Qualcomm's Neural Processing SDK Plugin for Unreal Engine

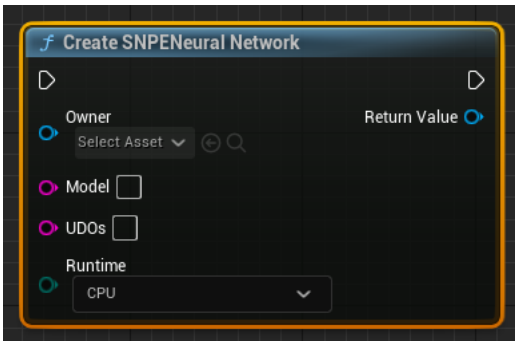
Blueprint reference guide

SNPE Neural Network Object

UObject that encapsulates a SNPE neural network and provides functionality to run inference synchronously and asynchronously.

Create SNPE Neural Network.

Creates a SNPE neural network object using a dlc model for Qualcomm's Neural Processing SDK. Using this object, you can run inference calls.



Owner: Object owning the SNPE Neural Network class.

Mode: path to DLC file

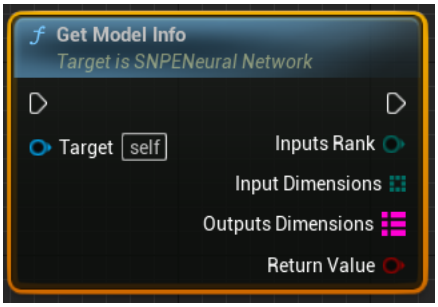
UDO's: string specifying the UDO (User defined Operation) files separated by comma.

Runtime: Target runtime (CPU, GPU, DSP)

Return Value: SNPE Neural Network object.

Get Model Info (SNPENeural Network)

Retrieve the information of the inputs and the outputs of a SNPENeural Network model.



Target: SNPE Neural Network Object from which the information will be retrieved.

Inputs Rank: Number of dimensions the input has

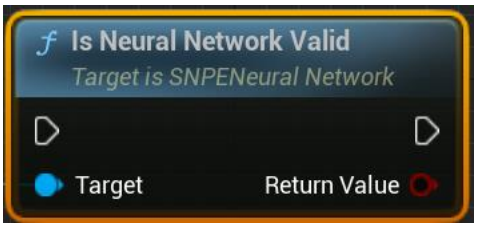
Input Dimensions: Array of dimension values of the input

Outputs Dimensions: Map containing the output name and a string of values separated by commas representing the output dimensions.

Return Value: returns true or false if the request was successful or not.

Is Neural Network Valid (SNPENeural Network)

Verifies if the neural network is initialized correctly and ready to run infererence

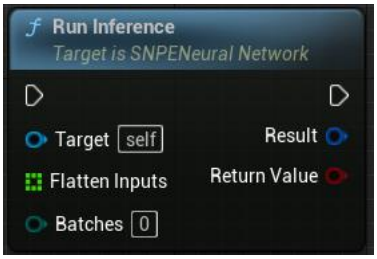


Target: SNPE Neural Network Object from which the information will be retrieved.

Return Value: True or false if the neural network is ready or not.

Run Inference(SNPENeural Network)

Execute a synchronous inference request.



Target: SNPE Neural Network Object on which the inference request will be performed.

Flatten Inputs: Input for the Inference request in a flatten array.

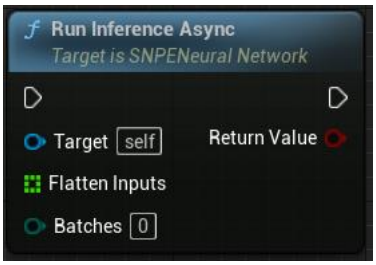
Batches: Number of batches included in the Flatten Inputs.

Result: Struct containing the result values and information of the outputs.

Return Value: returns true or false if the request was successful or not.

Run Inference Async (SNPENeural Network)

Execute an asynchronous inference request. When the result is available SNPENeural Network will raise an event “ResultCallbackBP” with the result information. You should bind a function to “ResultCallbackBP” delegate to retrieve the information of the request.



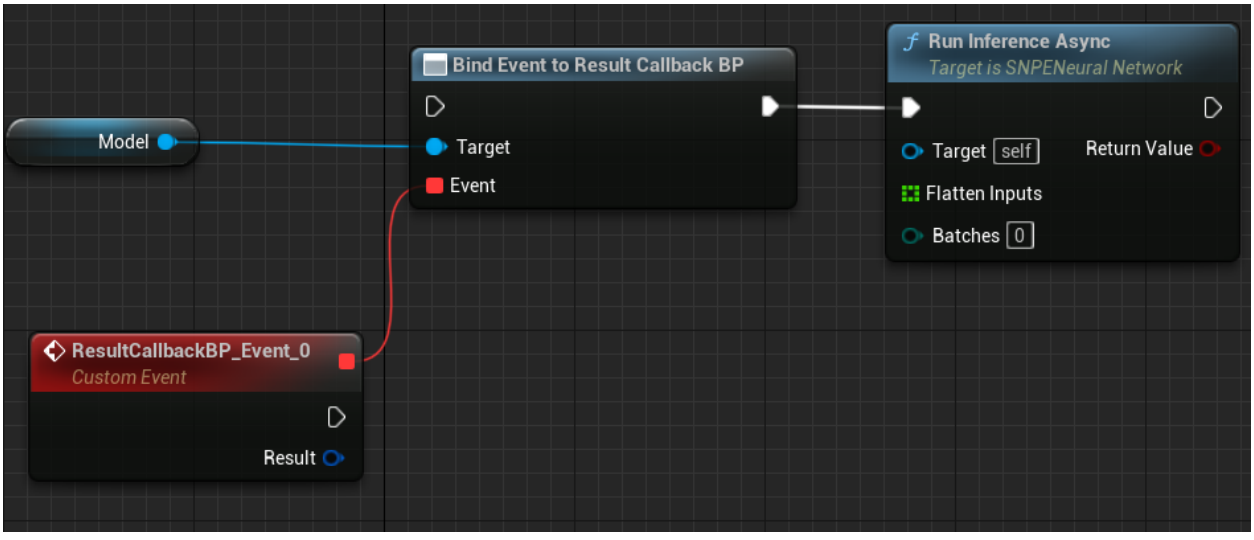
Target: SNPE Neural Network Object on which the inference request will be performed.

Flatten Inputs: Input for the Inference request in a flatten array.

Batches: Number of batches included in the Flatten Inputs.

Return Value: returns true or false if the request was successful or not.

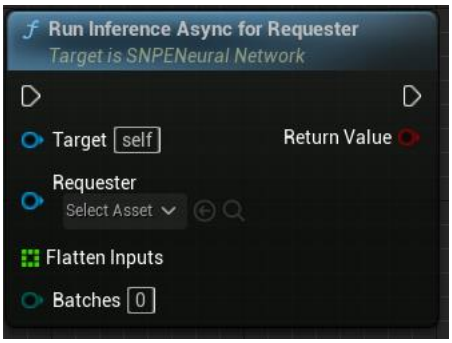
a. *How to bind a callback function to Run Inference Async:*



ResultCallbackBP_Event_0 will be called after the Run Inference Async finishes its request, with the result exposed on the pin “Result”

Run Inference Async for Requester (SNPENeural Network)

Execute an asynchronous inference request, that will be handled by a requester UObject other than the SNPENeural Network. The UObject must implement the C++ interface ISNPEInferenceRequestHandler.



Target: SNPE Neural Network Object on which the inference request will be performed.

Requester: UObject that will handle the callback. It must implement the C++ interface ISNPEInferenceRequestHandler and override the function ISNPEInferenceRequestHandler:: InferenceResponseHandling(const FSNPE_Result& Result) to retrieve the result.

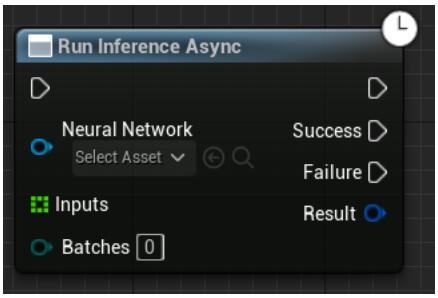
Flatten Inputs: Input for the Inference request in a flatten array.

Batches: Number of batches included in the Flatten Inputs.

Return Value: returns true or false if the request was successful or not.

Run Inference Async (Static function)

An alternative way to call async inference is to use the static function Run Inference Async which provide directly two delayed async execution pins for success of failure of the call.



Target: SNPE Neural Network Object on which the inference request will be performed.

Inputs: Input for the Inference request in a flatten array.

Batches: Number of batches included in the Flatten Inputs.

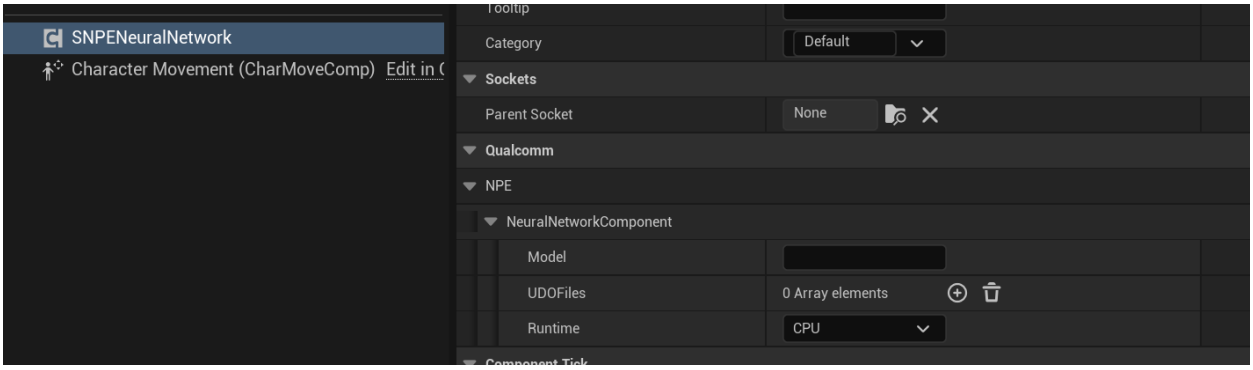
Success: Event raised when the results are ready and the operation was successful.

Failure: Event raised when the operation finishes unsuccessfully.

Result: Struct containing the result values and information of the outputs.

SNPE Neural Network Component

Actor Component that encapsulates an SNPE Neural network and can be used to run inference synchronously and asynchronously.



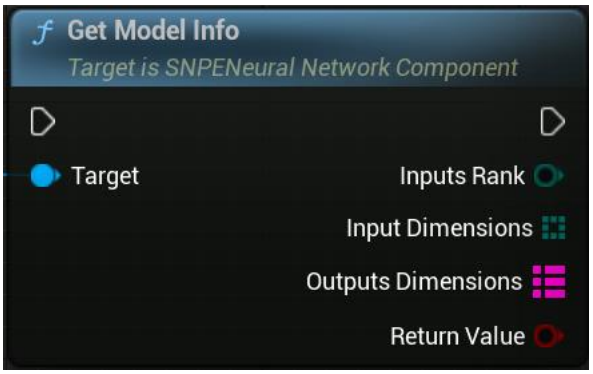
When added, you can configure the parameters of the SNPE model.

Model: path the DLC model file.

UDOFiles: Array of string representing the UDO (User defined operations) used by your model.

Runtime: Execution target (CPU, GPU, DSP)

Get Model Info (SNPENeural Network Component)



Target: SNPE Neural Network Component from which the information will be retrieved.

Inputs Rank: Number of dimensions the input has

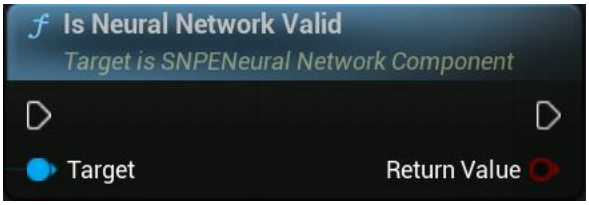
Input Dimensions: Array of dimension values of the input

Outputs Dimensions: Map containing the output name and a string of values separated by commas representing the output dimensions.

Return Value: returns true or false if the request was successful or not.

Is Neural Network Valid (SNPENeural Network Component)

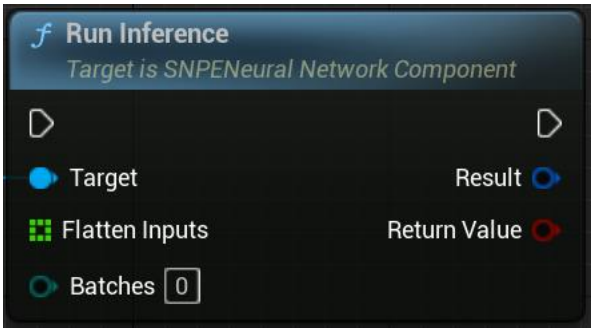
Verifies if the neural network is initialized correctly and ready to run infererence



Target: SNPE Neural Network Component from which the information will be retrieved.

Return Value: True or false if the neural network is ready or not.

Run Inference (SNPENeural Network Component)



Target: SNPE Neural Network Component on which the inference request will be performed.

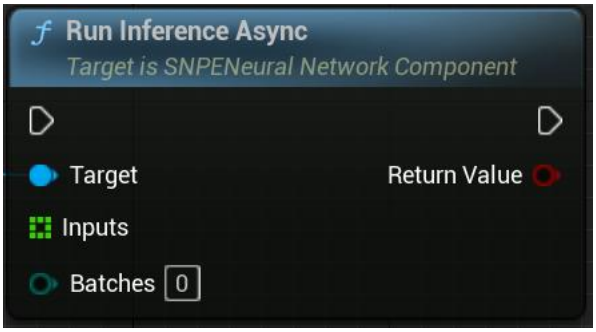
Flatten Inputs: Input for the Inference request in a flatten array.

Batches: Number of batches included in the Flatten Inputs.

Result: Struct containing the result values and information of the outputs.

Return Value: returns true or false if the request was successful or not.

Run Inference Async (SNPENeural Network Component)



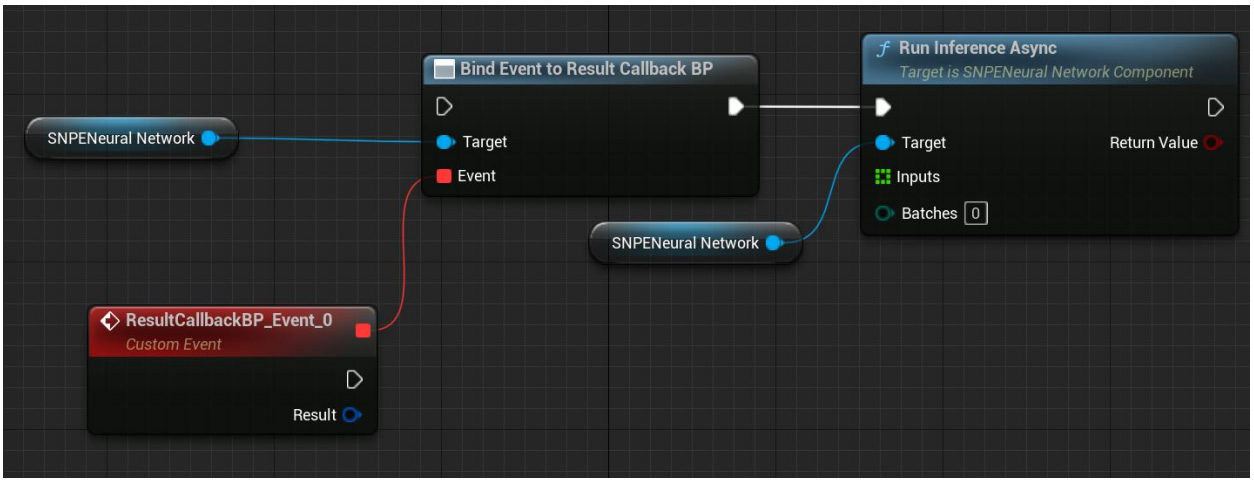
Target: SNPE Neural Network Component on which the inference request will be performed.

Flatten Inputs: Input for the Inference request in a flatten array.

Batches: Number of batches included in the Flatten Inputs.

Return Value: returns true or false if the request was successful or not.

As with the SNPENeural Network Object, you will need to bind a callback function to the component to retrieve the result.



SNPE Result

The inference results are returned in a struct called SNPE Result. Its members are:

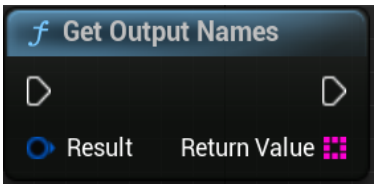
TMap<FString, TArray<TArray<float>>> Outputs = Map of the result values of each output.

TMap<FString, TArray<uint8>> OutputsDimensions = Map containing the dimensions of each output.

bool Success = Result of the inference request.

There are helper functions to retrieve the result information.

Get Outputs Names (SNPE Result)



Result: SNPEResult struct.

Return Values: Array containing the outputs names.

Get Flatten Values At (SNPE Result)

Result: SNPEResult struct.

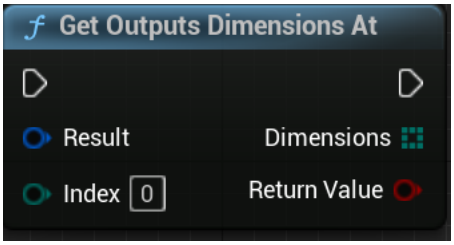
Index: Index in the map of outputs from which the values are requested.

Output: Array containing the values for the output requested (flatten)

Batches: If the output contains multiple batches, this value indicate how many batches are included in the output. This value is the same number as the batches value passed in the inference request.

Return Value: Returns true if the request is valid.

Get Outputs Dimensions At (SNPE Result)



Result: SNPEResult struct.

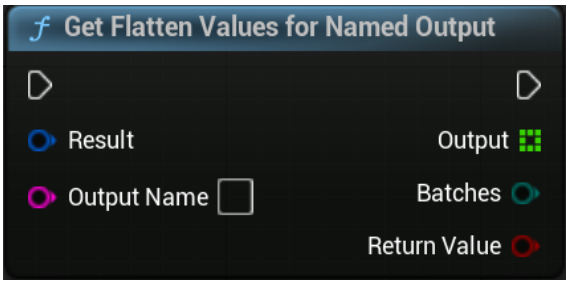
Index: Index in the map of outputs from which the values are requested.

Output: Array containing the dimensions for the output requested.

Batches: If the output contains multiple batches, this value indicate how many batches are included in the output. This value is the same number as the batches value passed in the inference request.

Return Value: Returns true if the request is valid.

Get Flatten Values for Named Output (SNPE Result)



Result: SNPEResult struct.

Output Name: Name of the output from which the values are requested.

Output: Array containing the values for the output requested (flatten)

Batches: If the output contains multiple batches, this value indicate how many batches are included in the output. This value is the same number as the batches value passed in the inference request.

Return Value: Returns true if the request is valid.

Get Outputs Dimensions for Named Output (SNPE Result)

Result: SNPEResult struct.

Output Name: Name of the output from which the values are requested.

Output: Array containing the dimensions for the output requested.

Batches: If the output contains multiple batches, this value indicate how many batches are included in the output. This value is the same number as the batches value passed in the inference request.

Return Value: Returns true if the request is valid.