходом второго слоя. В данной работе слова представляются как частота вхождения каждого слова в словарь (1-из-*N*, где *N* – мощность словаря). Для обучения использовались две видеокарты NVIDIA GTX 980

4GB, каждая из которых имеет 2048 CUDA ядер. Первая видеокарта использовалась для обучения первого слоя LSTM, а вторая – второго слоя. Для предобработки изображений и выражений использовался ЦП.

Двухслойная LSTM модель обучается предсказывать следующее слово S_{w_t} в комментарии на основе вектора визуальных признаков и предыдущего *t-1* слова, $P(S_{w_t}|V, S_{w_1}, \dots, S_{w_{t-1}})$. Функция 1 рассчитывается и оптимизируется на всем тренировочном наборе данных с использованием стохастического градиентного спуска. На каждом временном шаге вход x_t подается в LSTM наряду с состоянием h_{t-1} предыдущего шага, и LSTM выдает следующий вектор состояния h_t и слово. Для первого слоя LSTM x_t – конкатенация визуального вектора признаков и предыдущие закодированное слово. Для второго слоя LSTM x_t - z_t первого слоя. Выбираем слово с максимальной вероятностью на каждом временном шаге и устанавливаем его на вход LSTM сети для следующего временного шага, пока не получим маркер конца строки.

В процессе выполнения эксперимента использовался набор видеоданных Microsoft Research Video Descriptor Corpus (MSVD) [3]. Этот набор данных представляет собой набор из 1970 отрывков YouTube видео. Длительность каждого ролика составляет от 10 до 25 с, как правило, изображающих одну активность или короткую последовательность. Для каждого видео доступно порядка 40 описаний на английском языке. Для нашей задачи выбираем 1200 видео, которые будут использованы в качестве обучающих данных, 100 видео для проверки и 670 для тестирования.

Так как количество видео в наборе данных мало по сравнению с набором данных, используемых LSTM моделями в других задачах, то используем набор данных Flickr30k для инициализации весов LSTM сети, чтобы увеличить ее скорость и точность обучения на видеоданных. Flick30k содержит 30000 изображений, каждому из которых соответствуют 5 или более комментариев. Выбираем 1000 изображений для проверки, а остальные для обучения. Эксперименты проводим по обучению моделей на каждом наборе данных как отдельно, так и в комбинации.

Используемые модели генерации комментариев:

1. FGM – Factor Graph Model [4]. Этот подход, используя алгоритм SIFT для распознавания объектов и фактор-граф, определяет наиболее вероятный субъект, глагол, объект и сцену. Затем используется простой шаблон для генерации предложения.

2. Предлагаемые две основные LSTM модели:

а) LSTM-YT – двухслойная LSTM сеть, обученная на наборе данных MSVD;

б) LSTM-YTflickr – модель, сначала обученная на наборе данных Flickr30k, а затем тонко настроенная на MSVD наборе.

Для сравнения результатов экспериментов использовали алгоритмы автоматического оценивания качества машинного перевода BLEU [5] и METEOR [6].

Результаты

Использование глубоких моделей генерации комментариев к видео дает прирост эффективности (табл. 1). Особенно это заметно на метрике BLEU, разница на метрике METEOR не столь высока в связи с тем, что модель FGM хорошо оптимизирована на поиск субъекта, глагола и объект в видео.

Предварительное обучение предлагаемой модели на изображениях повышает производительность в обоих метриках. Модель LSTM-YTflickr, которая была претренирована на наборе данных Flickr30k, показывает прирост порядка 1% на обоих метриках по сравнению с LSTM-YT. Это указывало на то, что предлагаемая модель может использовать вспомогательный источник данных для улучшения предсказания. Разница между результатами двух наших моделей не столь велика из-за переобучения.

Таблица 1

Результаты экспериментов

Модель	BLEU (%)	METEOR (%)
FGM	13.68	23.90
LSTM-YT	31.18	26.85
LSTM-YTflickr	32.03	27.89

На рис. 4 можно увидеть примеры генерации комментариев различных моделей. Точность предлагаемых моделей выше, чем метода FGM, однако точность крайне низка, например, на 3 картинке неправильно определяются объекты на видео.

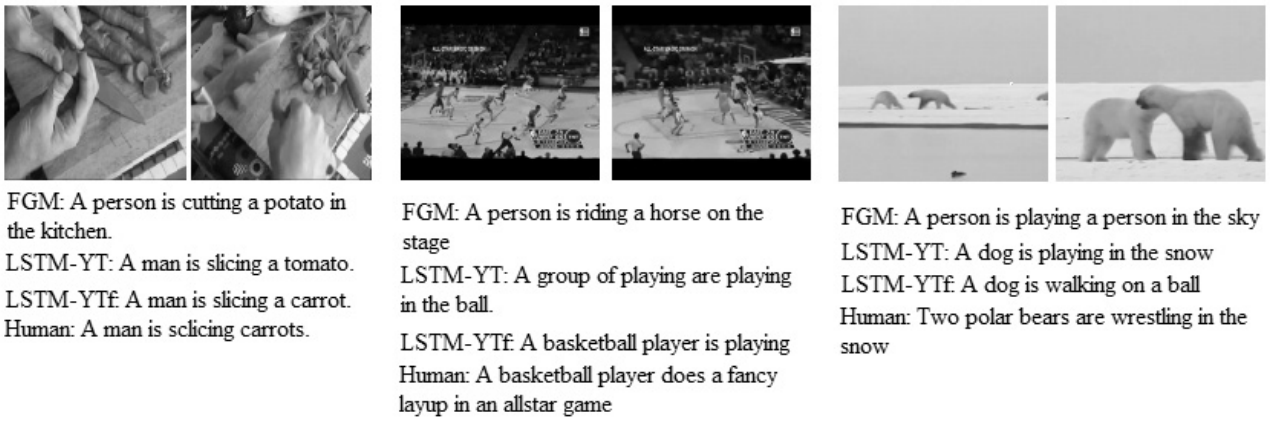

Рис. 4. Примеры работы моделей

Выводы

В данной работе предложена модель для описания видео, которое использует глубокие нейронные сети. В экспериментальной оценке показано, что предлагаемый подход создает лучшие комментарии, чем “классические” методы. Также показано, что использование размеченных изображений повышает производительность по сравнению с применением только видеоданных. Однако предложенный подход не учитывает временную информацию в видео, что является перспективным направлением исследований будущей работы.

Библиографический список

1. **Krizhevsky, A.** ImageNet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G.Hinton [Электронный ресурс]. – Canada, 2012. URL: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
2. **Hochreiter, S.** Long short-term memory / S. Hochreiter, J. Schmidhuber - [Электронный ресурс]. – Germany, 1997. URL: http://dl.acm.org/citation.cfm?id=1246450
3. **Chen, D. L.** Collecting highly parallel data for paraphrase evaluation / D. L. Chen, W.B. Dolan - [Электронный ресурс]. – USA, 2011. – URL: http://www.cs.utexas.edu/~ai-lab/downloadPublication.php?filename=http://www.cs.utexas.edu/users/ml/papers/chen.acl11.pdf&pubid=127065
4. Integrating language and vision to generate natural language descriptions of videos in the wild / Thomason J. [et al.] - [Электронный ресурс]/ – USA, 2015. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.650.7265&rep=rep1&type=pdf,

5. **Papineni, P.** BLEU: a method for automatic evaluation of machine translation / P. Papineni [et al.] - [Электронный ресурс]. – USA, 2002. URL: http://www.aclweb.org/anthology/P02-1040.pdf
6. **Elliott, D.** Comparing automatic evaluation measures for image description / D. Elliott, F. Keller - [Электронный ресурс]. – UK, 2014. URL : http://acl2014.org/acl2014/P14-2/pdf/P14-2074.pdf

Дата поступления
в редакцию 10.08.2017

A.S. Podkladkin, N.E. Posobilov

USING DEEP LEARNING NEURAL NETWORKS FOR VIDEO CAPTIONING

Nizhny Novgorod state technical university n.a. R.E. Alekseev

Purpose: The goal is to generate language sentences for videos.
Design/methodology/approach: This paper presents a model to generate captions for videos. For this we exploit recurrent neural networks, specifically LSTMs, which have demonstrated state-of-the-art performance in image caption generation. Our LSTM model is trained on video-sentence pairs and learns to associate a sequence of video frames to a sequence of words in order to generate a description of the event in the video clip. The model consists of two components: deep convolutional neural network for learning powerful video representation, a deep RNN for generating sentences.
Findings: The model achieves state-of-the-art performance on the MSVD dataset. Despite its conceptual simplicity, our model significantly benefits from additional data, suggesting that it has a high model capacity, and is able to learn complex sematic structure of videos.
Research limitations/implications: However, the proposed approach does not take into account the temporal information in the video, which is a promising direction for research of future work. Moreover, the video description generation might be significantly boosted if we could have sufficient labeled video-sentence pairs to train a deeper RNN.
Originality/value: The model outperforms the FGM model with a significantly large margin on both BLEU and METEOR language generation metrics.

Key words: neural network, deep learning, natural language processing, video captioning.