Partners in Crime: Multi-view Sequential Inference for Movie Understanding

Nikos Papasarantopoulos[♠] **Lea Frermann**[◇] **Mirella Lapata**[♠] **Shay B. Cohen**[♠]
[♠]School of Informatics, University of Edinburgh, UK
[◇]School of Computing and Information Systems, University of Melbourne, Australia
nikos.papasa@ed.ac.uk, lea.frermann@unimelb.edu.au
mlap@inf.ed.ac.uk, scohen@inf.ed.ac.uk

Abstract

Multi-view learning algorithms are powerful representation learning tools, often exploited in the context of multimodal problems. However, for problems requiring inference at the token-level of a sequence (that is, a separate prediction must be made for every time step), it is often the case that single-view systems are used, or that more than one views are fused in a simple manner. We describe an incremental neural architecture paired with a novel training objective for incremental inference. The network operates on multi-view data. We demonstrate the effectiveness of our approach on the problem of predicting perpetrators in crime drama series, for which our model significantly outperforms previous work and strong baselines. Moreover, we introduce two tasks, crime case and speaker type tagging, that contribute to movie understanding and demonstrate the effectiveness of our model on them.[1]

1 Introduction

While many natural language processing (NLP) problems concern exclusively textual or speech data, the integration of multimodal information (such as images, video or audio) is beneficial for a variety of problems. For example, visual information has been used in affect analysis (Kahou et al., 2016), sentiment analysis (Morency et al., 2011) and machine translation (Calixto et al., 2017; Lala and Specia, 2018). This is also the case for problems which are sequential in nature, such as video summarization (Smith and Kanade, 1998), continuous prediction of affect (Nicolaou et al., 2011) or engagement level prediction (Rehg et al., 2013).

Most existing multi-view representation learning approaches are tested in an unsupervised setup where the multi-view representations are learned separately from the task, and are designed to accommodate the learning of representation for monolithic (albeit multi-view) data points, not sequences (see Wang et al. 2015 for a survey). In this paper, we propose a neural architecture coupled with a novel training objective that integrates multi-view information for sequence prediction problems. Our model creates a multimodal embedding for every element of a sequence by using the correlational Gated Recurrent Unit (corrGRU; Yang et al. 2017) and makes token-level predictions based on those embeddings. Our training objective combines a supervision-guided term (cross-entropy) with a multi-view correlation objective on the available modalities.

We demonstrate the effectiveness of our model in an incremental inference setup (Figure 1), where it makes predictions on the fly without encoding the sequence in full, a more realistic scenario of interacting with data. This is a critical feature for online applications such as simultaneous translation (interpretation) and also a desirable behavior for movie processing models that mimic a human viewer watching a movie for the first time.

We evaluate our architecture on three tasks pertaining to movie understanding. Specifically, we use the recently introduced dataset of Frermann et al. (2018), which consists of episodes of the Crime Series Investigation (CSI) television series, segmented and aligned for three different modali-

[1] Our code is available at https://github.com/papagandalf/multiview_csi.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2057–2067,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

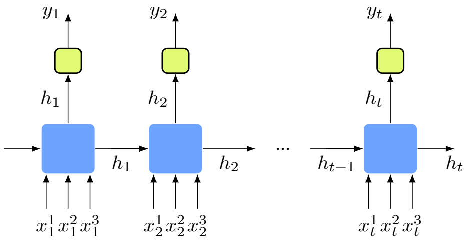


Figure 1: General unfolded overview of our model for a multimodal example x with three modalities: at each time step $t \in \{0, 1, ..., T\}$, representations x_t^k for each of three views are fed to a cell. The cell outputs a joint representation h_t for all the views which is fed to a softmax layer that generates a prediction y_t, depending on the task at hand. Training is performed with a correlation objective between the representations of x_t^k.

ties: image, audio and text. Originally, the dataset was introduced to train models in perpetrator mention identification, that is, viewing each episode as a sequence of multimodal elements, predicting a binary label indicating whether the perpetrator of a crime case is mentioned in each element. Leveraging the annotations of the dataset, we model two additional tasks: case segmentation (episodes can alternate between more than one crime case) and speaker type tagging (each utterance in the video can come from one type of speaker: detective, perpetrator, suspect, extra or none). Successfully modeling those two tasks provides structural information about episodes and informs perpetrator identification, a task solved on a per case and not on a per episode basis. The three tasks can be regarded as a form of *shallow movie understanding*, since they perform analysis on the structural level (cases), the dialogue level (speaker type) and the plot level (perpetrators).

We show that our multi-view model consistently outperforms models that integrate multimodal information by concatenating the representations of available views, across all three tasks. As a comparison, we also describe a non-multi-view variant of our architecture, equipped with a supervised multi-head attention module that can take advantage of two levels of annotation (e.g. token-level and sentence-level).

The contributions of this paper are the following:

- We propose a multi-view sequential inference neural architecture and use a novel training objective for training an RNN consisting of correlational GRU cells.
- We highlight the importance of multi-view fusion for multimodal applications, by comparing our model with a non-multi-view variant that employs multi-head supervised attention to make use of both the sentence-level and the token-level perpetrator annotations of the dataset.
- We introduce two novel tasks pertaining to shallow movie understanding that can be tackled in the context of television series data. We empirically show the effectiveness of our architecture and training objective on the perpetrator mention identification and the two newly introduced tasks, by using the Crime Scene Investigation (CSI) television series dataset. Notably, for the perpetrator identification task, our model significantly outperforms previous state of the art.

2 Background

Series Understanding[2] The abundance of series video data can benefit multimodal machine learning research and applications. Series commonly span many episodes (organized in loosely or tightly connected seasons), providing a large amount of data that data-hungry models can take advantage of. The sheer volume of data gives rise to several practical problems that machine learning models can tackle, such as the segmentation of continuous video streams to semantically coherent fragments (Del Fabro and Böszörmenyi, 2013) and speaker diarization (Miró et al., 2012; Bredin et al., 2014).

Work on movie/series analysis can be classified in three broad categories. *Deep semantic understanding* includes tasks that require thorough content analysis and reasoning, for example movie question answering (Tapaswi et al., 2016; Kim et al., 2017) or movie description (Rohrbach et al., 2016). *External understanding* refers to tasks whose end goal is not the analysis of the video content itself, but meta-information extraction relevant to preferences and recommendations for consumers of the videos (Bennett et al., 2007; Shi et al., 2013; Yang et al., 2012).

Shallow understanding refers to tasks that operate on the content level and extract content-related information, albeit without requiring complex reasoning. Their output is more tailored to structured prediction. Example tasks include speaker iden-

[2] We use the term “series” to refer to episodic shows, broadcast via television or other channels.

tification (Knyazeva et al., 2015), movie segmentation (Liu et al., 2013) and perpetrator mention identification (Fremann et al., 2018). We choose to tackle three problems of shallow understanding, namely a crime case sequence tagging problem, a speaker type sequence tagging problem and the problem of perpetrator mention identification.

Multi-view Learning Conventional machine learning algorithms treat all characteristics of training examples as features describing the input. When more than one modality or more than one sources of data (for example, images taken from different viewpoints) are available, view fusion can be achieved by early, late or hybrid fusion methods (Atrey et al., 2010). A simple concatenation of representations at the feature level is referred to as *early fusion*, while the integration of outputs of different modality-specific modules is called *late fusion*. Concatenation may cause overfitting in the case of a small size training samples and at the same time it is not intuitive, since the statistical properties of each view can be lost in the learning process. Multi-view learning algorithms extend early fusion, as they create sophisticated representations in which all available views are fused.

Our work extends the set of multi-view counterparts of popular sequence models. Rajagopalan et al. (2016) propose a general architecture that provides a degree of flexibility in designing different multi-view LSTM cells according to the application at hand and experiment with behavior recognition and image captioning. Ren et al. (2016) propose a multi-modal variant of LSTM and apply it to the task of speaker identification. Zadeh et al. (2018a) use an attention module and a multi-view gated memory to capture and summarize inter-modality interactions. Our proposed model enforces correlation between the representations of the available modalities; a technique that has been studied also for non-sequential neural models (Wang et al., 2015; Chang et al., 2018).

While the term *multi-view* does not always refer to multimodal settings, ideas from multi-view representation learning research are especially compelling for multimodal applications. Our model can be applied to sequential multi-view problems which are not necessarily multimodal.

Attention Models Attention mechanisms (Bahdanau et al., 2015) in various forms have been used in several multimodal applications, such as sentiment analysis, speaker trait recognition and emotion recognition (Zadeh et al., 2018b), machine translation (Caglayan et al., 2016), image (Xu et al., 2015) and video description (Hori et al., 2017). Broadly, an attention mechanism modifies the output of a sequence representation, based on the coherence of each of the elements of the sequence to a specific "query". Information learned by the attention mechanism may have a distinct conceptual importance (e.g. alignments in machine translation) or simply indicate which elements of the input contribute more to the final output representation.

Attention mechanisms can be learned along with the rest of the network in an end-to-end fashion, or can be explicitly supervised by providing the model with pre-calculated attention scores. Supervised attention has been shown to boost the performance of models for machine translation (Mi et al., 2016), constituency parsing (Kamigaito et al., 2017), event detection (Liu et al., 2017b) and aspect-based sentiment analysis (Cheng et al., 2017). Furthermore, image attention mechanisms guided by weak or direct supervision have been proposed for the tasks of image (Liu et al., 2017a) and video captioning (Yu et al., 2017).

Multi-head attention mechanisms (Vaswani et al., 2017) employ more than one, independent, attention mechanisms, boasting multiple areas of focus on the input sequence. The main idea behind them is that a single attention head may not prove adequate to capture all the different types and positions of information that are important to the end task. Our attentive model variant uses supervised multi-head attention to take advantage of annotations in two levels of granularity.

3 Multi-view Sequential Inference

Taking together the idea of multi-view learning with incremental sequence labeling, we formulate our problem as follows.

We assume a set $\mathcal{X}$ of M examples. Every $X_j \in \mathcal{X}, j \in \{1, \ldots, M\}$, consists of T elements, which form a sequence $X_j = [x_{j1}\ x_{j2}\ \ldots\ x_{jT}]$. Sequences with less than T elements are padded to be of length T. Each of the elements in the sequence is paired with a label from a set $\mathcal{Y}$ (binary or multi-class), which is the desired output. Lastly, for every x_{jt}, information from a set $\mathcal{V}$ of different views is available.

More specifically, we consider a dataset where each X_j is a video and distinguish and model three different views/modalities: image, audio and text (from aligned script or subtitles, if available). Each video X_j is represented as a sequence of short, semantically coherent snippets x_{jt} (for instance, snippets may correspond to subtitle sentences). For each sequence element and for each of the views $\mathcal{V} = \{\text{image, audio, text}\}$ we have a feature vector $(x_{jt})^k$ where $k \in \{1, 2, 3\}$ (indexing the different modalities). In addition, we have labels y_{jt}. The problem is to infer the correct label y_{jt} for each x_{jt} at the time it presents itself: data points appear sequentially in a time series and the label prediction should be done using information from the past and the present elements only.

We use a sequence model for incremental modeling of sequential multi-view data. The overall architecture of such a system is that of a Recurrent Neural Network (RNN). A general time-unfolded overview of the architecture for a single example $x \in \mathcal{X}$ is shown in Figure 1. Input segments x_t^k of different views are fed to a cell at every time step, and a single vector h_t combining information from all views is generated. This embedding is fed to an output layer, which in turn outputs a prediction.

View Fusion One of the most important and distinguishing features of multi-view models is the framework it uses for the fusion of the different views. We use a multimodal GRU cell (Yang et al., 2017) as a base for experimentation. The cell takes embeddings x_t^k for each view k and time step t and after passing each view from a designated GRU cell, it calculates a multi-view embedding for the time step by passing information of all views through the gates of a separate GRU. The output of the view-specific GRUs is used to calculate the Pearson correlation between the different views; a calculation whose output is added as a separate term to be maximized in the total loss.

Since correlation can be calculated only for a pair of variables, we calculate the total correlation loss as the sum of the correlation losses between all pairs of elements of $\mathcal{V}$. Formally, in each step t, the correlation between views is calculated as

$$c_t = \sum_{\substack{k,\ell \in \mathcal{V} \\ k \neq \ell}} \frac{\sum\limits_{i=1}^{L} (h_{it}^{(k)} - \overline{H}_t^{(k)})(h_{it}^{(\ell)} - \overline{H}_t^{(\ell)})}{\sqrt{\sum\limits_{i=1}^{L} (h_{it}^{(k)} - \overline{H}_t^{(k)}) \sum\limits_{i=0}^{L} (h_{it}^{(\ell)} - \overline{H}_t^{(\ell)})}},$$

where i spans over the L elements of each mini-batch, $h_{it}^{(k)}$ is the hidden state calculated by the view-specific GRU for the example i, view k, at timestep t. Moreover, $\overline{H}_t^{(k)} = \frac{1}{L} \sum\limits_{i=1}^{L} h_{it}^{(k)}$.

The correlation term is

$$\mathcal{L}_{\text{corr}} = -\frac{1}{|T|} \sum_{t \in T} c_t,$$

that is the average of the loss calculated for every time step t. In order to maximize correlation, the negative sum is used.

In conclusion, view fusion is achieved not only by the design of the multimodal GRU cell itself (weighted sum of view representations and one common hidden representation), but also by the maximization of correlation between the views.

Learning The network is trained by jointly minimizing the following objective

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{corr}},$$

where $\mathcal{L}_{\text{CE}}$ stands for cross-entropy loss, $\mathcal{L}_{\text{corr}}$ is the correlation loss term defined in Section 3 and λ weights the contribution of $\mathcal{L}_{\text{corr}}$ to the total loss.

This compound objective function is one of the distinctive features of our approach. It enables the model to take advantage of the labels available from the dataset, and at the same time optimize for the correlation between the available views. The underlying idea is that the constraint of the correlation will push the model to create more informative embeddings than those it would create if only the cross-entropy loss was used.

Multi-head Attention Each of the videos of the dataset is assumed to be divided to snippets, with the sequence model operating and making predictions on the snippet level. For each snippet, the text modality may contain a different number of tokens and a fixed-length text representation is generated by a text encoder. We use an RNN as text encoder and the encoded text representation for each snippet is weighted by attention scores calculated over its tokens. Both "query" and "token" representations come from the same sequence (self-attention; Yang et al. 2016). In cases where the dataset contains, apart from snippet-level, also token-level annotations, the attention module can be directly supervised: we add a term

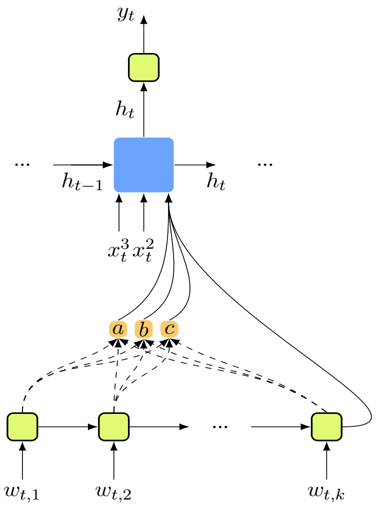


Figure 2: Hierarchical multi-view recurrent model with multi-head attention. Each sentence is encoded with an RNN and three attention heads (a, b and c) calculate attention scores for each of the tokens w_{ti} of the t-th sentence of the script.

to the model's loss that minimizes the error of the attention scores, with respect to token-level annotations. Specifically, we consider a case where the token-level annotations are of three different types and we create a supervised attention head for each type, averaging their outputs. A schematic depiction of this model can be seen in Figure 2.

4 Experiments

We describe in this section experiments with our proposed architecture. We use the CSI dataset (Frermann et al., 2018), which consists of 39 episodes of the television series *CSI: Crime Scene Investigation*. In each episode, a team of detectives undertake the solution of one (in 51% of the episodes) or two (49%) crime cases. The three modalities included are text scripts (dialogue subtitles and background scene descriptions), image snapshots from the video and audio segments. Each sentence of the script is aligned with image and audio[3] excerpts. Also, sentences are annotated with the case they belong to (binary label), perpetrator mention labels (binary) and the name of the speaker that uttered them ("None" for scene descriptions). Each speaker belongs to one of the types *detective, perpetrator, suspect, extra* (*none* for scene descriptions). An annotated example excerpt is shown in Figure 3.

4.1 Experimental Setup

For all experiments, we adopted an experimental setup similar to that of Frermann et al. (2018). For text, we use 50-dimensional GloVe vectors (Pennington et al., 2014) and a convolutional text encoder with maxpooling (filters of sizes 3, 4 and 5, each returning a 75-dimension output). Image features are generated by the final hidden layer of the inception-v4 Szegedy et al., 2017 model (dimensionality of 1,546). Audio features are constructed by concatenating five 13-dimensional Mel-Frequency Cepstral Coefficient (MFCC) feature vectors for each interval. For perpetrator mention identification, we use the case level splits, whereas the speaker and case tasks are performed on the episode level. All LSTM and GRU variants have one layer of length 128 and a dropout probability of 0.5 is used. We set the value of λ to 0.001 and train for 150 epochs with the Adam optimizer (Kingma and Ba, 2014), setting the initial learning rate at 0.001.

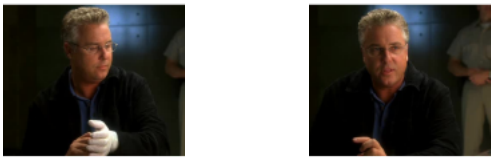
4.2 Perpetrator Mention Identification

In order to investigate the effectiveness of our model in the sequential multimodal inference setup, we conduct the following set of experiments on the task of perpetrator mention identification.

Multi-view Model The effectiveness of different architectures is shown in the first section of Table 1. The multi-view model (using corrGRU) is compared to early fusion models (using LSTM and GRU cells), for which the input is the concatenation of the feature vectors of the three modalities, passed through a ReLU activation. It can be seen that the multi-view model outperforms all other models.

Incremental Inference *Incremental sequence labeling* refers to making predictions on an incoming sequence when "streamed" in an online fashion. For example, if the sequence is a sentence, we are not allowed to encode the whole sentence first, but instead have to output a relevant label for each word as it arrives in the sequence. Incrementality underlies fundamental human cognition and is essential for scaling systems to large datasets and real-time inference, necessary, for example, in simultaneous translation (interpretation; Bangalore et al., 2012; Yarmohammadi et al., 2013; Cho and Esipova, 2016).

[3] Speech has been stripped from the audio track, leaving it only with audio effects and music (so that the text modality will not be deemed redundant and the dataset does not contain overlapping information).



NONE	GRISSOM
Grissom doesn't look worried. He takes his gloves off and puts them on the table.	You ever been to the theater, **Peter**?
CASE: 1	CASE: 1
PERP: 0	PERP: 1
CASE TAG: B-1	CASE TAG: I-1
SP. TAG: O	SP. TAG: B-D

Figure 3: An excerpt of the CSI dataset, where the image and text modalities are present. The case and speaker tags use the BIOU (Beginning, Inside, Outside, Unit) format (see 4.3 for details). In this case, "Peter" is the name of the perpetrator (Perp: 1). Both snapshots belong to the first case. The first snapshot does not have a speaker (screenplay description; speaker "None"and speaker type tag "O") and starts a chunk of utterances belonging to the first case (B-1). The second snapshot continues in the first case (I-1) and starts a chunk where the speaker is a detective (B-D).

In order to assess the effectiveness of the incremental inference capabilities of our model, we contrast the output of incremental models (forward-pass unidirectional) to that of similar models that do not perform incremental inference (bidirectional), in Table 1. Both multi-view and non-multi-view bidirectional models look ahead in the sequence, gathering information that is potentially useful for temporal inference. The bidirectional correlational model was trained with an extra correlation loss term, calculated exactly as $\mathcal{L}_{\text{corr}}$, with data from the backward pass. The bidirectional multi-view model does not score as high as the unidirectional multi-view one, though it balances better between precision and recall. Interestingly, results suggest that the incremental multi-view model outperforms early fusion bidirectional models (biLSTM and biGRU). Across experiments, there is a trend of higher recall. This can be justified by the fact that is the dataset is not balanced: positive examples of perpetrator mentions are significantly fewer than negative examples. This trend is consistent with the findings of Fremann et al. (2018), where an early fusion LSTM is used.

	MODEL	pr	re	f1
UNI-DIR	EF (LSTM)	42.8	51.2	46.6
	EF (GRU)	39.4	60.4	47.7
	MV (corrGRU)	41.3	**63.4**	**50.0**
BI-DIR	EF (biLSTM)	40.0	62.7	48.8
	EF (biGRU)	43.6	58.1	49.8
	MV (biCorrGRU)	**49.6**	49.4	49.5

Table 1: Precision (pr), recall (re) and F1 scores for detecting the minority class (perpetrator mentioned) on the held-out dataset. EF stands for early-fusion, while MV for multi-view. The first section of the table reports scores for unidirectional (incremental) models and the second for bidirectional (non-incremental) models. The result for the simple unidirectional LSTM model is the one reported by Fremann et al. (2018).

	MODALITY			pr	re	f1
	T	I	A			
THREE	✓	✓	✓	41.3	**63.4**	**50.0**
TWO	✓	✓		**41.4**	49.6	45.1
	✓		✓	39.6	50.4	44.3
		✓	✓	38.7	5.1	9.0
ONE	✓			41.9	47.3	44.4
		✓		28.4	6.7	10.8

Table 2: Ablation experiment assessing the contribution of each modality for our multi-view model. Precision (pr), recall (re) and F1 scores for detecting the minority class (perpetrator mentioned) on the held-out part of the dataset are reported. The modalities are denoted by T (text), I (image) and A (audio).

Contribution of the different modalities We conduct an ablation experiment assessing the contribution of each modality to the final prediction. The results of this experiment can be found in Table 2. Evidently, all three modalities contribute to the good performance of the multi-view model. We note that the *text* modality is the most informative; models taking text into account score consistently better, in both multi-view and single view setups. Results of the single-view video model suggest that the *image* modality alone provides very little information about the perpetrator's identity. Results on *audio* only are not reported, since the audio modality contains only music and audio effects and is not expected to generate useful representations by itself.

Supervised Multi-head Attention The CSI dataset includes token-level perpetrator mention annotations: every token of the script sentences is tagged as being a mention of the perpetrator or not. An example can be found in Figure 3 (right), where "Peter" is tagged as being a reference to the perpetrator (boldface). The two levels of annotation in the dataset follow a compositional structure: the presence of at least one token annotated with 1, results in the whole sentence annotated with 1. The sentence-level annotations used throughout the previous experimental section and throughout the work of Fremann et al. (2018) are generated by aggregating token-level annotations.

We distinguish three types of perpetrator token mentions in the dataset: *first person pronoun tokens* (the perpetrator is the speaker and speaks in first person), *pronoun tokens* (other characters refer to the perpetrator by using pronouns) and *other type of tokens* (perpetrator is mentioned by their name or other attributes). We replace the original binary token-level annotation with three binary annotation streams reflecting the three different types (first person pronoun mention, other person pronoun, other type of mention).

We replace the convolutional encoder of the previous experimental setup with an LSTM and three attention heads and run experiments comparing attentive architectures with non-attentive ones. The results can be found in Table 3. Unsurprisingly, models that make use of the extra information in the form of attention supervision score better than their counterpart that does not take token-level annotations into account. Interestingly, the complexity and diversity of token-level annotations is reflected in the results of single-head attention models: the supervised single-head model scores lower than the one that is free to learn any attention scores. Ultimately, the choice to split the annotations to three streams and use more than one attention heads, each focusing on different types of mentions, leads to better performance.

However, even the multi-head supervised attentive model, does not score as well as the multi-view (non-attentive) model. This result highlights the, sometimes disregarded, importance of modality fusion: creating a fused representation out of the available modalities led to a model that outperforms one with significantly more information in its disposal.

4.3 Episode Structure Tagging

Extracting knowledge about a movie by relying on simplified tasks can be challenging and may require assumptions about the input. Specifically, casting the task of perpetrator identification as a binary classification task, is based on the premise that there is, at most, one perpetrator. This assumption does not always hold, since, some episodes contain two cases and consequently, two perpetrators. Moreover, new perpetrators are introduced in every episode and data sparsity makes multi-class classification difficult. For the experiments described in the previous sections, we alleviate this obstacle by performing binary inference on the case level, using the annotated case splits of the dataset.

MODEL	pr	re	f1
EF	**42.8**	51.2	46.6
MV	41.3	**63.4**	**50.0**
EF+ATT	39.95	58.70	47.32
EF+SUPATT	40.72	56.11	47.15
EF+MULTIHEAD	40.25	59.19	47.88

Table 3: Comparing the performance of early fusion (EF) and multi-view (MV) models with attentive early fusion models. Three different attention schemes are compared: simple attention (ATT), supervised attention (SUPATT) where the network's loss includes an error term for the attention scores with respect to the token-level annotations, and multi-head supervised attention (MULTIHEAD) where the token-level annotations are divided conceptually into three groups and each head is supervised by the scores of one of the groups.

In order to enable more robust movie understanding, we investigate the automatic segmentation of episodes to coherent chunks by experimenting with tagging utterances with tags of two levels of granularity: case and speaker type. The former refers to associating each utterance with the crime case it belongs to, while the latter to labeling each utterance as coming from one type of speaker (detectives, perpetrators, suspects, extras) or none (scene descriptions).

The two tagging tasks are closely related, since a shift from a speaker type (e.g. a conversation between detectives) to another (a conversation between extras) may indicate a shift in the focus of the episode, hinting a case change. The presence of more than one related tasks makes our setup ideal for testing our model in a multi-task setting. Sharing representations between tasks is justified by the notion that information from similar tasks can aid in solving the task at hand faster and more accurately (Caruana, 1998).

MODEL	SPEAKER TYPE				CASE			
	acc	**pr**	**re**	**f1**	**acc**	**pr**	**re**	**f1**
EF	50.55	20.28	24.18	20.66	61.00	0.01	0.01	0.01
MV	57.66	19.87	35.95	25.29	61.65	0.03	2.86	0.05
EF+CRF	49.70	15.70	14.22	14.48	62.75	3.10	15.21	4.97
MV+CRF	51.27	14.89	16.71	14.96	73.53	**11.72**	27.75	11.24
MULTI-TASK (SPEAKER+CASE)								
EF	45.07	19.83	27.53	21.82	61.75	0.00	0.00	0.00
MV	46.11	18.17	22.36	19.09	61.02	0.06	0.08	0.06
EF+CRF	47.04	**22.02**	17.42	18.60	73.95	1.09	1.52	1.11
MV+CRF	**60.07**	21.36	**39.25**	**27.61**	**79.49**	8.29	**44.17**	**13.72**

Table 4: Macro-average scores for accuracy (acc), precision (pr), recall (re) and F1 scores for early fusion (EF) and multi-view (MV) models on speaker type and case tagging. The top section of the table refers to single-task setup, while the bottom on multi-task setup (training jointly on speaker type and case tagging).

To facilitate tagging, we convert the case and speaker annotations of the dataset to annotations employing the BIOU (Beginning, Inside, Outside, Unit) format, derived from the BIO scheme proposed for text chunking (Ramshaw and Marcus, 1999) and heavily used in the CoNLL shared tasks[4] for sentence tagging. An example of the labels on which the model operates can be found in Figure 3. We modify the architecture of our model, so that the output of the sequence model cell is fed to different output layers, one for each task. Training proceeds by summing the loss terms for both tasks. In the case of our multi-view model, the loss consists of two cross-entropy terms and one correlation term. Moreover, we experiment with adding a Conditional Random Field (CRF) on top of the sequence models, based on recent work that achieves state-of-the-art performance in tagging tasks, such as Named Entity Recognition (Lample et al., 2016).

The results for speaker type and case tagging can be found in Table 4, where our model (MV) is compared with an LSTM early fusion model. We use a variant of the evaluation script used for the CoNLL shared tasks[5] and report average scores. Our multi-view model consistently outperforms early fusion models. Interestingly, the multi-task MV+CRF model trained exhibits the best performance, suggesting that jointly solving the two tasks improves the capabilities of the model.

5 Related Work

Inference on multimodal sequences can take the form of inferring a label for a whole sequence, or a label for each of the parts of it. The multimodal LSTM of Ren et al. (2016) is applied in a sequential inference setting, however, it does not produce a joint representation for all modalities, but rather, different (albeit informed about each other) modality representations are used for inference.

Our approach is more related to that of Yang et al. (2017), where a multimodal encoder-decoder model for representation learning of temporal data is described. Our method uses their corrGRU cell with a distinct architecture and loss function: first, their network is used as an unsupervised sequence representation learning tool trained to generate an embedding for a whole sequence, which in turn is used for classification tasks, while our model outputs embeddings and makes predictions at the token-level (for every element of a sequence). Secondly, they use a decoder which reconstructs the original representations of each view, while our model does not include autoencoding.

Casting correlation maximization between three or more variables as the maximization of the sum of the correlation between all pairs of available variables has been previously used in extensions of CCA for more than two views (Benton et al., 2019), or other multi-view learning works (Kumar et al., 2011). Yang et al. (2017) mention it in their paper, although they do not experiment with it.

The multi-head attention component of our model bears similarities in spirit to the recent work of Strubell et al. (2018), where an attention head is replaced by a model trained to predict syntac-

[4] http://www.conll.org/previous-tasks

[5] https://github.com/spyysalo/conlleval.py

tic dependencies (Dozat and Manning, 2017). In contrast, our model uses explicit supervision for all self-attention heads and is trained to predict the correct attention scores in a multi-task fashion.

6 Conclusions

We describe a neural multi-view sequential architecture, paired with a novel objective that takes advantage of supervision, while at the same time, maximizes the correlation between views. We test our approach on the task of perpetrator mention identification of the CSI dataset, on which we show that it outperforms state of the art. Also, we introduce two shallow movie understanding tasks, crime case and speaker type tagging, and show that our model yields consistently better results than early fusion models, highlighting the importance of careful fusion of modalities in sequential inference.

Acknowledgments

The authors gratefully acknowledge the support of the European Research Council (Lapata; award number 681760, "Translating Multiple Modalities into Text"), the H2020 EU project SUMMA (award number 688139/H2020-ICT-2015; Papasarantopoulos, Cohen) and Bloomberg (Cohen).

References

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 437–445. Association for Computational Linguistics.

James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, NY, USA.

Adrian Benton, Huda Khayrallah, Biman Gujral, Drew Reisinger, Sheng Zhang, and Raman Arora. 2019. Deep generalized canonical correlation analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019*, pages 1–6.

Hervé Bredin, Anindya Roy, Nicolas Pécheux, and Alexandre Allauzen. 2014. "Sheldon speaking, bonjour!": Leveraging multilingual tracks for (weakly) supervised speaker identification. In *Proceedings of the 22nd ACM International Conference on Multimedia*.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *arXiv preprint arXiv:1609.03976*.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1913–1924.

Rich Caruana. 1998. Multitask learning. In *Learning to learn*, pages 95–133. Springer.

Xiaobin Chang, Tao Xiang, and Timothy M Hospedales. 2018. Scalable and effective deep cca via soft decorrelation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1488–1497.

Jiajun Cheng, Shenglin Zhao, Jiani Zhang, Irwin King, Xin Zhang, and Hui Wang. 2017. Aspect-level sentiment classification with HEAT (hierarchical attention) network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 97–106. ACM.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.

Manfred Del Fabro and Laszlo Böszörmenyi. 2013. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia systems*, 19(5):427–454.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Lea Frermann, Shay Cohen, and Mirella Lapata. 2018. Whodunnit? crime drama as a case for natural language understanding. *Transactions of the Association for Computational Linguistics*, 6:1–15.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R. Hershey, and Tim K. Marks. 2017. Attention-based multimodal fusion for video description. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4203–4212.

Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111.

Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2017. Supervised attention for sequence-to-sequence constituency parsing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 7–12.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022. AAAI Press.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Elena Knyazeva, Guillaume Wisniewski, Hervé Bredin, and François Yvon. 2015. Structured prediction for speaker identification in tv series. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Abhishek Kumar, Piyush Rai, and Hal Daume. 2011. Co-regularized multi-view spectral clustering. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1413–1421. Curran Associates, Inc.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Eleventh International Conference on Language Resources and Evaluation*, pages 3810–3817, Miyazaki, Japan.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 260–270.

Cailiang Liu, Dong Wang, Jun Zhu, and Bo Zhang. 2013. Learning a contextual multi-thread model for movie/tv scene segmentation. *IEEE Transactions on Multimedia*, 15:884–897.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. 2017a. Attention correctness in neural image captioning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4176–4182.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017b. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1789–1798.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2283–2288.

Xavier Anguera Miró, Simon Bozonnet, Nicholas W. D. Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:356–370.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 169–176. ACM.

Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105.

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)*, pages 1532–1543.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*, pages 338–353. Springer.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

James M Rehg, Gregory D Abowd, Agata Rozga, Mario Romero, Mark A Clements, Stan Sclaroff, Irfan Essa, Opal Y Ousley, Yin Li, Chanho Kim, et al. 2013. Decoding children's social behavior. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3414–3421. IEEE.

Jimmy SJ Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun, and Qiong Yan. 2016. Look, listen and learn-a multimodal lstm for speaker identification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Joseph Pal, Hugo Larochelle, Aaron C. Courville, and Bernt Schiele.

2016. Movie description. *International Journal of Computer Vision*, 123:94–120.

Yue Shi, Martha Larson, and Alan Hanjalic. 2013. Mining contextual movie similarity with matrix factorization for context-aware recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):16.

Michael A Smith and Takeo Kanade. 1998. Video skimming and characterization through the combination of image and language understanding. In *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pages 61–70. IEEE.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. pages 5027–5038.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, volume 4, page 12.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4631–4640.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A. Bernal, and Jiebo Luo. 2017. Deep multimodal representation learning from temporal data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiwang Yang, Harald Steck, and Yong Liu. 2012. Circle-based recommendation in online social networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1267–1275. ACM.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1480–1489.

Mahsa Yarmohammadi, Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Baskaran Sankaran. 2013. Incremental segmentation and decoding strategies for simultaneous translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1032–1036.

Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. 2017. Supervising neural attention models for video captioning by human gaze data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii*, pages 2680–29.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5634–5641.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5642–5649.